Next-generation sequencing and genome analysis in dimorphic fungi and human: using genomic variation to recognize and understand disease

Juan Esteban Gallo Bonilla

Ph.D. thesis in Biomedical Sciences

Doctoral Program in Biomedical Sciences

Universidad del Rosario

Bogotá, Colombia

Tutor

Oliver Keatinge Clay, Ph.D.

Professor, School of Medicine and Health Sciences,

Universidad del Rosario, Bogotá, Colombia

COLEGIO MAYOR DE NUESTRA SEÑORA DEL ROSARIO

ESCUELA DE MEDICINA Y CIENCIAS DE LA SALUD

BOGOTA, DC

2017

Dedication

I dedicate this thesis to my family, wife and son.

Acknowledgements

There are many individuals who have left an important impression on my scientific career during the time of my doctoral studies, and without whom this thesis would not have been possible. I am deeply thankful for all input received by all of those who have helped me along the way.

First and foremost, I would like extend my most sincere thanks to my advisors and mentors. Specifically, I am indebted to Dr. Oliver Keatinge Clay for his valuable help, guidance, patience and teachings from the beginning of my work at CIB. I would also like to thank Dr. Juan McEwen for his insightful guidance and help, as well as for the opportunity to work in his research group. I thank Dr. Angela Restrepo, founder of the CIB, for her willingness to help me join the CIB team. My most sincere thanks to Dr. Beatriz Gomez for her encouragement and valuable input in our primer design studies.

I am very thankful for my colleagues at CIB, specifically the BCM group for teaching me the value of friendship and collaboration. I would like to thank Ana Maria Garcia, Orville Hernandez, Diana Tamayo, Oscar Gomez, Ana Aristizabal, and Luisa Gomez. I am grateful for the valuable contributions of Jose Fernando Muñoz and Elizabeth Misas, my bioinformatics colleagues at CIB. I specifically thank Isaura Torres for her insightful guidance in preparation and revision of this thesis.

I would like to extend my gratitude to the Universidad del Rosario for their academic support and doctoral scholarship. I thank Dr. Leonardo Palacios, Dr. Luisa

Matheus, Dr. Juan Posada, Dr. James Richardson, Dr. Gustavo Quintero and Dr. Juan Manuel Anaya.

I thank the bioinformatics and genomics group at Georgia Tech, specifically Professors Dr. King Jordan and Dr. Fredrick Vannberg for allowing me to work in their labs during my Fulbright fellowship. I would like to extend my thanks to Lavanya Rishishwar for his valuable contributions during my stay at Georgia Tech.

I thank Dr. Mary Brandt and Dr. Anatasia Litvintseva for allowing me to work in the Mycotic Diseases Branch lab during my Fulbright Fellowship. I also thank the Mycotic Diseases Branch group at CDC for their kind hospitality during our visit.

I am forever grateful to my family for their unlimited support and energy throughout these years, and for always pushing me to strive for the best. I thank God for giving me the opportunity, health, energy and the willingness to start and finish my doctoral studies.

Table of Contents

Abstract

The work of this thesis has a common focus on bioinformatics, comparative genomics of fungal genomes and clinical genomics of human chronic disease. We primarily focused on using genomic data of dimorphic fungal pathogens in order to obtain a better perspective and understanding of how commonly used assembly, annotation and genomic comparison bioinformatics programs dealt with genomic data.

We began by exploring the advantages and disadvantages of using raw read data for comparative genomics applications and found that the often sought after goal of a finished assembly can be a utopian illusion. We proposed that familiar, gene-sized sequences, including but not limited to nuclear protein-coding genes, would provide feasible consensus benchmarks allowing simple visualization. The exploratory analysis of such familiar candidate loci constituted a step toward finding, testing and establishing familiar, biologically interpretable consensus benchmark sequences for fungal and other eukaryotic genomes. We assembled the genomes of various strains of *Paracoccidioides brasiliensis* while varying *k*-mer lengths from 17 to 127 bp using SOAPdenovo and Velvet, and found a striking example of how certain *k*-mer lengths (between 49 and 51 bp) caused one of these programs to include a substantial amount of repetitive content as an artifact, which misrepresents the true fungal genomic content.

Our group re-sequenced the reference strains that had been used for the existing assemblies and annotations of *Paracoccidioides* spp., and used the new, higher quality reads to substantially improve the reference assemblies and annotations

of these pathogenic fungi. We also sequenced *de novo* the species *Emmonsia crescens* and *E. parva*, which are closely related to the causal agent of blastomycosis, *Blastomyces dermatitidis*, but non-pathogenic or with low virulence. We performed comparative analyses of gene content and structure between various strains of *B. dermatitidis, E. crescens* and *E. parva.* We screened a sample of 11 fully sequenced fungal mitochondrial genomes by observing where exact *k*-mer repeats occurred several times and found that such screening may serve as an efficient expedient for gaining a rapid but representative first insight into the repeat landscapes of sparsely characterized mitochondrial chromosomes. We extended our studies of *Paracoccidioides* spp. and sequenced, assembled and annotated 31 isolates representing the phylogenetic, geographic, and ecological breadth of the genus. These samples included clinical, environmental and laboratory reference strains of the S1, PS2, PS3, and PS4 lineages of *P. brasiliensis* and also isolates of the species *Paracoccidioides lutzii.* We completed the first annotated genome assemblies for the PS3 and PS4 lineages and found that gene order was highly conserved across the major lineages, with only a few chromosomal rearrangements.

Using available sequences, we then designed and analytically validated two primer pairs from regions that are unique to *Histoplasma capsulatum* but present across diverse strains of this species, and can therefore be utilized to detect the presence of *H. capsulatum*. Using the same approach, we also designed and analytically validated three primer pairs of high confidence for the amplification of sequence fragments that are unique to the genus *Paracoccidioides.* We designed and

implemented an algorithm that takes any given sequence(s) and splits the sequence into fragments in order to query the unique percentage of the fragment against a group of sequences that are closely related to the query as well as outgroups that may be relevant in clinical settings, including human.

In the section of human genomics, we genotyped 67 selected SNPs within the 9p21.3 locus of the human genome, motivated by its proven association with cardiovascular disease, for a Colombian cohort of 357 healthy individuals with data collected for detailed hemodynamic and other phenotypic traits. We found consistent associations of SNPs in a 60 kb subregion inside this locus with the phenotypic category of blood pressure. We also sequenced the exome of a patient with familiar hypercholesterolemia. Our findings helped the guide the patient's treatment using an alternative next generation cholesterol lowering drug which showed promising results in treatment.

Resumen

Este trabajo de tesis tiene un enfoque común en bioinformática, genómica comparativa de genomas fúngicos y genómica clínica de la enfermedad crónica humana. Nos centramos principalmente en el uso de datos genómicos de patógenos fúngicos dimórficos con el fin de obtener una mejor perspectiva y comprender como de manera común los programas bioinformáticos de ensamblaje, anotación y los estudios de genómica comparativa utilizan los datos genómicos.

Nuestro estudio inicia explorando las ventajas y desventajas de usar datos crudos de lecturas "reads" para aplicaciones en genómica comparativa, y encontramos que el objetivo de lograr un ensamblaje terminado puede ser una ilusión utópica. Propusimos entonces que secuencias familiares o conocidas, con características de genes nucleares codificantes de proteínas y otras regiones no codificantes, servirían como un punto de referencia de secuencias consenso útiles para la visualización simple de estas secuencias. El análisis exploratorio de este tipo de loci candidato familiar constituyó un paso hacia la búsqueda y prueba de estas secuencias usadas como punto de referencia con significado biológico para hongos y otros genomas eucariotas. Ensamblamos los genomas de varias cepas de *Paracoccidioides brasiliensis*, variando las longitudes de los *k*-mer de 17 a 127 pb usando los programas SOAPdenovo y Velvet, y encontramos un ejemplo sorprendente de como ciertas longitudes *k*-mer (entre 49 y 51 pb) causaron que uno de estos programas incluyera una cantidad sustancial de secuencias repetitivas como artefactos, lo cual tergiversa el ensamblaje del genoma fúngico.

Nuestro grupo realizó la re-secuenciación de varias cepas de referencia que habían sido utilizadas para los ensamblajes y las anotaciones existentes de *Paracoccidioides* spp. Como producto de esta re-secuenciación se obtuvieron lecturas "reads" de alta calidad que mejoraron sustancialmente los ensamblajes y las anotaciones de referencia de estos hongos patógenos. También secuenciamos *de novo* las especies *Emmonsia crescens* y *E. parva*, las cuales se hallan estrechamente relacionadas con el agente causal de la blastomicosis, *Blastomyces dermatitidis*, siendo estas especies de *Emmonsia* no patógenas o con baja virulencia. Se realizaron análisis comparativos del contenido de genes y su estructura entre diversas cepas de *B. dermatitidis*, *E. crescens* y *E. parva*. Se realizó una selección de 11 genomas mitocondriales de hongos totalmente secuenciados determinando la ubicación exacta donde ocurrieron repeticiones de los *k*-mer y se encontró que tal determinación puede servir como un recurso eficiente para obtener una primera vista rápida pero representativa del panorama de secuencias repetidas en las cromosomas mitocondriales escasamente caracterizados.

Extendimos nuestros estudios de *Paracoccidioides* spp.: se secuenciaron, ensamblaron y anotaron 31 aislamientos que representan la amplitud filogenética, geográfica y ecológica del género. Estas muestras incluyeron cepas de referencia clínicas, ambientales y de laboratorio de los linajes S1, PS2, PS3 y PS4 de *P. brasiliensis* y también aislamientos de la especie *Paracoccidioides lutzii*. Completamos los primeros ensambles del genoma anotado para los linajes PS3 y PS4

y encontramos que el orden de los genes estaba altamente conservado a través de los principales linajes, con sólo unos pocos reordenamientos cromosómicos.

Utilizando las secuencias disponibles, hemos diseñado y validado analíticamente dos pares de cebadores que hibridan en regiones que son únicas para *Histoplasma capsulatum*, las cuales están presente en las diversas cepas conocidas de esta especie, y por lo tanto, pueden utilizarse para detectar la presencia de *H. capsulatum*. Siguiendo el mismo enfoque, también diseñamos y validamos analíticamente tres pares de cebadores de alta confianza para la amplificación de fragmentos de secuencia que son únicas para el género *Paracoccidioides*. Hemos diseñado e implementado un algoritmo que toma cualquier secuencia dada y la divide en fragmentos con el fin de determinar el porcentaje único del fragmento comparado con un grupo de secuencias que están estrechamente relacionados con la secuencia de interés así como grupos externos "outgroups" que pueden ser relevantes en un escenario clínico, incluidos los humanos.

En la sección de genómica humana, se realizó la genotipificación de 67 SNPs seleccionados dentro del locus 9p21.3 del genoma humano, motivado por su asociación comprobada con las enfermedades cardiovasculares, para una cohorte colombiana de 357 individuos sanos de los que se tenia información acerca de sus rasgos hemodinámicos detallados y otros rasgos fenotípicos. Encontramos asociaciones consistentes de SNPs en una subregión de 60 kb dentro de este locus con la categoría fenotípica de la presión arterial. También se secuenció el exoma de un paciente con hipercolesterolemia familiar. Nuestros hallazgos ayudaron a guiar el

tratamiento del paciente usando un fármaco alternativo de nueva generación para reducir el colesterol el cual mostró resultados prometedores en su tratamiento.

Chapter 1

Presentation of the work

1.1 Preamble

The work of this thesis has a common focus on bioinformatics, comparative genomics of fungal genomes and clinical genomics of human chronic disease. Although genomics of fungal infectious diseases and human chronic disease may sound worlds apart, a common ground underlies both studies. The shared denominator is the eukaryotic genome and the importance of unraveling its genomic structure and content to better understand pathogenesis. Bioinformatics and genomics have helped to bridge the gap between the genomic data produced in research and their clinical interpretation in the last decade. This applies to infectious and human chronic disease advances. In fungal genomics, when a given genomic sequence of a pathogenic fungal species is available, one can compare its genome with the genomes of phylogenetically related species in order to better understand its genomic structure and function. The feasible application of whole-genome sequencing, assembly, and comparative analyses to one of the simplest eukaryotic forms, unicellular fungi, makes this an ideal group of model organisms to choose, in order to fine-tune techniques and develop methods that can be applied in genomics research. Hence, I primarily focused on using genomic data of dimorphic fungal pathogens in order to obtain a better perspective and understanding of how commonly used assembly, annotation and genomic comparison bioinformatics programs dealt with genomic data. Throughout the course of my studies, our groutook on two large research projects funded by Colciencias: (1) *A gene atlas for human pathogenic fungi* and (2) *Population analysis and molecular understanding of the aging diseases locus 9p21.3 in a Colombian cohort: a multi-level study*. The work

presented in this thesis falls within the scope of both projects. In order to facilitate comprehension for the reader, I present the fungal genomics sections first, followed by the sections corresponding to human chronic disease.

1.2 Fungal genomics

1.2.1 Overview

Several thermal dimorphic fungi in the Onygenales order cause respiratory diseases, especially in immunocompromised patients. The clinical relevance of genomic studies of fungal pathogens is supported by the need to understand their genes, the regulation of the genes, and the proteins they encode, and then to apply insights from that understanding to the development of methods for detection or treatment. As a final result, we were able to describe, improve and compare various genomic assemblies and annotations of species within the Ajellomycetaceae family. We also accomplished a total of 5 novel molecular detection designs via conventional and real time PCR that were analytically validated with promising clinical and environmental applications.

1.1 Presentation of fungal genomics chapters

In these fungal genomic sections, our overall objective was to utilize bioinformatics tools to better understand the process of how genomic differences contribute to pathogenicity of thermal dimorphic pathogenic fungi, as well as how one can improve current genomic sequences and their annotation in order to contribute to current molecular detection methods, based on unique genomic region amplification. Such methods may open possibilities for designing novel primer pairs that could potentially

achieve higher levels of specificity and sensitivity using conventional molecular detection techniques without sacrificing cost.

Initially, the main research interest of our group was to sequence, assemble and compare the genomes of several pathogenic fungal species in the Onygenales order. Initial results allowed us to better understand the complex nature of the genomic organization and content of fungal genomes. We began this journey by studying how the valuable and informative content of fungal genomic reads can be quickly queried to search for gene content. We explored the ups and downs of using raw read data for comparative genomics applications and found that the often sought after goal of a finished assembly can be a utopian illusion. We published the work titled *"The eukaryotic genome, its reads and the unfinished assembly"* in FEBS Letters in 2013, found in chapter 2 (Muñoz et al., 2013).

Next, we considered that assembly analyses could be greatly facilitated by using simple DNA sequence benchmarks, i.e., standard test sequences that could monitor or help to predict ease or difficulty of (a) short-read sequencing and (b) *de novo* assembly of the sequenced reads. We proposed that familiar, gene-sized sequences, including but not limited to nuclear protein-coding genes, would provide feasible consensus benchmarks allowing simple visualization. The exploratory analysis of such familiar candidate loci constituted a step toward finding, testing and establishing familiar, biologically interpretable consensus benchmark sequences for fungal and other eukaryotic genomes. We published this work titled *"Limits to Sequencing and de novo Assembly: Classic Benchmark Sequences for Optimizing*

Fungal NGS Designs" in Advances in Computational Biology in 2014, found in chapter 3 (Muñoz et al., 2014).

As a result of the previous two publications, and now understanding the complex nature of raw genomic data and their subsequent assembly, we were interested in better understanding how some commonly used *de novo* assembly algorithms scaffold reads when modifying kmers choices. In this study, we assembled the genomes of various strains of *Paracoccidioides brasiliensis* while varying *k*-mer lengths from *k*=17 to *k*=127 using SOAPdenovo and Velvet. We found a striking example showing how certain *k*-mer lengths, specifically between *k*=49 and *k*=51 in the SOAPdenovo package, caused a substantial amount of repetitive content to be included in the genomic assemblies as potential artifacts, which misrepresents the true fungal genomic content. We published the work titled *"The complex task of choosing a de novo assembly: Lessons from fungal genomes"* in the Journal of Computational Biology and Chemistry in 2014, found in chapter 4 (Gallo et al., 2014).

At this point, our group had re-sequenced the reference strains used in the assemblies and annotations of *Paracoccidioides* spp., as well as sequencing *de novo* the species *Emmonsia crescens* and *E. parva*, which are closely related to the causal agent of blastomycosis, *Blastomyces dermatitidis*, but non-pathogenic or with low virulence. Although the genomes of *Paracoccidioides* spp. were previously published in 2011, the assemblies and annotations could be greatly improved by re-sequencing the reference strains, as well as other representative strains of these species. We improved the assembly and annotation of the *Paracoccidioides* spp. and published the

work titled *"Genome Update of the Dimorphic Human Pathogenic Fungi Causing Paracoccidioidomycosis"* in PLOS Tropical Neglected Diseases in 2014, found in chapter 5. In parallel, we continued the comparative analysis of gene content and structure between various strains of *B. dermatitidis, E. crescens* and *E. parva.* We published the work titled *"The Dynamic Genome and Transcriptome of the Human Fungal Pathogen Blastomyces and Close Relative Emmonsia"* in PLOS Genetics in 2015, found in chapter 6 (Muñoz et al., 2015).

Up to this point, we had only considered nuclear genomic DNA, yet mitochondrial DNA (mtDNA) plays a very important role in the biological functions of all species. The presence of repetitive or non-unique DNA persisting over sizable regions of a eukaryotic genome can hinder the genome's successful *de novo* assembly from short reads: ambiguities in assigning genome locations to the non-unique subsequences can result in premature termination of contigs and thus overfragmented assemblies. Fungal mtDNA genomes are compact (typically less than 100 kb), yet often contain short non-unique sequences that can be shown to impede their successful *de novo* assembly *in silico*. Such repeats can also confuse processes in the cell *in vivo*. We screened a sample of 11 fully sequenced fungal mitochondrial genomes by observing where exact *k*-mer repeats occurred several times. We found that such screening may serve as an efficient expedient for gaining a rapid but representative first insight into the repeat landscapes of sparsely characterized mitochondrial chromosomes. We published the work titled *"From NGS assembly challenges to instability of fungal mitochondrial genomes: a case study in genome complexity"* in

Computational Biology and Chemistry in 2016, found in chapter 7 (Misas et al., 2016).

We continued our studies of *Paracoccidioides* spp. and sequenced, assembled and annotated 31 isolates representing the phylogenetic, geographic, and ecological breadth of the genus. These samples included clinical, environmental and laboratory reference strains of the S1, PS2, PS3, and PS4 lineages of *P. brasiliensis* and also isolates of the species *Paracoccidioides lutzii.* We completed the first annotated genome assemblies for the PS3 and PS4 lineages and found that gene order was highly conserved across the major lineages, with only a few chromosomal rearrangements. Our analyses provided insight into the recent evolutionary events highlighting genetic differences between the lineages that could impact the distribution, pathogenicity, and ecology of *Paracoccidioides*. We published the work titled *"Genome Diversity, Recombination, and Virulence across the Major Lineages of Paracoccidioides"* in mSphere in 2016, found in chapter 8 (Muñoz et al., 2016).

The previous publications marked a milestone in our research as we felt confident about the experience we had gained in fungal genomics and considered a clinical application of this knowledge. We chose one of the pathogenic species within this Onygenales clade with high public health impact, *Histoplasma capsulatum,* and revisited the genomic sequences utilized in molecular detection of the fungus. Although the genome of *H. capsulatum* has not been updated, using all available sequences allowed us to design primer pairs with high confidence. Specifically, our

aim was to generate a set of sequence fragments that are unique to diverse *H. capsulatum* strains that can be utilized to detect the presence of *H. capsulatum*.

In order to re-analyze all of the available sequences of *H. capsulatum*, we designed and implemented an algorithm that takes any given sequence(s) and splits the sequence into fragments in order to query the unique percentage of the fragment against a group of sequences that are closely related to the query as well as some outgroups, including human. In this case, we compared *H. capsulatum* against other Onygenales fungal species. This algorithm has not been published and is currently used as an in-house script for the design of novel molecular detection assays by our laboratory. The algorithm and its description can be found in as an appendix to chapter 9.

As a proof of principle, we used our algorithm, as well as other bioinformatic methods, for the design of primer pairs used for the molecular identification for *H. capsulatum*. We analytically validated the primer pairs designed using isolated DNA from 62 *H. capsulatum* strains. We obtained promising results with 100% analytical specificity and 100% analytical sensitivity. The work titled "*Design and standardization of a conventional and a real time PCR assay based on novel species-specific genomic regions of the fungal pathogen Histoplasma capsulatum"* has been submitted to PLOS ONE and can be found in chapter 9.

In parallel, and utilizing the same protocols, we also designed and validated primer pairs used for the molecular identification of *Paracoccidioides* spp*.,* fungal pathogens that cause paracoccidioidomycosis, which is endemic in Latin America. We

obtained promising results with 100% analytical specificity and 100% analytical sensitivity. The work titled "*Design and standardization of a conventional and a real time PCR assay based on novel genus-specific genomic regions of the fungal pathogens Paracoccidioides* spp.*"* is in the last stage of preparation and will be submitted for publication, and is included as a draft manuscript in chapter 10.

This concludes my work in the area of fungal genomics. The work on the application of the algorithm and the design of potential diagnostic assays was a result of my international internship as a Fulbright Fellow at Georgia Institute of Technology and the Centers for Disease Control and Prevention (CDC). The work done in the design and implementation of the algorithm was appraised by international referees who awarded me the recognition of *MIT Technological Reviews Innovator Under 35 Colombia 2015* as well as *BBC World: BBC: Young Latin Americans revolutionizing science and medicine with their inventions 2016*.

1.3 Human cardiovascular genomics

1.3.1 Overview

Human cardiovascular genomic studies are of high importance worldwide due to the high mortality rates caused by cardiovascular diseases observed in low, middle and high income countries. Several publications have demonstrated that a specific region of the human genome, the 9p21.3 locus, is highly involved in cardiovascular disease, inflammation and cancer. At the molecular and genetic levels, the six major disease groups, which all affect the higher age groups within and beyond Colombia are: (1) diseases of heart, (2) malignant neoplasms, (3) chronic lower respiratory diseases, (4)

cerebrovascular diseases, (5) Alzheimer's disease, and (6) diabetes mellitus (in particular, type 2). Insights into these diseases at the molecular and genetic levels are likely to also help the understanding of metabolic syndrome, and of other constellations of abnormalities that are found linked to it via associations. We studied a 100 kilobase (kb) region in the human genome, together with its immediately surrounding regions, which has consistently appeared, independently and in original studies done by different research groups, as being strongly associated with two or more of the above 6 disease types and/or their established precursor signs/markers. The 9p21.3 locus spans a region next to and including the cyclin-dependent kinase inhibitor gene *CDKN2A/B*. This region, called *ANRIL*, corresponds to a long non coding antisense RNA of the *CDKN2A/B* gene. It is one of the most interesting, repeatedly 'rediscovered' and now frequently discussed loci reported for cardiovascular disease association (Paynter et al., 2009, Stefansson et al., 2009, Jarinova et al., 2009, Pasmant et al., 2011, Scheffold et al., 2011). It has also appeared in lists of loci preferentially associated with malignant neoplasms (Sherborne et al., 2010, Lucioni et al., 2011, Savola et al., 2007), type 2 diabetes mellitus (Silander et al., 2009, Cheng et al., 2011) and late-onset Alzheimer's disease (Zuchner et al., 2008). The gene *CDKN2A*, which is within the locus, is also of interest for longevity studies, because of its apparent role in aging-related processes, notably tumor suppression and/or cellular senescence (Halaschek-Wiener et al., 2009).

We took on the task of genotyping 67 SNPs within this region that had previously been reported as highly associated with coronary heart disease and type 2

diabetes for a cohort of 394 healthy individuals. These studies are important for understanding risk associated with disease, highly important as public health indicators of the population. Nevertheless, studies in affected populations provide insights of the disease itself. Human cardiovascular genomics also has an important application as a diagnostic aid for patients with cardiovascular diseases that may have an inherited component.

1.2.1 Presentation of the human cardiovascular genomics chapters

We began our work on human cardiovascular genomics with a study involving genotyping of 67 selected SNPs within the 9p21.3 locus of the human genome motivated by its proven association to cardiovascular disease. In collaboration with a prestigious Cardiology group in Medellin, we focused on a Colombian cohort of 357 healthy individuals with data collected for 44 basic phenotypic traits and of which 181 of those individuals had additional data on 157 detailed hemodynamic and other phenotypic traits. The phenotypic categories were mainly linked to cardiovascular disease. Preliminary results using population genomics algorithms indicated consistent associations of several SNPs in this locus with the phenotypic category of blood pressure. These studies are still underway and the final analysis has not yet been published. We recently published intermediate results applying our *k*-SNP analysis methods to another locus (12q24), titled *"Towards multiple-SNP motif analyses of loci associated with phenotypic traits"* in the Journal of American College of Cardiology in 2017, found in chapter 11.

The real-world implementation of human genomics in the clinical setting is slowly making its way into Colombia. As part of the experience I gained throughout my doctoral studies, I was given the opportunity to be the founder and scientific director of a clinical genomics laboratory, GenomaCES at Clínica CES, a university hospital of Universidad CES. I have applied the bioinformatics and genomics pipelines we implemented early in my doctoral studies for clinical application on human samples of various chronic diseases. The implementation of whole exome sequencing has allowed me to effectively diagnose mutations, insertions and deletions within the coding sequences of patients with genetic conditions such as cardiomyopathy, innate metabolism diseases, neuromuscular pathologies and various cancers. We concluded that use of exome sequencing has a high impact on cost-effective diagnosis and lowers cost to the national health system when compared to traditional single gene sequencing. We present some of these results in the publication titled *"Current state of cardiovascular genomics in Colombia"* in Revista Colombiana de Cardiologia in 2016, found in chapter 12 (Gallo et al., 2017).

As a proof of principle, we sequenced the exome of a patient with familiar hypercholesterolemia. The objective was to determine which genetic variations underlie this inherited cardiovascular pathology. Our findings helped the cardiology group focus the patients' treatment using an alternative next generation cholesterol lowering drug which showed promising results in treatment. The work titled *"Hipercolesterolemia familiar heterocigota en manejo con anti-PCSK9"* was accepted in Revista Colombiana de Cardiologia in 2017, found in chapter 13.

References

1. Muñoz JF, Gallo JE, Misas E, McEwen JG, Clay OK. The eukaryotic genome, its reads, and the unfinished assembly. FEBS Lett. 2013 Jul 11;587(14):2090–3.
2. Muñoz JF, Misas E, Gallo JE, McEwen JG, Clay OK. Limits to Sequencing and de novo Assembly: Classic Benchmark Sequences for Optimizing Fungal NGS Designs. In: Castillo LF, Cristancho M, Isaza G, Pinz n AS, Rodr guez JMC, editors. Advances in Computational Biology. Cham: Springer International Publishing; 2014. pp. 221–30. (Advances in Intelligent Systems and Computing; vol. 232).
3. Gallo JE, Muñoz JF, Misas E, McEwen JG, Clay OK. The complex task of choosing a de novo assembly: lessons from fungal genomes. Comput Biol Chem. 2014 Dec;53 Pt A:97–107.
4. Muñoz JF, Gauthier GM, Desjardins CA, Gallo JE, Holder J, Sullivan TD, et al. The Dynamic Genome and Transcriptome of the Human Fungal Pathogen Blastomyces and Close Relative Emmonsia. Haridas S, editor. PLoS Genet. 2015 Oct;11(10):e1005493.
5. Misas E, Muñoz JF, Gallo JE, McEwen JG, Clay OK. From NGS assembly challenges to instability of fungal mitochondrial genomes: A case study in genome complexity. Comput Biol Chem. 2016 Apr;61:258–69.
6. Muñoz JF, Farrer RA, Desjardins CA, Gallo JE, Sykes S, Sakthikumar S, et al. Genome Diversity, Recombination, and Virulence across the Major

Lineages of Paracoccidioides. Mitchell AP, editor. mSphere. American Society for Microbiology Journals; 2016 Sep;1(5):e00213–6.

7. Paynter NP, Chasman DI, Buring JE, Shiffman D, Cook NR, Ridker PM. Cardiovascular disease risk prediction with and without knowledge of genetic variation at chromosome 9p21.3. Ann Intern Med. NIH Public Access; 2009 Jan 20;150(2):65–72.
8. Stefansson H, Ophoff RA, Steinberg S, Andreassen OA, Cichon S, Rujescu D, et al. Common variants conferring risk of schizophrenia. Nature. 2009 Aug 6;460(7256):744–7.
9. Jarinova O, Stewart AFR, Roberts R, Wells G, Lau P, Naing T, et al. Functional analysis of the chromosome 9p21.3 coronary artery disease risk locus. Arterioscler Thromb Vasc Biol. American Heart Association, Inc; 2009 Oct;29(10):1671–7.
10. Pasmant E, Sabbagh A, Vidaud M, Bièche I. ANRIL, a long, noncoding RNA, is an unexpected major hotspot in GWAS. FASEB J. Federation of American Societies for Experimental Biology; 2011 Feb;25(2):444–8.
11. Scheffold T, Waldmüller S, Borisov K. A case of familial hypertrophic cardiomyopathy emphasizes the importance of parallel screening of multiple disease genes. Clin Res Cardiol. Springer-Verlag; 2011 Jul;100(7):627–8.
12. Sherborne AL, Hosking FJ, Prasad RB, Kumar R, Koehler R, Vijayakrishnan J, et al. Variation in CDKN2A at 9p21.3 influences

childhood acute lymphoblastic leukemia risk. Nat Genet. 2010 Jun;42(6):492–4.

13. Lucioni M, Novara F, Fiandrino G, Riboni R, Fanoni D, Arra M, et al. Twenty-one cases of blastic plasmacytoid dendritic cell neoplasm: focus on biallelic locus 9p21.3 deletion. Blood. American Society of Hematology; 2011 Oct 27;118(17):4591–4.

14. Savola S, Nardi F, Scotlandi K, Picci P, Knuutila S. Microdeletions in 9p21.3 induce false negative results in CDKN2A FISH analysis of Ewing sarcoma. Cytogenet Genome Res. 2007;119(1-2):21–6.

15. Silander K, Tang H, Myles S, Jakkula E, Timpson NJ, Cavalli-Sforza L, et al. Worldwide patterns of haplotype diversity at 9p21.3, a locus associated with type 2 diabetes and coronary heart disease. Genome Med. BioMed Central; 2009 May 12;1(5):51.

16. Cheng X, Shi L, Nie S, Wang F, Li X, Xu C, et al. The same chromosome 9p21.3 locus is associated with type 2 diabetes and coronary artery disease in a Chinese Han population. Diabetes. 2011 Feb;60(2):680–4.

17. Züchner S, Gilbert JR, Martin ER, Leon-Guerrero CR, Xu P-T, Browning C, et al. Linkage and association study of late-onset Alzheimer disease families linked to 9p21.3. Ann Hum Genet. Blackwell Publishing Ltd; 2008 Nov;72(Pt 6):725–31.

18. Halaschek-Wiener J, Amirabbasi-Beik M, Monfared N, Pieczyk M, Sailer C, Kollar A, et al. Genetic variation in healthy oldest-old. Mary Bridger J, editor. PLoS ONE. 2009 Aug 14;4(8):e6641.

19. Gallo JE. Current state of cardiovascular genomics in Colombia. Revista Colombiana de Cardiologia. 2017 Jan;24(1):e1–e2.

Chapter 2

The eukaryotic genome, its reads and the unfinished assembly

FEBS Letters 587 (2013) 2090–2093

ELSEVIER

FEBS Letters

journal homepage: www.FEBSLetters.org

Hypothesis

The eukaryotic genome, its reads, and the unfinished assembly

José Fernando Muñoz [a,b], Juan Esteban Gallo [a,c], Elizabeth Misas [a,b], Juan Guillermo McEwen [a,d], Oliver Keatinge Clay [a,e,*]

[a] Cellular & Molecular Biology Unit, Corporación para Investigaciones Biológicas, Medellín, Colombia
[b] Institute of Biology, Universidad de Antioquia, Medellín, Colombia
[c] Doctoral Program in Biomedical Sciences, Universidad del Rosario, Bogotá, Colombia
[d] School of Medicine, Universidad de Antioquia, Medellín, Colombia
[e] School of Medicine and Health Sciences, Universidad del Rosario, Bogotá, Colombia

ARTICLE INFO

Article history:
Received 9 February 2013
Revised 9 May 2013
Accepted 20 May 2013
Available online 30 May 2013

Edited by Takashi Gojobori

Keywords:
Next generation sequencing
Eukaryotic genomics
Assembly-free genome analysis
Object-view separation
Microbial strain collection

ABSTRACT

In recent years, readily affordable short read sequences provided by next-generation sequencing (NGS) have become longer and more accurate. This has led to a jump in interest in the utility of NGS-only approaches for exploring eukaryotic genomes. The concept of a static, 'finished' genome assembly, which still appears to be a faraway goal for many eukaryotes, is yielding to new paradigms. We here motivate an object-view concept where the raw reads are the main, fixed object, and assemblies with their annotations take a role of dynamically changing and modifiable views of that object.

1. Introduction

Advances in next-generation sequencing (NGS) technology in recent years have increased length and accuracy of short read sequences that are produced as primary sequence data. A few years ago, Illumina/Solexa reads, which typically allow good coverage at affordable cost, still measured only some 36 base pairs (bp), of which the last 6 bp or so were often of poor quality. Now, Illumina read lengths are typically at least 100 bp, and the quality is often excellent throughout.

The one-pass, automated nature of current sequencing workflows runs allows NGS to reliably deliver a single set of text or binary files that contain the full set of fixed-length reads or read-pairs for a genome of interest. Such stand-alone or modular output is attractive, and many groups now deposit their primary read data in short read archives at NCBI or the European Nucleotide Archive for others to use. In an NGS project, the standardized output files, in FASTQ or equivalent format, arrive from the sequencer ready for quality control, assembly and then annotation.

As NGS technologies advance, the way we think of the primary output from a sequencer is changing, and the time may have come to reassess a way of looking at sequencing processes that we have retained from past decades.

In the past, the gap-free, 'finished' assembly (which may still be a utopia for many genomes, even for the human genome, in spite of its paramount importance for human health) was seen as a prime object or trophy. The hypothesis we explore here is that until such a goal comes closer, it might help us to think more clearly, pragmatically, and phenomenologically about NGS if we de-emphasize the goal of a static, 'best' assembly and consider, instead, the initial read set as the primary and reliable object. Possible assemblies, with their respective annotations, would then become dynamic, modifiable views of that primary object or 'observable', although at any given time a single state-of-the-art assembly could serve as a reference. Our working hypothesis is that shifting the object-view boundary in this way could bring advantages, if short-read approaches remain a stable norm.

In most of the following considerations we will keep individual, previously unsequenced, unicellular fungi in mind as conceptual test genomes. Unicellular fungi are intermediate, in genome size

Abbreviations: NGS, next generation sequencing; GC, guanine and cytosine level

* Corresponding author at: Cellular & Molecular Biology Unit, Corporación para Investigaciones Biológicas, Medellín, Colombia.
E-mail address: oliver.clay@gmail.com (O.K. Clay).

0014-5793/$36.00
http://dx.doi.org/10.1016/j.febslet.2013.05.048

and complexity, between the prokaryotes (bacteria and archaea) and much larger, metazoan eukaryotes such as human, so their NGS reads do not need massive storage (as do, for example, some extensive human population resequencing studies or large-scale metagenomics projects). Many unicellular fungi are of wide interest, either because they are pathogenic to human, other animals, or plants, or because they serve as model fungi or close relatives of well-characterized fungi. Many unicellular fungi can now be sequenced at good coverage in a single lane (or even a fraction of a lane) of an Illumina sequencer, especially if one is interested mainly in the genes. The storage space that is occupied by the primary NGS reads can hardly be considered expensive, and in future it will presumably become cheaper.

2. Uncurated de novo assemblies can lose information

The strict de novo genome assembly problem belongs to a class of combinatorial inverse problems exemplified by Humpty Dumpty's rhyme.[1] A strictly de novo assembly of a genome from an NGS read set (for single or paired reads) can never contain more information than is present in that original read set. Automatic de novo assembling without human supervision or curation will never create new sequence-specific information, although it may skillfully extract or infer information present in the reads, and it may lose information that was originally present in the raw reads.

First, contiguity information is usually lost when a eukaryotic genome is chopped into small pieces. The smaller the pieces, or the higher the repetitiveness, the worse is the loss. In some genomes, only parts of the original genome can be reliably assembled from the short pieces de novo, because the short pieces' sequences are not all unique in the actual genome, so their sequence context cannot be reconstructed [2–4]. Already the 86 kb, circular mitochondrial genome of baker's yeast [5], which contains substantial repeats and low complexity regions, provides a good example of this difficulty for NGS-only approaches.

Second, most assembly programs, such as Velvet [6] or SOAPdenovo with GapCloser [7], must make some evidence-based decisions. Such decisions may leave no trace of their risk or location in the resulting assembly when it is released, so if the decision was wrong, further information may be lost.

Third, not only do typical NGS assemblies released to the public omit the quality information provided for each read, but the local coverage by reads, another measure of confidence (number of reads covering a given position, and degree of agreement among those reads' sequences), is also missing from the resulting contigs or scaffolds. Indeed, with current NGS technology, most loci in a genome are likely to be covered by many reads, if they are correctly mapped. Especially where guanine and cytosine level (GC) is neither very high nor very low, the reads' coverage depth (coverage profile), quality and consistency contain information. To give one example: such local read alignments can help one to detect, a posteriori, where an unsupervised assembly might have erroneously collapsed nearly identical paralogs (as exist in some mammalian interferons; [8]) onto a single composite 'gene' that does not exist.

3. The idea of a final assembly is often a utopia

For reasons including those mentioned above for de novo assembly, the long, assembled and annotated chromosomal sequences (contigs or scaffolds) that are obtained for eukaryotes are usually hypotheses, not facts. This conclusion applies also where the assembly is not a de novo assembly but a reference assembly, because reference assemblies inherit mistakes from the genome sequence(s) to which they refer. Indeed, reference assemblies or annotations are ultimately based, possibly via a chain of recursive referencing, on some 'first' reference genome that was assembled or annotated de novo. Assembly and annotation errors can propagate along such a chain, especially when one does not interleave reference strategies with de novo strategies. 'Snowball effects' of this kind can be a problem not only for reference assembly or when finding genes [9,10], but also when assigning functions to genes via reference using programs such as Blast2GO.

Even some of the most important eukaryotic genomes' assemblies remain unfinished, and the goal of a perfect, final genome sequence is likely to remain a utopia for many eukaryotic species in the near future. This is mainly because of repetitive non-protein coding sequences (a genome's protein-coding exons are often well covered by NGS-derived contigs [2]). Thus, the public human genome assembly, which was formally declared "finished" in its euchromatic parts in 2004 [11], is actually still incomplete: the hg19 sequence continues to lack large expanses of heterochromatin, as well as some euchromatic regions such as the (repetitive) ribosomal DNA on chromosomes 13, 14, 15, 21 and 22.

It is likely that several existing annotated assemblies of eukaryotic genomes will be updated again at some future time, as users discover inconsistencies between assembled and annotated sequences of related species/strains, succeed in assembling previously missing regions, re-curate automated output from gene callers, or sequence transcripts. The influx of information on individual genes coming from molecular biology experiments will continue, so it is to be expected that the best assembly will ultimately incorporate them and thus continue to change.

Much as one can now print books on demand and reduce the need for archiving, one can in principle perform automatic assemblies or annotations of genomes on demand. Algorithms for de novo or reference assembly and annotation continue to evolve and improve, together with the databases they access and the hardware they use. As a result, assemblies and their annotations are likely to become increasingly replaceable, transitory, quick to alter, automated, and cheap to repeat or remaster.

We propose that it is natural to consider the read set as the master reference, template or object, from which assemblies and their respective annotations are generated as dynamic, modifiable and refreshable 'views'. Variants of the *object-view* (or *thing-view*) metaphor are commonly used in software design, where it is good practice to clearly separate a basic 'thing' or its model from possible views of it, both conceptually and when coding [12,13]. Clearly separating out the true observable is also a necessary practice in quantum mechanics, where an often-followed protocol is "'what is observed, certainly exists; about what is not observed we are still free to make suitable assumptions.' This freedom then is used to avoid paradoxes" [14].

What is firm, and possibly irreproducible at a later time, is the set of text or binary files containing the primary read sequences and their quality tracks. This set of files encodes an experiment and a DNA sample, captured in a momentary snapshot of an individual organism at a particular time in the evolutionary history of the strain or population to which the organism belongs. In such a read file, the nucleotide sequences are delivered together with the quality symbol for each nucleotide. For practical purposes, such as browsing or searching by users, annotation database organizing and consistent communication among researchers, the read set should at any given time be accompanied by a single state-of-the-art assembly chosen as a reference and with a version number, as is the case for human genome releases, but with the understanding

[1] "Humpty Dumpty sat on a wall / Humpty Dumpty had a great fall / All the king's horses and all the king's men / couldn't put Humpty together again" [1]. Humpty Dumpty is often depicted as an egg with a face, hands and feet; the problem is to piece the egg back together from its fragments.

that this is one assembly view of the reads and is likely to be replaced in future.

4. A read set is a precious object

When NCBI came close to permanently closing its Sequence Read Archive for new submissions at the beginning of 2011, it became clear to many how important it is to keep a central service for receiving and carefully maintaining read data. This was an important issue: still today, major genome sequencing centers do not all offer, for public downloading, the original read sets they used for the assemblies on their web servers.

Some researchers, however, voiced the opinion that everything can simply be resequenced later, consequently reads are not precious (and they take up gigabytes of storage), so they can be discarded. A related opinion would be that the reads are not and should not be the object; perhaps a fungal strain that has been assigned a strain number in a strain collection might deserve that role, but not its reads. Although such stances might seem reasonable at first sight, there are three reasons to think otherwise.

The first reason is reproducibility. If NGS reads are used as a foundation for building an assembly and annotation, and insights and findings are in turn built on top of those and published, the basic principles of scientific conduct dictate that the reads will remain of vital importance to check reproducibility. If ever an inquiry should be needed later because someone notices a strange result, was it the processing of the reads that was strange, or was something wrong with the reads themselves? If the reads were discarded or misplaced, there is no way to solve this problem.

The second reason is a practical one. From a purely project-management perspective, consider the actual ordering and obtaining of samples, extracting of DNA, preparing of insert libraries, and waiting (sometimes for months) in queues for time on sequencers that are shared by an institution or community, together with the actual cost of sequencing. In addition, one must invest human time, attention and insistence in order to make sure everything is done well. Contrast that bill with the simple running of a re-assembly or re-annotation task in background mode on one's own server during a weekend, possibly using a more recent assembly or annotation program.

A third reason for keeping read sets comes from an evolutionary or identification perspective. The metaphor of a unique, time-stamped snapshot is justified because populations, even strains, get lost or change. Microbiologists working in microbial identification who re-order (or follow over time) a strain of a microbe from a strain collection may occasionally notice a change in phenotypic properties (assuming no strains were mixed up, which also sometimes happens). When collected microbes or cell lines are followed in time, genes that are no longer used, or are no longer under their previous selection pressure, can show expression anomalies or corresponding epigenetic changes in methylation patterns or chromatin configuration [15]. After many generations, such changes can in turn lead to changes in the observable genome sequence, for example when a gene mutates without negative consequences for the cell's survival or replication.

We mentioned in the Introduction that, in recent years, the usable or effective read length one can expect from readily affordable NGS has approximately trebled, from less than 36 bp to over 100 bp. This change, although it may seem a modest step, has brought clear advantages for both the ease of de novo assembly and the ease of locating the individual reads on a conspecific or related reference genome. Although for some genomes a usable read length of 30 bp may suffice to obtain long contigs, i.e., assemblies with high N50 values, there are other genomes in which assembly quality or reliability increases very noticeably when one increases read lengths to 100 bp. A quantitative analysis comparing read lengths and their effects on assembly quality in selected genomes is presented in Ref. [3]. The use of paired-end reads, separated via an insert library by a fairly fixed distance of a few hundred base pairs, then further improves reliability. Indeed, a read in a small repetitive region has a better chance of being disambiguated by its mate: even if a read is lost in the repeats, its mate standing on firm, unique DNA some distance away can in principle localize both. An example of a fungal genome paper in which assembly results are shown first after using paired-end 36 bp Solexa/Illumina reads, and then again after including also longer, single-end 454 reads, is the paper describing the *Sordaria macrospora* genome project [16]. Trebling the effective sequence length from around 30 bp to 100 bp can, similarly, facilitate the assignment of an individual read to its position in an external reference assembly. Such considerations strengthen the notion that NGS read sets have now become precious objects in their own right.

5. Assembly-free uses of reads

Dedicated assembly/annotation projects that include human curation components have enormous merit. For the progress of genome biology, it is crucial that they continue to receive decent funding. We are still far from being able to replace, by any unsupervised pipeline or view, the dedicated human curating, expert decision-making, quality honing, and careful resolution of biological inconsistencies that are part of a serious genome assembly and annotation project. However, experiences made since the advent of NGS no longer sustain the opinion that, prior to human curation, it need be "the initial alignment or assembly that determines whether an experiment has succeeded and provides a first glimpse into the results" [17]. First glimpses can also be obtained from the reads without an assembly. One can now, for example, directly view or search the raw read data for a task at hand, quickly create an ad hoc, local assembly of reads around a guide or test gene of interest, or compare sites where there are single-nucleotide polymorphisms (SNPs) within a population.

Biological analyses can be done in principle, and sometimes also in practice, using the raw reads directly, bypassing global assemblies and/or annotation. This statement is more obvious than it may seem. Today, many global assemblies and annotations are almost entirely automated. The products of such software runs can therefore be represented, or conceptually replaced, by the processes themselves, which can be piped and optimized. In other words, there is no conceptual need for a 'thing' or intermediate product called an assembly or an annotation. This simple observation, and its potential for exploiting when one designs algorithms or combinatorial methods, has not received much attention in the literature. An exception has been the research of Peterlongo and his colleagues on pre-assembly or assembly-free, direct analysis of NGS reads. They have written and presented dedicated, efficient proof-of-concept programs for local or targeted assemblies, SNP calling and other biological analyses that do not require prior whole-genome assembly or annotation [18,19]. As one of their presentations aptly states in its title: "Biological information is in the reads" [20].

Some basic and useful direct analysis or data extraction procedures ('pedestrian' tasks) can sometimes be quite easily and tractably performed on commodity hardware using familiar general-purpose programs such as BLAST or BLAT and/or basic Linux/Unix commands, although newer, dedicated NGS programs such as BWA [21] or Bowtie 2 [22] have advantages for pipelines. An example is a similarity search for a gene of interest against 50 million read

pairs of a fungal genome. Such checks can be extremely valuable when one cannot find the ortholog of a known gene in a newly obtained genome assembly, and wants a reliable proof of its absence in the actual genome.

The possibilities of directly using reads for analysis provide an opportunity to reflect on the timeliness of a historic linear pipeline topology, still widely used as a paradigm when designing genome projects. Its direction goes from sequencing through assembly and annotation to analysis and then usually to publication and project termination, but typically not back again to reassembly or re-annotation unless there is a formal follow-up project. Much as the waterfall-Gantt model of project management, which corporate and other organizations have used as a guideline for decades [23,24], and with parallels to the Central Dogma of molecular biology [25], current genome hosting often does not anticipate how feedback from outside scientific communities could be efficiently integrated after the end of a genome project, when user communities wish to suggest further changes or corrections to genome assemblies or annotations on a routine, ongoing basis. A wish might be to frequently refresh assembly views or their annotations under a set of constraints representing user-supplied knowledge items. Although it is not yet clear how this would be implemented, user feedback would be treated as a certain event, or even actively solicited, and corresponding checkpoints would be hardwired into the plan (Supplementary Material, Box S1). In other words, the views could be freed to change dynamically while the central object, the original genome snapshot being viewed, stays accessible in the reads.

6. Conclusion

In this contribution, we motivate an unconventional way of envisaging genomics processes, which has helped us to clarify and improve our own conceptual workflows in a fungal genomics lab. Although individual points mentioned here have been raised or discussed informally by others in conferences or on web sites, we have seen few previous publications (which we cite) that were dedicated to centrally addressing them and outlining a coherent perspective for a broad readership. It is difficult to predict for how long short-read sequencing technology will stay the main approach, and it is clear that many of the considerations presented here would need to be changed if much longer reads (e.g., along the lines anticipated for Oxford Nanopore sequencing [4,26]) become the popular choice. Until then, we hope that clear thinking along the lines we sketch here will stimulate conceptual and practical advances in genomics and genome informatics.

Acknowledgement

The present paper evolved from the authors' thoughts and input during early NGS assemblies, annotations and analyses for the project "Comparative genomics and virulence in the pathogenic fungus *Paracoccidioides brasiliensis*", supported by Colciencias grant 2213-48925460. We thank John W. Taylor and Emily A. Whiston (University of California, Berkeley) for high-quality insert library preparation, sequencing and conceptual work that motivated some of the ideas presented here, and for discussions. We also thank two anonymous reviewers for careful reading of the manuscript and for constructive criticism.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.febslet.2013.05.048.

References

[1] Opie, I. and Opie, P. (1997) The Oxford Dictionary of Nursery Rhymes, Oxford University Press, Oxford. 2nd ed., pp. 213–215.
[2] Kingsford, C., Schatz, M.C. and Pop, M. (2010) Assembly complexity of prokaryotic genomes using short reads. BMC Bioinformatics 11, 21.
[3] Whiteford, N., Haslam, N., Weber, G., Prügel-Bennett, A., et al. (2005) An analysis of the feasibility of short read sequencing. Nucleic Acids Res. 33, e171.
[4] Nagarjan, N. and Pop, M. (2013) Sequence assembly demystified. Nat. Rev. Genet. 14, 157–167.
[5] Foury, F., Roganti, T., Lecrenier, N. and Purnelle, B. (1998) The complete sequence of the mitochondrial genome of *Saccharomyces cerevisiae*. FEBS Lett. 440, 325–331.
[6] Zerbino, D.R. and Birney, E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18, 821–829.
[7] Luo, R., Liu, B., Xie, Y., Li, Z., et al. (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience 1, 18.
[8] Henco, K., Brosius, J., Fujisawa, A., Fujisawa, J.-I., et al. (1985) Structural relationship of human interferon α genes and pseudogenes. J. Mol. Biol. 185, 227–260.
[9] Jabbari, K., Cruveiller, S., Clay, O., Saux, J.L. and Bernardi, G. (2004) The new genes of rice: a closer look. Trends Plant Sci. 9, 281–285.
[10] Cruveiller, S., Jabbari, K., Clay, O. and Bernardi, G. (2003) Compositional features of vertebrate genomes for checking predicted genes. Brief. Bioinform. 4, 43–52.
[11] IHGSC (International Human Genome Sequencing Consortium) (2004) Finishing the euchromatic sequence of the human genome. Nature 431, 931–945.
[12] Reenskaug, T. (1979) Thing-Model-View-Editor – an example from a planningsystem. Technical Report 1979-05-MVC, Xerox PARC, <http://heim.ifi.uio.no/trygver/1979/mvc-1/1979-05-MVC.pdf>.
[13] King, T., Reese, G., Yarger, R. and Williams, H.E. (2002) Managing and Using MySQL, O'Reilly Media, Sebastopol, CA. 2nd ed..
[14] von Weizsacker, C.F. (1971) The Copenhagen InterpretationQuantum Theory and Beyond: Essays and Discussions arising from a Colloquium, pp. 25–32, Cambridge University Press.
[15] Antequera, F., Boyes, J. and Bird, A. (1990) High levels of *de novo* methylation and altered chromatin structure at CpG islands in cell lines. Cell 62, 503–514.
[16] Nowrousian, M., Stajich, J.E., Chu, M., Engh, I., et al. (2010) *De novo* assembly of a 40 Mb eukaryotic genome from short sequence reads: *Sordaria macrospora*, a model organism for fungal morphogenesis. PLoS Genet. 6, e1000891.
[17] Flicek, P. and Birney, E. (2009) Sense from sequence reads: methods for alignment and assembly. Nat. Methods 6, S6–S12.
[18] Peterlongo, P., Schnel, N., Pisanti, N., Sagot, M.-F. and Lacroix, V. (2010) Identifying SNPs without a reference genome by comparing raw reads. Lecture Notes in Computer Science, 6393 (String Processing and Information Retrieval), 147–158.
[19] Peterlongo, P. and Chikhi, R. (2012) Mapsembler, targeted and micro assembly of large NGS datasets on a desktop computer. Bioinformatics 13, 48.
[20] Peterlongo, P. (2011) Biological information is in the reads. Bioinformatics and High Throughput Sequencing, Institut Pasteur, Paris, France. < http://www.lirmm.fr/~rivals/HTS-2011/RES/Peterlongo-abs.html>.
[21] Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760.
[22] Langmead, B. and Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359.
[23] Brooks, F.P. (1995) The Mythical Man-Month: Essays on Software Engineering, Addison-Wesley, Reading, MA. with four new chapters, Anniversary edition, pp. 264 ff.
[24] Chromatic (2003) Extreme Programming Pocket Guide, O'Reilly Media, Sebastopol, MA.
[25] Crick, F. (1970) Central dogma of molecular biology. Nature 227, 561–563.
[26] Loman, N.J., Constantinidou, C., Chan, J.Z.M., Halachev, M., et al. (2013) High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. Nat. Rev. Microbiol. 10, 599–606.

Chapter 3

Limits to sequencing and *de novo* assembly: Classic benchmark sequences for optimizing fungal NGS designs

Limits to Sequencing and *de novo* Assembly: Classic Benchmark Sequences for Optimizing Fungal NGS Designs

José Fernando Muñoz[1,2], Elizabeth Misas[1,2], Juan Esteban Gallo[1,3], Juan Guillermo McEwen[1,4], and Oliver Keatinge Clay[1,5, ⋆]

[1] Cellular and Molecular Biology Unit, Corporación para Investigaciones Biológicas, Medellín, Colombia
[2] Institute of Biology, Universidad de Antioquia, Medellín, Colombia
[3] Doctoral Program in Biomedical Sciences, Universidad del Rosario, Bogotá, Colombia
[4] School of Medicine, Universidad de Antioquia, Medellín, Colombia
[5] School of Medicine and Health Sciences, Universidad del Rosario, Bogotá, Colombia
{josejfmg,Elizabeth.misas,galloucf,oliver.clay}@gmail.com, mcewen@une.net.co

Abstract. Planning of pipelines for next-generation sequencing (NGS) projects could be facilitated by using simple DNA sequence benchmarks, i.e., standard test sequences that could monitor or help to predict ease or difficulty of (a) short-read sequencing and (b) *de novo* assembly of the sequenced reads. We propose that familiar, gene-sized sequences, including but not limited to nuclear protein-coding genes, would provide feasible consensus benchmarks allowing simple visualization. We illustrate our proposal for fungi with candidates from ribosomal DNA (rDNA, used in phylogeny and identification/diagnostics), mitochondrial DNA (mtDNA), and combinatorially constructed conceptual (synthetic) DNA sequences. The exploratory analysis of such familiar candidate loci could be a step toward finding, testing and establishing familiar, biologically interpretable consensus benchmark sequences for fungal and other eukaryotic genomes.

Keywords: Next generation sequencing, Eukaryotic genomes, *De novo* assembly, Benchmarking.

1 Introduction

When one plans pipelines for next-generation sequencing (NGS) projects, it would often be helpful to have available simple reference DNA sequences or benchmarks, i.e., standard or consensus test sequences that could monitor or help to predict ease or difficulty of (a) short-read sequencing and (b) *de novo*

⋆ Corresponding author.

L.F. Castillo et al. (eds.), *Advances in Computational Biology*, Advances in Intelligent Systems and Computing 232,
DOI: 10.1007/978-3-319-01568-2_32,

assembly of the sequenced reads. Indeed, a genome or genomic region can have patches that are inherently refractory to either endeavour.

We consider separately two main groups of tasks in a genomics pipeline leading to an assembly, which lend themselves to separate benchmarking: tasks that are done before the official read set is obtained, and tasks that are done after one has the reads. The former, *sequencing-related* tasks include choice of sequencing technology and choice of parameters such as read length and insert size; the preparation and possible selection of insert libraries; actual sequencing; and any routine censoring or other processing of raw sequencer output that is then performed in order to arrive at a set of official or presentable read files. The latter, *assembly-related* tasks include choice of assembly program(s) and choice of parameters such as k-mer length (for de Bruijn graph-based programs); actual assembling into contigs and/or scaffolds; and any subsequent censoring or elimination of short or otherwise doubtful contigs or scaffolds from the assembly. The tasks listed in these two groups are the ones we can influence, for which we seek guidance, and for which existing and new benchmarks could be valuable. We propose that familiar sequences, of the length of one or a few known genes, could provide useful and realistic external references that would be helpful in such contexts.

If one looks only at whole-genome assemblies of eukaryotes without considering the structural and functional features that exist along the chromosomes, benchmarks already exist. Well-known examples are N50, NG50 or NG90; state-of-the-art benchmarks of this category are listed and used in the recent Assemblathon 2 report [5]. If one restricts one's attention to protein-coding genes of the nuclear genome, benchmarking can be done by assessing how well one's assembly covers conserved or 'core' genes that one expects to find, e.g., using CEGMA [28,29,5]. We propose, in the case of fungi, to complement such existing benchmarks with benchmarks based on ribosomal DNA (rDNA, used in phylogeny and identification/diagnostics), mitochondrial DNA (mtDNA, of which for example the gene for cytochrome c oxidase subunit 1, *CO1*, is used in barcoding), and combinatorially constructed conceptual or synthetic DNA sequences. We present our proposal using selected examples.

2 Materials and Methods

Reference sequences were selected after performing exploratory studies on DNA sequences of model and pathogenic fungi (*Saccharomyces cerevisiae*, *Aspergillus fumigatus*, and dimorphic fungi from the Onygenales order) that are represented by an intrinsic interest in the scientific community and may therefore already be available prior to a strain's whole-genome sequencing. We retrieved real DNA reference sequences from public databases (GenBank, Broad Institute), and created synthetic DNA reference sequences by generating DNA stretches that have no repeats of length $\geq w$. Reads used in the exploratory studies were taken from Illumina paired-end short-read ($l \geq 100$ bp) files downloaded from the NCBI Sequence Read Archive (SRA), from our own Illumina paired-end reads ($l \geq 100$

bp) or, in the case of ideal reads (see [20]), created from longer public sequences by generating all possible subsequences of a fixed length l ($l \geq 100$ bp). Further details of presented examples are given in Figure legends and descriptions in the main text.

3 Results and Discussion

3.1 Benchmarks from Ribosomal DNA for Optimizing Sequencing-Related Tasks

Although ribosomal DNA (rDNA) of a eukaryotic organism can seriously resist assembly because of its tandemly repeated nature (for example, some of the human genome's known rDNA regions are still missing from the hg19 sequence), we focus here on inherent resistance to sequencing and/or related tasks that lead to the production of an official read set.

Generally, patches along a chromosome that are truly refractory to sequencing (thus causing unsatisfactory read depths) can result from various factors, of which one seems to be extreme or unusual GC (guanine-cystosine) levels [33, and refs. therein]. The rDNA loci in eukaryotes, including fungi and also human [16, Genbank accession U13369], can exhibit some of a genome's most severe GC changes and/or extremely high levels of GC; in the interphase nucleus, rDNA does not reside in typical chromosome territories but in nucleoli [1].

Figure 1 shows an example of how a classic rDNA reference sequence of a fungus can be used as a potential benchmark, to monitor patches along the sequence where reads from a closely related organism (here, another strain of the same species) are thinly spread. If the organism represented by the reads is close enough to the organism represented by the reference sequence, and if one chooses a matching method having appropriate sensitivity and specificity (e.g., general-purpose matching programs such as BLAST or BLAT or more dedicated NGS programs such as BWA [22] or Bowtie 2 [21]), then patches of shallow reads must have been caused during sequencing or sequencing-associated steps.

3.2 Benchmarks from Mitochondrial DNA for Optimizing Assembly-Related Tasks

Mitochondrial genomes of some fungi can be surprisingly difficult to assemble *de novo*, despite their small sizes, which are typically less than 100 kb. For example, two groups that used shotgun methods to sequence strains of yeast (*Saccharomyces cerevisiae*) appear to have encountered problems in obtaining satisfactory assemblies of the mitochondrial genome [12,25].

A short sequence of less than 100 kb might, at first sight, seem like child's play to assemble, but it turns out that some far bigger nuclear genomes are routinely and quite successfully assembled and annotated while the same organisms' mitochondral genome is not. Mitochondrial genomes can contain seriously repetitive and/or low complexity DNA interspersed in a number of noncoding

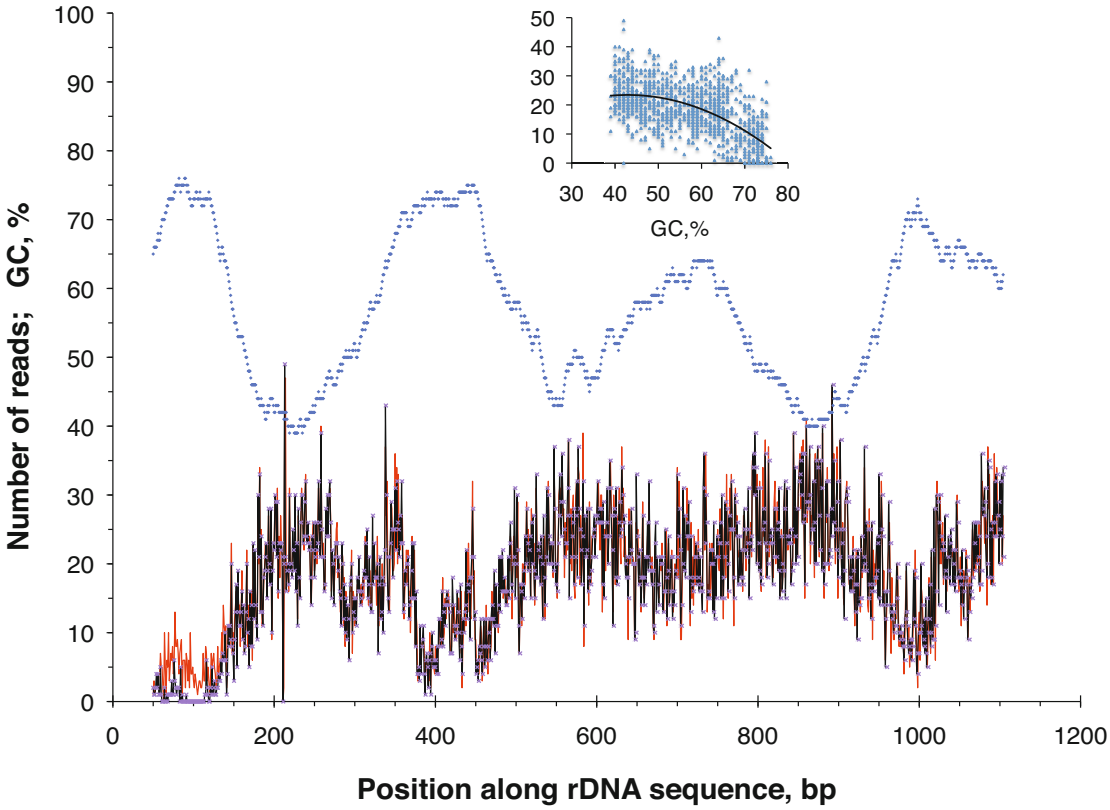


Fig. 1. Positional variation of read counts of *Aspergillus fumigatus* strain A1163 (Illumina GAII paired-end reads of length 101 bp, insert size 300; NCBI SRA accession SRX028559) that match along an rDNA reference sequence. The reference sequence is from *A. fumigatus* strain NRRL 35223 (GenBank accession EF634403, 1154 bp). Matching was done using BLASTN with default settings (black read depth curve). A moving-window GC plot (window 101 bp, step 1 bp) is shown above the match counts, anticorrelating with them. Matching was repeated with the DUST [27] low complexity filter switched off (`-dust no` , light curve superposed on and essentially coinciding with black read depth curve), in order to confirm that the anticorrelation with GC is not an artifact of low-complexity patches. Since the first two valleys of the read coverage plot are found within internal transcribed spacer regions (ITS1: positions 13–195, ITS2: positions 353–521), we also verified that these sequences are highly conserved (close to 100% identity) among *A. fumigatus* strains, i.e., that the valleys are not due to more pronounced ITS sequence divergence as is sometimes seen in higher-level, interspecies alignments [17,18]; we have observed similar fluctuations of read depth when mapping our own Illumina 101 bp reads for an *Emmonsia parva* strain to a reference rDNA sequence for that strain. Positions are midpoint of BLAST match for read match counts and window midpoint for GC. Inset: Scatterplot and quadratic polynomial fit sketching the negative relation between read depth and GC ($R = 0.604$; cf. also a similar general relation observed in [33, Suppl. Figure 3]).

stretches along the sequence, which can confuse assembly programs. In fact, repeats can confuse not only assembly programs, but also (biological) homologous recombination, leading to mitochondrial genome instabilities that can become visible as petite mutant colonies in yeast [3,4,12], so if one is looking for a tough benchmark sequence this may be a good candidate: Figure 2 shows that even an ideal paired-end strategy with no sequencing errors and providing uniform, good coverage by reads cannot bridge the *ori* repeats, and suggests that the resulting (modest) fragmentation may be a fundamental limit [20] that is not dependent on the program used for assembling. This small genome has relatively few genes, of which the *CO1* gene has been proposed for barcoding [31], and it is well-studied and familiar to all biologists. Use of mitochondrial genomes as benchmarks might correspond to benchmarking ideals of a "toughest competitor" [7, p. 10], or to agile practices advocating that one write a test that fails, then be prompted by the failing test to write code that will pass that test ([9, pp. 25–29], [6, Chapters 16–19]).

3.3 Comments on Benchmarks from Nuclear Protein-Coding DNA

Protein-coding genes of the nuclear genome are the most abundant genes of a eukaryotic organism, but they are often less difficult to sequence or assemble than other genomic regions. On the other hand, in some sequencing and assembly projects one may be primarily interested in obtaining the full set of nuclear protein-coding sequences, and much less interested in other regions. A recent reminder that it is important to be clear about one's priorities was given by the results of Assemblathon 2, in which some of the least successful assemblies of vertebrate genomes, as measured by genome-wide metrics, were successful in covering a majority of genes ([5]; see also [20]).

The frequently observed or rediscovered success of simple next-generation sequencing strategies in assembling most protein-coding sequences suggests that the notion of *gene space* is appropriate here [29], a term that was originally used to characterize the dramatic enrichment of genes in a relatively small interval of the GC distribution of genomes such as those of cereal plants [8].

The set of genes that was used to assess coverage by the Assemblathon 2 entries came from a collection of core eukaryotic protein-coding genes (CEGs), proposed by Parra et al. [28,29] as a composite benchmark. Their rigorous filtering protocol yielded 458 genes [28] present in 6 eukaryotic model organisms including human and baker's yeast, of which 248 genes were found to be generally present as single copy genes [29]; the 248 genes were then further divided into 4 groups according to their conservation. Based on mapping of CEGs in 25 previously characterized eukaryotic gen he most conserved group (group 4), consisting of 65 genes, has been proposed as an estimator of the percentage of total protein-coding genes covered by the assembly that should be useful even for highly divergent genomes; a CEGMA mapping package is available for analyzing coverage of arbitrary eukaryotic genomes using CEG genes [29].

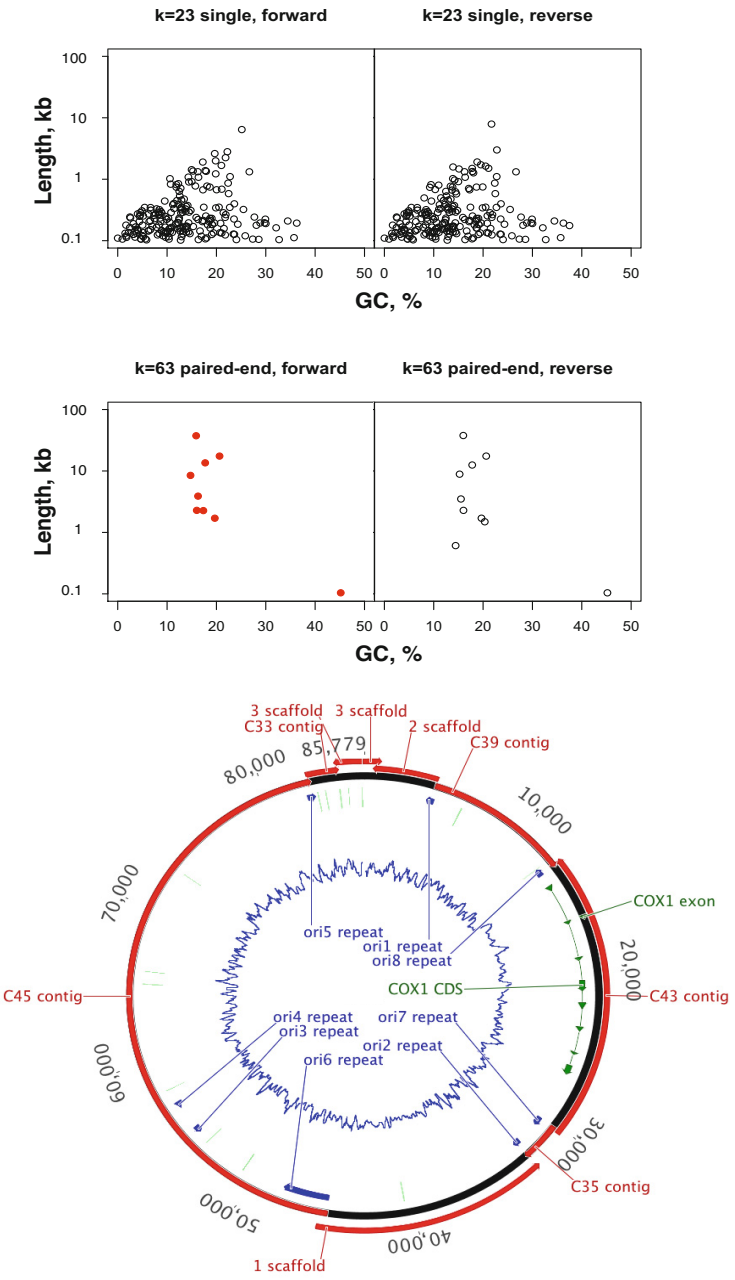


Fig. 2. Simulated NGS strategy designs for the yeast mitochondrial genome. Results of feeding all possible subsequences of length 100 bp of the *Saccharomyces cerevisiae* reference mtDNA genome sequence [14, GenBank acc. AJ011856], as simulated ideal reads at 100× [20], to the SOAPdenovo/GapCloser assembly pipeline [24]. Scaffold/contig properties are shown for oppositely directed sample runs (sense and antisense, for consistency checking) in two contrasting sample scenarios or strategies. The first strategy is single reads at a low *k*-mer size (23 bp), and is completely unsuccessful; the second strategy is paired-end reads separated by 512 bp at a high *k*-mer size (63 bp), and is more successful, but still all scaffolds/contigs are prematurely terminated at *ori* repeats, suggesting a fundamental limit of the NGS sequencing design (lower left scatterplot, and corresponding arcs in circular Geneious [19] display with inner ring showing GC). One sample mitochondrial gene, *CO1*, is shown.

3.4 Benchmarks from Artificially Generated DNA Symbol Sequences for Optimizing Assembly Programs

The last class of benchmarks we discuss comes from artificial DNA. Although the use of such artificial or conceptual DNA (DNA *in silico*) is not backed by a tradition of study by molecular biologists (with an few exceptions), it is backed by a classic branch of thinking and proving theorems about repetitiveness in combinatorics, i.e., in mathematics. This tradition usually bears the name of de Bruijn, although it has its roots in the first graph-theoretic result, Euler's 1766 solution of the Königsberg bridge problem [23, p. 40], [2, Chapter 11].

If mitochondrial genomes, such as that of yeast shown in Figure 2, are some of the toughest benchmarks for assembly-related tasks because of their repeats and low-complexity regions, then we can, conversely, aim to find lenient benchmarks that could serve as a minimal requirement for assembly programs by eliminating repeats or, alternatively, by artificially constructing a library of synthetic DNA (syDNA) sequences that are guaranteed to have no repeats longer than a given length on either strand. For example, the first sequence in such a library could be free of repeats greater than some length w, the second sequence in the library could be free of repeats greater than $w+1$, and so on. Different assembly programs could be assessed or 'acceptance-tested' based on their performance profiles when fed such a synthetic sequence library in read-sized fragments (possibly paired, and/or with simulated sequencing errors to assess robustness).

The problem of constructing, or generating, a repeat-free double-stranded DNA sequence for a given word length w was formulated and addressed already early, in an appendix [15] of a 1966 paper addressing the danger of ectopic (illegitimate, out-of-register) homologous recombination [32]. That appendix presented maximal attainable lengths ($\approx 4^w/2$) and theorems pertaining to those lengths.

The generating of *single-stranded* DNA (ss-DNA) sequences without repeats is relatively straightforward, and even implemented in the general-purpose mathematics software Sage (`DeBruijnSequences(4,`w`).an_element()` and then replace 0,1,2,3 by A,C,G,T; for theory see [2, Chapter 11], and for the elegant generation of a basic example for a given w used by Sage see [30, section 7.3], [13]). However, the study of algorithms for generating *double-stranded* (ds-DNA) repeat-free sequences, in which no words or their reverse complements encounter a match, has not received much attention since 1966. One reason is that the ss-DNA problem is traditionally solved by considering a well-studied class of directed graphs, the de Bruijn graphs, which are used also in several NGS assembly programs. When one considers ds-DNA, however, the underlying structure is less simple: a word needs to be 'glued' to its reverse complement to form a single node or vertex consisting essentially of two node-chambers, connected to other nodes or their chambers by separate directed or doubly-directed edges (bi-directed graph, bi-flow, conjunction product; [26]; cf. also [11]).

Although it would be useful to have an efficient algorithm for directly generating maximal nonrepetitive ds-DNA sequences, e.g., via an analog of the Lyndon word approach used by Sage, one can create nonrepetitive ds-DNA sequences that are not maximal but often sufficiently long for one's purposes, by generating

maximal nonrepetitive ss-DNA sequences and then censoring or editing them. In such editing, visualization using the familiar dotplot method is helpful.

We end this subsection with a toy example. The following randomly generated, 1028-nt de Bruijn sequence, viewed as a simple sequence of letters from a 4-letter alphabet, has no repeats of length 5 nt or more:

```
GCCGAGTCATTTTTATATAGGCTGCCTGCTATACCACGTCTCGGCCCAGCCAAAAAGCTGTTTTCGCTACAGTCTGATCACGGAATGCAAGCA
ACAATGACGTTGACACCCACCAGGTTTGCACCTAAGGTGGGCCGGTCGATGCTCTTATCGCATCCAGTGTAGGGCGTGTTGCCATTCCCGCAA
TCTCACCGTTACTTCATGAGACGATCCGTGAACATCTGTACTAGCTCAACTAAACAGGGACATGTCTTGGTATCCCCAATAAAGGCATAAGAC
AGCGATACACTTTAGTTGTTATGGGAGTATTAGCCTATTCAAGATAGCGCACTATGATGGAGAGGCCAGATCTTTCACTCGTATGTGAGTGAC
CATAGACTACCGGGCTTTTGGACTGACTTAGATTGTGTGCGCGGGTCTAATGTATAATCAGTACAACCAACGGCAAAGAACTTGTAAGTCGGA
TATGCCGCTGGGTTGGCCTCTAGGAGCTAGTGCCCTTGAAGCCGTCACACAAGGGTAACGCAGGAAGGACGCTTGCGGCGAAACTGCGTAATA
CTCATATTGATTCTGCAGTTCAGACCGCGTTTCTATCATCAATTCGAGGGGTGCATGGCGGTGATAACACGCCACTGGCAGAAGTGGCTCCAA
GTTAGGTACCTTACCCCCGAATAGTCCCATGCGAGCCCCTACGTACGGGGGATTTGTCAGCAGCTTCGGTAGTAGAAAGTAAATATCTACTGT
CCACAGAGCGGAGGATGAATCGTGCTTAAAATGGTCAAATCCTGTGGATCGGGAAATTAACCCGTAGCACATTACGCGACAAACCGATTATTT
CCATCGAAGAGTTTACATACGAGAATTTAAGCGTCCTTCTTCCGACTCTCCTCCGCCTTTGAGCATTGGGGCACGAACCTCGCGCTCGACCTG
AAAACGACGGTTCCTAGAGATGTTCGTCGCCCGGCTAACTCCCTGGTGTCGTTCTCTGGAACGTGGTTAATTGCTGAGGTCCGGACCCTCAGG
CGCCG
```

However, if we interpret the sequence as one strand of ds-DNA, the situation changes. Indeed, GAAGG should not occur twice and its reverse complement CCTTC should never occur, but the first two underlined regions show that this condition is not met. Palindromes, i.e., sense-antisense 'self-repeats' (third underlined region), should in principle also be absent, as a short read assembly program could get confused about the direction in which it should continue growing a contig. The underlined regions are the only ones with such problems.

We ran this example through the assembly program `succinctAssembly/gossamer`, which implements efficient algorithms for assembly sub-tasks with few additional heuristics [10], on all of the sequence's 16-bp ds-DNA subsequences choosing a k-mer size of 10. This procedure is analogous to the one used for yeast mtDNA illustrated in Figure 2. As expected, the places where the contigs broke off were precisely the 3 underlined regions.

4 Conclusion

Exploratory analyses of familiar reference sequences, such as those we present here as candidate benchmark loci, could take us a step further toward finding, testing and establishing familiar, modestly sized, biologically interpretable consensus benchmark sequences for fungal and other eukaryotic genomes. The results suggest that we could use selected, known genes or genomic regions as guides to monitor, help predict, and thus optimize the success of sequencing and assembly pipelines or designs for whole genomes.

Acknowledgements. We acknowledge funding from Colciencias, Colombia for the project "Comparative genomics and virulence in the pathogenic fungus *Paracoccidioides brasiliensis*", 2213-48925460. We thank Drs. Natalie Fedorova, Ishwar

Chandramouliswaran, and William C. Nierman (J. Craig Venter Institute) for permission to use the *A. fumigatus* A1163 read set; their work was funded in whole or part with federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services under contract numbers N01-AI30071 and/or HHSN272200900007C (BioProject IDs 14003, 18733, 46347, 52783, 9521, and 67101).

References

1. Audas, T.E., Jacob, M.D., Lee, S.: Immobilization of proteins in the nucleolus by ribosomal intergenic spacer noncoding RNA. Mol. Cell 45, 147–157 (2012)
2. Berge, C.: Graphs. North Holland, Amsterdam (1989)
3. Bernardi, G.: Lessons from a small, dispensable genome: The mitochondrial genome of yeast. Gene 354, 189–200 (2005)
4. Bernardi, G.: Structural and evolutionary genomics: Natural selection in genome evolution. Elsevier, Amsterdam (2005)
5. Bradnam, K.R., Fass, J.N., Alexandrov, A., Baranay, P., Bechner, M., et al.: Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. Giga Science (submitted, 2013), preprint at `http://arxiv.org/abs/1301.5406`
6. Brooks, F.P.: The Mythical Man-Month: Essays on Software Engineering, with four new chapters, Anniversary edn. Addison-Wesley, Reading (1995)
7. Camp, R.: The Search for Industry Best Practices that Lead to Superior Performance, 1st edn. Productivity Press (2006)
8. Carels, N., Barakat, A., Bernardi, G.: The gene distribution of the maize genome. Proc. Natl. Acad. Sci. USA 92, 11057–11060 (1995)
9. Chromatic: Extreme Programming Pocket Guide. O'Reilly Media, Sebastopol (2003)
10. Conway, T.C., Bromage, A.J.: Succinct data structures for assembling large genomes. Bioinformatics 27, 479–486 (2011)
11. Deng, A., Wu, Y.: De Bruijn digraphs and affine transformations. Eur. J. Comb. 26, 1191–1206 (2005)
12. Dimitrov, L.N., Brem, R.B., Kruglyak, L., Gottschling, D.E.: Polymorphisms in multiple genes contribute to the spontaneous mitochondrial genome instability of Saccharomyces cerevisiae S288C strains. Genetics 183, 365–383 (2009)
13. Duzhin, S., Pasechnik, D.: Automorphisms of necklaces and sandpile groups. Preprint, arXiv:1304.2563v1 (2013)
14. Foury, F., Roganti, T., Lecrenier, N., Purnelle, B.: The complete sequence of the mitochondrial genome of *Saccharomyces cerevisiae*. FEBS Lett. 440, 325–331 (1998)
15. Fraenkel, A.S., Gillis, J.: Proof that sequences of A, C, G, and T can be assembled to produce chains of ultimate length avoiding repetitions everywhere. Prog. Nucleic Acid Res. Mol. Biol. 5, 343–348 (1966)
16. Gonzalez, I.L., Sylvester, J.E.: Complete sequence of the 43-kb human ribosomal DNA repeat: analysis of the intergenic spacer. Genomics 27, 320–328 (1995)
17. Henry, T., Iwen, P.C., Hinrichs, S.H.: Identification of Aspergillus species using internal transcribed spacer regions 1 and 2. J. Clin. Microbiol. 38, 1510–1515 (2000)
18. Hinrikson, H.P., Hurst, S.F., De Aguirre, L., Morrison, C.J.: Molecular methods for the identification of *Aspergillus* species. Med. Mycol. 43 (suppl. 1), S129–S137 (2005)

19. Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., et al.: Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28, 1647–1649 (2012)
20. Kingsford, C., Schatz, M.C., Pop, M.: Assembly complexity of prokaryotic genomes using short reads. BMC Bioinformatics 11, 21 (2010)
21. Langmead, B., Salzberg, S.L.: Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359 (2012)
22. Li, H., Durbin, R.: Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760 (2009)
23. Lovasz, L.: Combinatorial Problems and Exercises. North Holland-Elsevier, Amsterdam (1993)
24. Luo, R., Liu, B., Xie, Y., Li, Z., et al.: SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. GigaScience 1, 18 (2012)
25. Lynch, M., Sung, W., Morris, K., Coffey, N., Landry, C.R., et al.: A genome-wide view of the spectrum of spontaneous mutations in yeast. Proc. Natl. Acad. USA 105, 9272–9277 (2008)
26. Medvedev, P., Brudno, M.: Maximum likelihood genome assembly. J. Comput. Biol. 16, 1101–1116 (2009)
27. Morgulis, A., Gertz, E.M., Schäfer, A.A., Agarwala, R.: A fast and symmetric DUST implementation to mask low-complexity DNA sequences. J. Comp. Biol. 13, 1028–1040 (2006)
28. Parra, G., Bradnam, K., Korf, I.: CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 23, 1061–1067 (2007)
29. Parra, G., Bradnam, K., Ning, Z., Keane, T., Korf, I.: Assessing the gene space in draft genomes. Nucleic Acids Res. 37, 289–297 (2009)
30. Ruskey, F.: Combinatorial Generation. Working version 1j-CSC 425/520. Available at CiteSeer:10.1.1.93.5967 (2003)
31. Seifert, K.A., Samson, R.A., de Waard, J.R., Houbraken, J., Lévesque, C.A., et al.: Prospects for fungus identification using *CO1* DNA barcodes, with Penicillium as a test case. Proc. Natl. Acad. USA 104, 3901–3906 (2007)
32. Thomas Jr., C.A.: Recombination of DNA molecules. Prog. Nucleic Acid Res. Mol. Biol. 5, 315–337 (1966)
33. Wang, W., Wei, Z., Lam, T.-W., Wang, J.: Next generation sequencing has lower sequence coverage and poorer SNP-detection capability in the regulatory regions. Sci. Rep. 1, 55 (2011)

Chapter 4

The complex task of choosing a *de novo* assembly:

Lessons from fungal genomes

G Model
CBAC-6351; No. of Pages 11
ARTICLE IN PRESS

Computational Biology and Chemistry xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Research Article

The complex task of choosing a *de novo* assembly: Lessons from fungal genomes

Juan Esteban Gallo[a,b], José Fernando Muñoz[a,c], Elizabeth Misas[a,c], Juan Guillermo McEwen[a,d], Oliver Keatinge Clay[a,e,*]

[a] *Cellular & Molecular Biology Unit, Corporación para Investigaciones Biológicas, Medellín, Colombia*
[b] *Doctoral Program in Biomedical Sciences, Universidad del Rosario, Bogotá, Colombia*
[c] *Institute of Biology, Universidad de Antioquia, Medellín, Colombia*
[d] *School of Medicine, Universidad de Antioquia, Medellín, Colombia*
[e] *School of Medicine and Health Sciences, Universidad del Rosario, Bogotá, Colombia*

ARTICLE INFO

Article history:
Accepted 11 July 2014
Available online xxx

Keywords:
Genome assembly methods
Repetitive DNA
Next-generation sequencing

ABSTRACT

Selecting the values of parameters used by *de novo* genomic assembly programs, or choosing an optimal *de novo* assembly from several runs obtained with different parameters or programs, are tasks that can require complex decision-making. A key parameter that must be supplied to typical next generation sequencing (NGS) assemblers is the *k*-mer length, i.e., the word size that determines which de Bruijn graph the program should map out and use. The topic of assembly selection criteria was recently revisited in the Assemblathon 2 study (Bradnam et al., 2013). Although no clear message was delivered with regard to optimal *k*-mer lengths, it was shown with examples that it is sometimes important to decide if one is most interested in optimizing the sequences of protein-coding genes (the gene space) or in optimizing the whole genome sequence including the intergenic DNA, as what is best for one criterion may not be best for the other. In the present study, our aim was to better understand how the assembly of unicellular fungi (which are typically intermediate in size and complexity between prokaryotes and metazoan eukaryotes) can change as one varies the *k*-mer values over a wide range. We used two different *de novo* assembly programs (SOAPdenovo2 and ABySS), and simple assembly metrics that also focused on success in assembling the gene space and repetitive elements. A recent increase in Illumina read length to around 150 bp allowed us to attempt *de novo* assemblies with a larger range of *k*-mers, up to 127 bp. We applied these methods to Illumina paired-end sequencing read sets of fungal strains of *Paracoccidioides brasiliensis* and other species. By visualizing the results in simple plots, we were able to track the effect of changing *k*-mer size and assembly program, and to demonstrate how such plots can readily reveal discontinuities or other unexpected characteristics that assembly programs can present in practice, especially when they are used in a traditional molecular microbiology laboratory with a 'genomics corner'. Here we propose and apply a component of a first pass validation methodology for benchmarking and understanding fungal genome *de novo* assembly processes.

1. Introduction

"Complexity" is, itself, a complex concept. In a review of complexity and its measures in the context of DNA, Li (1997) has correctly pointed out the need to occasionally return to first principles when considering which, if any, of the more widely used (and sometimes not fully interchangeable) definitions of complexity is most relevant to a given task at hand. Indeed, the task itself can be used to define a measure of complexity of a genome (or other object), namely the difficulty of accomplishing that task on the genome (Li, 1997and refs. therein).

A task associated with eukaryotic genomics that has become increasingly important for individual laboratories as DNA sequencing costs continue their decrease is the task of obtaining a reliable, annotated *de novo* assembly of a genome of a species, isolate/strain/population or, especially in large metazoans, of an individual. Depending on the questions one wishes to answer by obtaining genome sequences, in some contexts one may be interested primarily in obtaining the sequences of all genes or coding

* Corresponding author at: Cellular & Molecular Biology Unit, Corporación para Investigaciones Biológicas, Medellín, Colombia.
E-mail address: oliver.clay@gmail.com (O.K. Clay).

http://dx.doi.org/10.1016/j.compbiolchem.2014.08.014
1476-9271/

G Model
CBAC-6351; No. of Pages 11
ARTICLE IN PRESS

regions of the genome being studied; in other contexts one may also need to obtain the unique intergenic regions that could be relevant for the genes' expression such as promoters and enhancers; finally, one's aim may be to offer to the community a 'full' reference sequence for a genome, including as much of its repeat landscape as is feasible given the available technology, time and funds. The optimal design or simulation of a sequencing strategy, and indeed the 'complexity of the genome' one faces, can vary depending on the precise task of interest.

Clearly it can be of much utility if one can design, compare and then choose a relatively optimal sequencing strategy together with the subsequent steps of assembly and annotation, before embarking on the actual library preparation and sequencing (Muñoz et al., 2014). For this one needs to know what one wants most, e.g., what cost-effectiveness, cost–utility or cost–benefit trade-offs (Drummond et al., 2005) one is prepared to make, and what sequencing technologies, options, and hardware configurations are available to one's laboratory.

There have been promises of what the next NGS generation (post-next generation sequencing, or third/fourth generation sequencing), will soon be able to offer. Some of the promises still need to be fulfilled, however. As a recent article in Science (Pennisi, 2014) has emphasized, because of various delays we have still not fully arrived at the goal of empowering small laboratories to enjoy reliable, stand-alone implementations of the projected new technologies that fit within their often modest budget.

In this article, we therefore focused on a second generation technology, Illumina short read sequencing (single or paired-end reads), a widely used solution that has consistently remained accessible and typically reliable now for several years, for which costs charged by some service providers have decreased, and for which read lengths have recently increased to about 150 bp. Insights obtained for this specific technology can be partly generalized or extrapolated, *mutatis mutandis*, to other read sequencing technologies that also use de Bruijn graph-based assembly methods (Compeau et al., 2011). Also, a recent analysis by Li et al. (2014) in the context of reference assembly, showing that mappability of reads (or *r*-mers) to reference sequences does not markedly improve above read lengths *r* of 150–200 bp, could suggest that at least some findings for *de novo* assembly of currently available (≈150 bp) reads might extrapolate well to longer Illumina reads that may be available in future.

We will restrict our genomic context to unicellular fungi, eukaryotes that have served as models of metazoan eukaryotes (including human) in many respects including disease (Bassett et al., 1996; Foury, 1997), and that represent an intermediate in size and complexity between the genomes of prokaryotes (bacteria or archea) and those of vertebrates and flowering plants. Since the prime model fungus, the baker's yeast *S. cerevisiae*, has a relatively compact nuclear genome, we break with tradition by choosing as our main example another, less compact fungal genome that has recently been fully sequenced, assembled and annotated (Desjardins et al., 2011), the thermally dimorphic human pathogenic fungus *Paracoccidioides brasiliensis*. Its nuclear genome has a larger size and contains more repetitive DNA than yeast, and in this respect it may be more representative of other eukaryotes and their assembly challenges.

In short read sequencing of such genomes, and the subsequent processing of the reads, there are at least three stable, key parameters that can currently be considered fundamental (Fig. 1). Other measures or parameters determining the success of an assembly include measures pertaining to read depth or coverage, and non-tunable measures reflecting the repeat structure or profiles of the genome itself.

Two of the key parameters depicted in Fig. 1 are sequencing-related, namely the read size *r* and, if paired-end reads are used as part of the design, also the insert size *I* of the insert library.

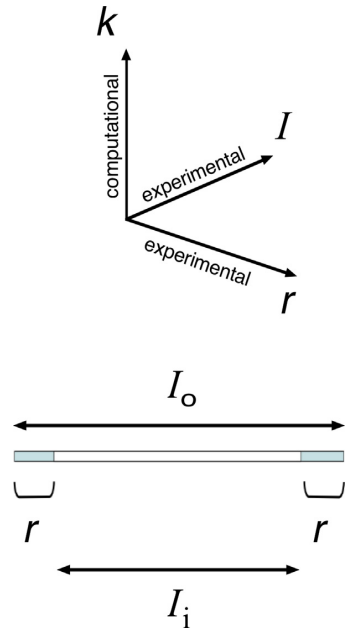


Fig. 1. Three key length parameters that can influence the structure of *de novo* assemblies obtained from Illumina short reads. Two parameters that are decided at the sequencing-design stage (horizontal plane in the scheme at top) are read length *r* and, where paired-end reads are used, insert length *I*. An additional key parameter that can be chosen or modified at the assembly stage (vertical axis in the scheme at top) is the *k*-mer length, which is a tunable parameter ($k < r$) in typical de Bruijn graph-based *de novo* NGS assembly programs, including SOAPdenovo2 (Luo et al., 2012), ABySS (Simpson et al., 2009), Velvet (Zerbino and Birney, 2008) and ALLPATHS (MacCallum et al., 2009). Insert lengths are often considerably longer than read lengths, and are therefore denoted here by an upper-case symbol; to communicate a design either the outer insert length I_o (the length of the insert between adapters, which includes the two end regions that are sequenced as reads) or, less frequently, the inner insert size $I_i = I_o - 2r$ (the distance between the two reads) have been used in the literature; clearly only one of them needs to be specified when the read size *r* is given (scheme at bottom). Insight into methodologies (experimental and/or computational) can be gained by choosing a single species/isolate, and then comparing sequencing designs (having different *r* and/or *I* values) and assembly designs (using different programs and a range of *k* values).

In actual practice, where there is a spread of insert sizes, one can take the mean insert size, i.e., the average fragment size $<I_o>$, to represent the insert size (see Fig. 1), or one can include more information ranging from the standard deviation to an estimate of the full frequency distribution of fragment lengths given by an electropherogram.

The third key parameter depicted in Fig. 1 is a bioinformatic parameter that is used only after the sequencing of the raw reads is done, the so-called *k*-mer length, or *k*. This parameter is the length of a read's fixed-length subsequences, or words, that are used in the de Bruijn graph-based assembly steps that form a part of most modern, widely used short-read assembly programs (Compeau et al., 2011). Given the value of *k*, the assembly program traces out from the reads the corresponding regions of a full de Bruijn graph over an alphabet of size 4 (i.e., 0,1,2,3 or A,C,G,T) and dimension *k* (Berge, 2001, p. 167 ff.). In this way, flows on the de Bruijn graph correspond to successive 1-step shifts of a window of size *k* traversing a sequence. At least conceptually, one needs to generalize the de Bruijn graph because reverse complements should not be counted twice (Medvedev and Brudno (2009), cf. also Deng and Wu (2005)). Implementations differ, but they all involve steps in which regions of a full de Bruijn or de Bruijn-like graph are traced out by the reads and the flows are recorded (see Muñoz et al.

Please cite this article in press as: Gallo, J.E., et al., The complex task of choosing a *de novo* assembly: Lessons from fungal genomes. Comput. Biol. Chem. (2014), http://dx.doi.org/10.1016/j.compbiolchem.2014.08.014

G Model
CBAC-6351; No. of Pages 11
ARTICLE IN PRESS

(2014) and refs. therein). Which de Bruijn-like graph is partially constructed is decided by the parameter *k*. The value of *k* can also be thought of as the 'window size' through which these steps of the assembly program 'see' the reads. A further consideration, which different assembly programs may deal with in different ways, is the additional constraint on the possible assemblies imposed by the pairing of the reads, i.e., by the knowledge of which reads are paired, together with the knowledge of the frequency distribution of the insert/fragment sizes in the library used for sequencing: this is where the experimental parameter *I* is needed.

The ability to wisely choose options and parameters ensuring a good *de novo* assembly is a valuable asset when designing a sequencing project for a fungal genome. One sometimes needs to consider bioinformatic options or parameters at the same time as one chooses experimental details for the initial read sequencing (as one would do when writing a grant proposal). Given a selected fungal organism (species and strain/isolate) and a particular experimental strategy, which for Illumina short-read sequencing will always involve choosing read length *r* and usually also insert length(s) *I*, what assembly and annotation quality could we expect in the best possible case? A closely related reference species' existing assembly or reads, where available, can often help answer this question *in silico*. Simulated reads can be obtained from the scaffolds by *in silico* 'sonication' and fed to a series of different assembly and annotation runs with different parameters/choices, thus averting unpleasant surprises later, such as the discovery that one's choice of read or insert length was inherently unsuited for the chosen genome and task. Low and high *k* values may have different advantages and drawbacks.

The problem addressed here, the choice of an assembler and '*k* strategy', or the choice of an assembly from a set of assemblies obtained from the same reads using different strategies, is indeed a complex one, depending also on the repeat complexities of the genome of interest, as is now finally being acknowledged (see, e.g., Haznedaroglu et al. (2012) and refs. therein). In a single contribution presenting original data it would be very difficult to do justice to this complex problem if one insists on also maintaining generality, providing an exhaustive review, or suggesting general guidelines that one cannot feasibly alpha-test. In this contribution, we therefore choose a restricted context, and report results from an initial exploration of fungal assemblies from real reads, using two widely used assembly programs or pipelines that were run over a wide range of *k*-mer lengths: the SOAPdenovo2 pipeline (up to but not including the final GapCloser run; Luo et al. (2012)) and the ABySS program (Simpson et al., 2009). We imagine a typical molecular microbiology laboratory endowed with a medium-sized grant to initiate constructive genomics work (e.g., a server with no more than 64 GB RAM), and having no bioinformatic 'insider information' or dedicated informatic personnel, i.e., no access to tips, tricks or 'tweaks' that are not available in the public domain. Although Illumina read-based quality reports and other analytical studies have been offered before, to our knowledge few if any focused on non-model fungi having repeat-rich genomes, or 150 bp read sets; including such realistic and timely features that are likely to reflect real in-house genomics projects allows us to obtain a wider picture.

2. Materials and methods

2.1. Sequencing

Paracoccidioides brasiliensis strains Pb113, Pb1445 and PbBAC were sequenced using an Illumina HiSeq 2500 sequencer, with 150 bp paired-end reads and a nominal length (insert size) of 620 bp. Insert library construction and sequencing were performed at the DNA Services facilities of the University of Illinois at Urbana-Champaign.

Saccharomyces cerevisiae strain S288c paired-end reads were downloaded from the NCBI Sequence Read Archive (SRA), Experiment ERX240932, Run ERR266425. The reads were sequenced using Illumina HiSeq 2000 with a nominal length (insert size) of 450 bp and a read length of 144 bp.

Although plots are shown here only for *P. brasiliensis* strain Pb113 as an example, full plots were also obtained for the other two *P. brasiliensis* strains and the *S. cerevisiae* strain, with corresponding findings.

For comparing SOAPdenovo2 metrics at the observed jump from 49 bp $\leq k \leq$ 51 bp for a third fungal genus, paired-end Illumina reads (read length 101 bp) of *Emmonsia crescens* strain UAMH 4076 and *Emmonsia parva* strain UAMH 130 were used. Insert library construction and sequencing with a HiSeq 2000 sequencer were performed at the Broad Institute, Cambridge, MA (collaborative paper in preparation; NCBI Bioproject PRJNA179100).

2.2. Assemblies

The program SOAPdenovo version 2.04-r240 (SOAPdenovo2; Linux binary for *k* values up to 127 bp) was used with all odd (i.e., allowed) *k*-mer values from 17 bp to 127 bp. The assemblies were run essentially using default parameters (except for the *k*-mer values), only modifying the config files with the corresponding read lengths and insert sizes, along with the maximum read length as the read length. The recommended SOAPdenovo2 pipeline also includes the programs KmerFreq and Corrector (SOAPec_v2.01), which are run prior to calling the SOAPdenovo2 program, and GapCloser, which is run after SOAPdenovo2 as a final step. Paradoxically, however, KmerFreq comes with no clear recommendations for pre-piping reads for which one wants SOAPdenovo2 to use high *k* values (e.g., around 50): KmerFreq_HA does not accept *k* values above 27, and KmerFreq_AR uses 4 GB RAM for *k*=17, 256 GB RAM for *k*=19, etc., and so is not useful in normal practice even at moderately high *k* values. In both variants of KmerFreq, the defaults were 17, so we used that value for all assemblies in KmerFreq_AR (*-k* 17 *-q* 33 *-t* 10) and Corrector_AR (*-k* 17 *-l* 3 *-Q* 33 *-r* 50 *-t* 12). Because we wanted to focus on the processes of the core assembly program, we did not attempt to close the gaps remaining after SOAPdenovo2 by running GapCloser.

ABySS version 1.3.7 was used with all odd *k*-mer values from 17 bp to 127 bp. The assemblies were run using default parameters.

2.3. Genes and repeats

BLASTX version 2.2.28+ was run locally with default settings using protein sets from reference annotations as the database, and the 17–127 bp *k*-mer assemblies were the query.

Augustus (Keller et al., 2011, http://bioinf.uni-greifswald.de/augustus/) was used for fully automated gene calling on each assembly.

RepeatMasker version open-4.0.3 was used to detect repetitive elements in each assembly using the fungal repeats data set from Repbase.

LTR_Finder version 1.0.5 (tlife.fudan.edu.cn/ltr_finder/) was used to find long terminal repeat (LTR) transposons in each assembly, with default settings.

3. Results

3.1. Metrics for assessing assembly of bulk DNA

We used the SOAPdenovo2 pipeline (without GapCloser post-processing) and the ABySS program/pipeline to generate a large

Please cite this article in press as: Gallo, J.E., et al., The complex task of choosing a *de novo* assembly: Lessons from fungal genomes. Comput. Biol. Chem. (2014), http://dx.doi.org/10.1016/j.compbiolchem.2014.08.014

G Model
CBAC-6351; No. of Pages 11
ARTICLE IN PRESS

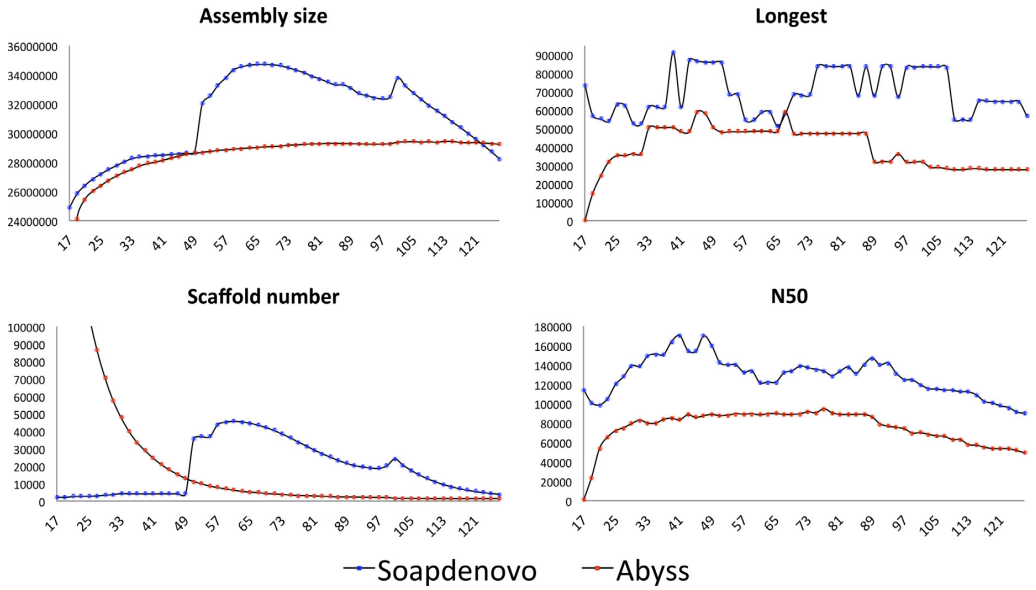


Fig. 2. Variation of four assembly statistics (metrics assessing assembly of bulk DNA) for *P. brasiliensis* strain Pb113, as the chosen *k*-mer length is increased: assembly size, size of longest scaffold, number of scaffolds, and N50. The *k*-mer lengths ranged from 17 to 127 bp, and the assembly pipelines used were SOAPdenovo2 and ABySS. The shape of the plot for SOAPdenovo2 (but not ABySS) assembly size resembles the contour of a mouse ('mouse model' or shape with its 'rump' at 50 bp and pointed 'ears' at 100 bp). A similar trend is seen when one chooses the total number of scaffolds in the assembly as the statistic. The plots of SOAPdenovo2 assemblies, also in the other three genomes studied, show an upward discontinutity or 'jump' in assembly size and number of scaffolds between *k*-mer lengths 49 bp and 51 bp (see Table 1). An upward jump at this *k*-mer length was present in the SOAPdenovo2 assemblies of all real genomes we studied, but in no ABySS assemblies; the smaller 'ears' jump was outside the *k* range for sets of shorter reads (101 bp, Table 1).

number of assemblies of paired-end Illumina reads for each fungal genome of interest, and then compare those assemblies via a series of bulk DNA based, gene-based and repeat-based metrics. To make descriptions simple and uniform, throughout the main text of this paper we will loosely refer to all final (i.e., maximally extended or joined) sequences of the assembly that are delivered by the two pipelines as *scaffolds*, although (at least for SOAPdenovo2) some of these long sequences are technically contigs.

The four fungal genomes we chose for detailed analytics were three strains of a human pathogen endemic to several regions of South America, *P. brasiliensis* (Pb113 from Brazil, Pb1445 from Argentina and PbBAC from Colombia; for the genome of other conspecific strains, see Desjardins et al. (2011)), and the model strain S288c of the baker's yeast *S. cerevisiae* (for the genome assembly and complexity analyses of this strain, see (Goffeau et al., 1996; Li et al., 1998; Román-Roldán et al., 1998; Li, 1997)).

For reads of 150 bp, we ran each of the two assembly programs for every possible odd *k*-mer length from 17 bp to 127 bp, resulting in a total of 112 assemblies for each genome. For reads shorter than 150 bp, we ran the assemblies over a correspondingly smaller range of *k* values.

The resulting plots of four simple bulk DNA metrics are shown for the assemblies from reads (150 bp) of *P. brasiliensis* isolate Pb113 in Fig. 2. It can be seen that whereas in all plots for the ABySS runs the *k*-dependence is relatively smooth, and in two metrics (assembly size and scaffold number) completely monotone, the shapes of all four SOAPdenovo2 plots are more jagged, suggesting underlying complexity in the ways this pipeline or some of its algorithms handle the Illumina reads at different *k*-mer sizes. In particular, such plots dispel the notion that any assembly program will invariably have a single, smooth local maximum near a traditional *k*-mer value, and that exploring assemblies at *k*-mer values farther away will merely give predictable or irrelevant results. Without having to delve into the details of an assembly program's source code, which is a complex venture even where the program is open source, we already have a first, graphic overview of how the program behaves and 'reacts' to changes in the key parameter *k*.

Two of the four SOAPdenovo2 plots exhibit at least one characteristic jump or 'cliff' for Pb113. Interestingly, N50, which has been one of the most widely used metrics for such plots, does not reveal the characteristic jump(s) here. Where longest scaffold is the metric, the 'Arizona butte' landscape one sees in this particular example was not a surprise, as scaffolds tend to terminate at 'weak spots' (e.g., non-unique regions) of the genome that are often far apart. By contrast, the plots for assembly size and scaffold number were a surprise, and have a characteristic 'mouse profile' shape; abrupt upward jumps are seen at 49 bp $\leq k \leq$ 51 bp in Fig. 2 and, to a smaller extent, at 99 bp $\leq k \leq$ 101 bp ('ears' of the mouse). Proportionally, the scaffold number has the most dramatic jump: the tiny change of *k* value from 49 bp to the next allowed (i.e., odd) value, 51 bp, caused an order-of-magnitude increase in the number of scaffolds, from around 4000 to around 35,000. This jump was caused by a large increase of short sequences of length 100 bp in the reported assembly (file `.scafSeq`; see Supplementary Fig. S3).

A similarly accidented dependence on *k* was seen for assembly size and scaffold number in the other SOAPdenovo2, but not ABySS, assemblies of the other three fungal genomes. We also ran assemblies at *k*=49 and 51 bp for reads of two *Emmonsia* species that were sequenced as part of another study (101 bp paired-end reads). The height of the jump at 49 bp $\leq k \leq$ 51 bp varied, however: the assembly size increase at that *k* increment, reported in Table 1, is smallest but still clearly present in the more compact (repeat-poorer) yeast genome, and largest in the less compact (repeat-richer) *Emmonsia* genomes.

To check that the SOAPdenovo2 program did not exhibit trivial or general discontinuities also in genomes where no repeats are present, we used ShortCAKE (Orenstein and Shamir, 2013) to synthetically generate a maximal linear sequence or 'chromosome' of

Please cite this article in press as: Gallo, J.E., et al., The complex task of choosing a *de novo* assembly: Lessons from fungal genomes. Comput. Biol. Chem. (2014), http://dx.doi.org/10.1016/j.compbiolchem.2014.08.014

G Model
CBAC-6351; No. of Pages 11
ARTICLE IN PRESS

Table 1
Discontinuity in the size of assemblies produced by SOAPdenovo2 between *k*-mer lengths 49 and 51 bp, for several fungal species and strains. The table also reports the design (*r*, *I*) and count of the read set for each strain. Pb113 is the *P. brasiliensis* strain depicted in the plots of this paper. Even higher jumps than those of *P. brasiliensis* (for which the reference assemblies for Pb18 and Pb03 have sizes 29.95 and 29.06 Mb, respectively) are seen in *Emmonsia* spp., presumably as a result of a higher repeat content. Conversely, the much more compact genome of baker's yeast *S. cerevisiae* strain S288c (reference assembly size 12.16 Mb) has a correspondingly low, but still clearly noticeable jump between *k*-mer lengths 49 and 51 bp; in the case of these (SRA public domain) yeast reads the discontinuity was more pronounced for scaffold number, which jumped from 8705 for *k*=47 and 9017 for *k*=49 to 15,315 for *k*=51 bp.

Species	Strain	Increase in assembly size (Mb) from *k*=49 to *k*=51	Number of reads	*r* (bp)	*I* (bp)
P. brasiliensis	Pb113	3.2	15,515,977 × 2	150	620
P. brasiliensis	Pb1445	3.2	16,165,496 × 2	150	620
P. brasiliensis	PbBAC	3.3	13,868,844 × 2	150	620
S. cerevisiae	S288c	0.7	50,352,236 × 2	144	450
E. crescens	UAMH 4076	7.4	48,133,047 × 2	101	157
E. parva	UAMH 130	9.6	47,986,642 × 2	101	158

length 2,097,152 bp engineered to have no repeats of size 11 bp or higher on either strand, i.e., a double-stranded de Bruijn sequence of word size 11 on a 4-letter alphabet (see Fraenkel and Gillis, 1966; Muñoz et al., 2014). We then created artificial paired-end reads from that sequence with a read size *r* of 150 bp and I_o of 770 bp, and fed them to the SOAPdenovo2 pipeline under the same conditions and *k*-mer values as we had done for the real sequences. Although to our surprise SOAPdenovo2 did not assemble the synthetic read set on our system for some intermittent (isolated) *k* values because of segmentation faults, assembly did succeed for many *k* values throughout the 17–127 bp range we used for the *P. brasiliensis* genomes. For all of the successful *k*-mer values, the SOAPdenovo2 assembly consisted of a single scaffold recreating the original sequence, i.e., corresponded to flat (horizontal) plots in which no metrics depend on *k*.

This simple negative control does not rule out that the upward jumps at 50 bp (and/or at 100 bp) in the real read sets were caused by internal thresholds in algorithm(s) or internal parameters that are used by SOAPdenovo2, or in the preprocessing tools KmerFreq_AR and Corrector_AR. Indeed, it seems likely that internal decisions inside program(s), such as if–then–else decision(s) involving cutoff(s), contributed to the discontinuities observed (see Section 4). In either case, the plots are relevant for this study: one of the aims of the present study was to investigate the utility of plotting metrics' *k*-dependence as a way of gaining insight into the pros and cons of choosing individual or combined *k*-mer lengths for sequence assembly, in actual practice. Variations with respect to *k* that a microbiologist using genomics may encounter when properly using off-the-shelf assembly software are of interest. By visualizing such real variation, in graphical displays of pertinent species' genome assemblies, one is less likely to make a choice of *k*-mer size that later turns out to be far from optimal.

3.2. Metrics for assessing gene presence in the assemblies

We next investigated the contributions of unique genes and repeats to the metric *vs*. *k* plots observed in Fig. 2.

Rather than attempting to create fine-tuned gene annotations in order to assess gene presence, we used two imperfect but straightforward, easily reproducible methods to roughly estimate relative presence of probable genes. Neither of these methods explicitly involved arbitrary 'manual' decisions, or arbitrary choices of tunable weights.

The first method was to choose two reference gene sets from other strains/ species, and track how many of their genes/proteins had one or more BLASTX matches in the 112 Pb113 assemblies obtained by the two different programs. For the two reference gene sets, we chose the previously published official gene set of *P. brasiliensis* Pb18, which is closely related to the reference strain Pb113, and a reference set of well-conserved core genes, CEGMA (Parra et al., 2007, 2009; Bradnam et al., 2013), for which the fungal reference species are *S. cerevisiae* and *Schizosaccharomyces pombe*. The resulting plots are displayed in Fig. 3. The proportion of reference genes that matched a region in the assemblies (a rough estimate of the gene-level coverage of the *P. brasiliensis* genome) was very high in all cases, with the percentage of hits never falling below 99 percent. The plots show no substantial jump in the relative number of matches in SOAPdenovo2 assemblies around *k*-mer values 50 bp or 100 bp; even though a fraction of the genes may match paralogs (i.e., similar but 'different' genes) rather than, or in addition to, orthologs (i.e., the 'same' or corresponding gene), the absence of a substantial jump indicates that gene number differences cannot explain the 'jump of the mouse', nor are they affected by it. The plots also suggest that the presence of reference genes is about as good at low *k* values as it is at high *k* values. BLASTX similarity histograms for different *k* values are shown in Supplementary Fig. S1.

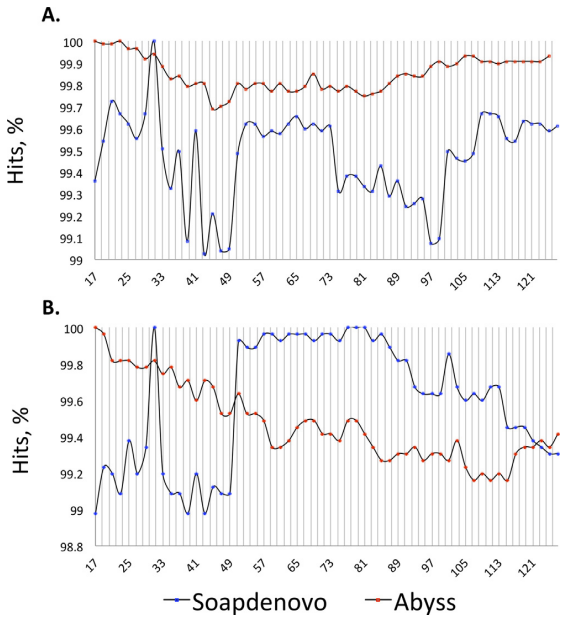


Fig. 3. Variation of two gene-number based assembly metrics using BLAST for *P. brasiliensis* strain Pb113, as the chosen *k*-mer length is increased from 17 bp to 127 bp. The plots show the percentage of genes from the fully annotated reference assembly of the closely related *P. brasiliensis* strain Pb18 (A, as subject) or from the CEGMA core reference genes (B, as subject/database) that had one or more BLASTX match(es) to the SOAPdenovo2 and ABySS assemblies of Pb113 (as query). The fluctuations seen appear amplified by the narrow range chosen for the vertical axis: the percentage of hits remains above 99% at all *k*-mer values for both assembly programs and both metrics.

Please cite this article in press as: Gallo, J.E., et al., The complex task of choosing a *de novo* assembly: Lessons from fungal genomes. Comput. Biol. Chem. (2014), http://dx.doi.org/10.1016/j.compbiolchem.2014.08.014

G Model
CBAC-6351; No. of Pages 11
ARTICLE IN PRESS

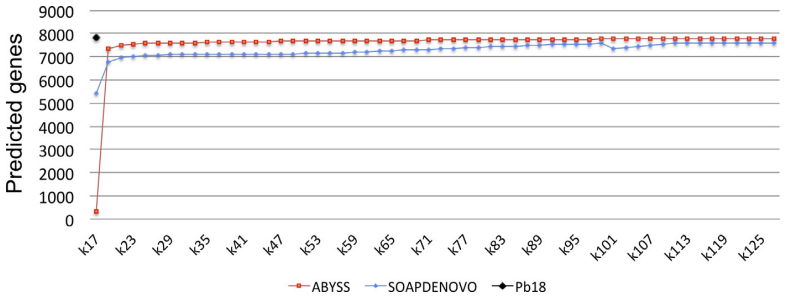


Fig. 4. Number of *ab initio* predicted genes present in assemblies obtained using different *k*-mer sizes (17 bp ≤*k*≤ 127 bp). The plot shows the number of genes predicted *ab initio* by Augustus in the *P. brasiliensis* strain Pb113 scaffolds assembled by SOAPdenovo2 and ABySS. The number of genes predicted by Augustus in the Pb18 reference assembly is also shown as a positive control (far left, black spot).

Our second method to estimate relative gene presence at different *k* values was to run the gene calling program Augustus on each *P. brasiliensis* Pb113 assembly, with the closely related fungus *Histoplasma capsulatum* as the training species, and use its raw single-pass annotation output to assess gene assembly. The number of genes found in the 112 re-runs of this program on different assemblies is shown in Fig. 4; the extents of the genes, and their breakdown into contributions from coding and intron DNA, are shown in Fig. 5. The data reveal only a tiny dependence on *k* of the number, lengths, coding region lengths or intron lengths. They show no trace of the main discontinuity (at *k*=50 bp) exhibited by the total coding + noncoding DNA of the Pb113 and other fungal assemblies.

3.3. Methods for assessing presence of previously characterized repeat groups in the assemblies

The presence of genes in the assemblies, which corresponds to representation or coverage of the gene space (Carels et al., 1995; Parra et al., 2009), did not vary appreciably with the *k*-mer length used for constructing the assemblies, and is therefore not affected by the bulk DNA metrics' sometimes strong or irregular dependence on *k* that we had observed (Fig. 2), so we next examined repeats. We directly assessed differences in repeat content using a program that detects a wide range of simple and interspersed repeats that are known, RepeatMasker. The number and extents of repeat-matching DNA (for RepeatMasker) are shown in Figs. 6 and 7. For the SOAPdenovo2 assemblies, both plots (*k*-dependencies) can be seen to covary with those for assembly size and scaffold number in Fig. 2 above, with accidents near the same *k* values (50 bp and 100 bp) and a similar 'mouse' shape. This parallelism, and the large numbers and total lengths of the repeats involved, indicate that the abrupt increase in assembly size and in scaffold number between *k*≤ 49 and *k*≥ 51 bp is at least partly a result of an increased proportion of repetitive DNA. By contrast, the repeat metrics for the ABySS assemblies showed very different *k* dependencies, i.e., plot shapes, compared to those for assembly size or scaffold number; both repeat number (Fig. 6) and total repeat length (Fig. 7) rose quickly to a local (and, for repeat number, global) maximum at around *k*≈ 25 bp; for the rest of the range, the number plot remained almost

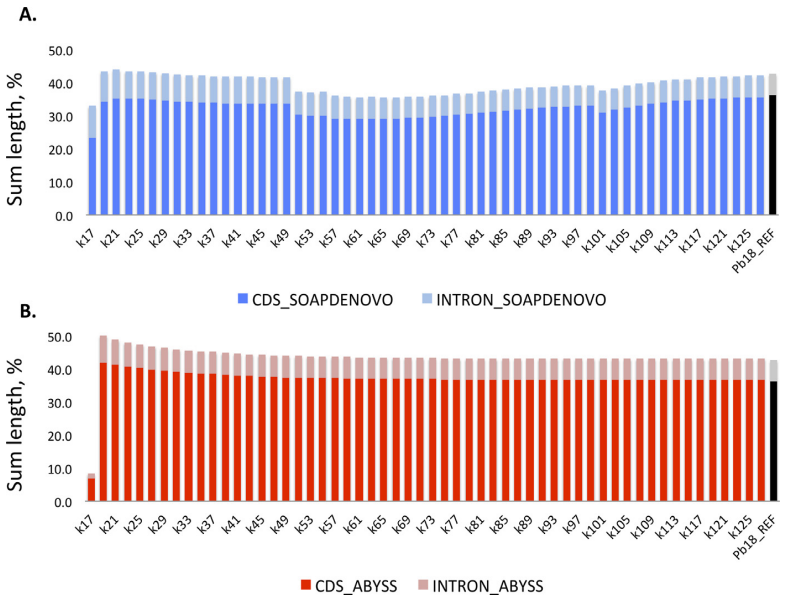


Fig. 5. Total lengths of genes predicted *ab initio* by Augustus that were present in the assemblies of *P. brasiliensis* strain Pb113 (17 bp ≤*k*≤ 127 bp; colored bars) and, for comparison, in the reference assembly of the closely related *P. brasiliensis* strain Pb18 (black bar at far right). The plots show the relative contributions (total lengths) of predicted coding (CDS; bold color) and intron sequences (faint color), in the assemblies obtained via SOAPdenovo2 (A) and ABySS (B). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Please cite this article in press as: Gallo, J.E., et al., The complex task of choosing a *de novo* assembly: Lessons from fungal genomes. Comput. Biol. Chem. (2014), http://dx.doi.org/10.1016/j.compbiolchem.2014.08.014

G Model
CBAC-6351; No. of Pages 11
ARTICLE IN PRESS
J.E. Gallo et al. / Computational Biology and Chemistry xxx (2014) xxx–xxx

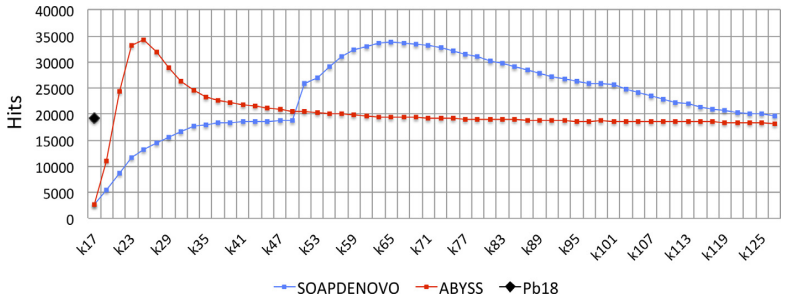


Fig. 6. Presence of repeats in scaffolds assembled using different *k*-mer values (17 bp ≤*k*≤ 127 bp) and programs, as illustrated by the number of matches (repeats) that were reported by RepeatMasker in *P. brasiliensis* strain Pb113 assemblies (colored curves) and, for comparison, in the Pb18 reference assembly (black spot at far left). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

flat from *k*≈ 70 bp onwards, while the total length plot maintained a slight upwards slope.

We also ran a more specialized program for detecting long terminal repeat transposons, LTR_Finder, and plotted the number presence of LTR transposons (Supplementary Fig. S2). Although the plots exhibited a consistent trend in both SOAPdenovo and ABySS assemblies, rising with increasing *k*, the numbers of detected LTR transposons per assembly were tiny, even in the Pb18 reference assembly (only 24 transposons), and apparent fluctuations were statistically unreliable.

4. Discussion

4.1. Effects of varying k-mer length

For the *de novo* assemblies created by ABySS, plots of essentially all metrics *vs*. the fundamental assembly software parameter *k* changed smoothly and slowly as *k* was increased, and for several metrics the dependency was monotone or close to monotone (i.e., either increasing or decreasing). As expected, one metric, the length of the longest scaffold in the assembly, was the exception, and its discontinuities are easily explained by the clustering of repetitive DNA in the genome. Indeed, the main reason for premature terminating of extending scaffolds or contigs is the presence of repetitive DNA: where subsequences occur several times in the genome in different places, *de novo* read assembly programs (and reference read assembly programs (Li et al., 2014)) can be faced with an ambiguous situation that they cannot resolve without access to additional contiguity information. Thus, the synthetic double-stranded sequence of >2 Mb we created using ShortCAKE so that no 11-mer would be repeated on either strand can be used as a benchmark (Muñoz et al., 2014): assembly programs using a *k*-mer size above 11 should never have problems assembling this sequence in one piece, and indeed this was confirmed (i.e., even with SOAPdenovo2 no assembly consisted of two or more scaffolds, regardless of *k*).

The ABySS scenario of the generally smooth or monotone *k*-dependencies was, however, not always found when the SOAPdenovo2 pipeline was used. In particular, two of the metrics, assembly size and number of scaffolds, that gave smooth *k*-dependence for ABySS gave an accidented or discontinuous *k*-dependence at *k*=50 bp, and to some extent at *k*=100 bp, for SOAPdenovo2. The most striking of those metrics was the number of scaffolds in the assembly, which for Pb113 was only 4004 at *k*=49 bp but jumped up to 35,368 at *k*=51 bp. Although initially we did not rule out that biological size thresholds of repeats might be contributing to such a discontinuity in principle, the magnitude and sharpness of the transition (occurring from one allowed *k* value to the next) and the suspiciously round numbers of their locations (50 and 100) made such an explanation improbable in this case. When we plotted the lengths of the scaffolds, it became clear that almost all of the additional 31,364 sequences appearing in Pb113 at *k*=51 bp had lengths of 100 bp or very close to 100 bp (see Supplementary Fig. S3). It therefore appeared that some switch in the code was letting in vast quantities of 100 bp-sized sequences and considering them as part of the assembly when *k* exceeded 50 bp; if these corresponded to reads that could not be extended to longer sequences, it seemed reasonable that they would be largely repetitive DNA, thus explaining the parallel 'mouse'-shaped curves in the assembly size (Fig. 2) and repeats plots (Figs. 6 and 7), as well as the larger jumps found for the more repetitive, larger genomes. We looked briefly through the source code of the main program in the SOAPdenovo2 suite we had used, `SOAPdenovo-127mer`, but on a first pass we found only one possibly relevant if–then–else decision (in `loadGraph.c`); when

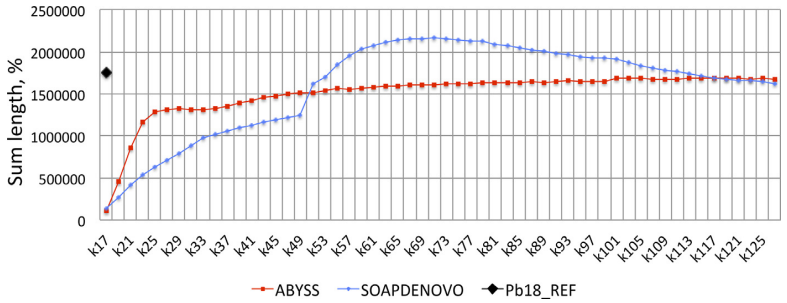


Fig. 7. Presence of repeats in scaffolds assembled using different *k*-mer values (17 bp ≤*k*≤ 127 bp) and programs, as illustrated by the total length of matches (repeats) reported by RepeatMasker in *P. brasiliensis* strain Pb113 assemblies (colored curves) and, for comparison, in the Pb18 reference assembly (black spot at far left). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Please cite this article in press as: Gallo, J.E., et al., The complex task of choosing a *de novo* assembly: Lessons from fungal genomes. Comput. Biol. Chem. (2014), http://dx.doi.org/10.1016/j.compbiolchem.2014.08.014

G Model
CBAC-6351; No. of Pages 11
ARTICLE IN PRESS

we changed it and recompiled, we saw that its effect was only small. We then reasoned that the KmerFreq and Corrector preprocessing steps, for which one cannot specify *k* values above 27 bp, might be involved in the jump, and re-ran SOAPdenovo2 again without these steps for *k*=49 and *k*=51. To our surprise, there was now a downwards jump instead of an upwards jump and, as expected from the lack of correction, the numbers of sequences in the assemblies were vastly bigger (1,517,856 sequences for the *k*=49 Pb113 assembly *vs.* 251,968 for the *k*=50 Pb113 assembly). Such checks would be compatible with an interpretation of the jump at *k*=50 bp as a consequence of what may be two alternate ways the SOAPdenovo2 pipeline effectively processes repeats (for $k \leq 49$ bp and $k \geq 51$ bp); such an interpretation would also be compatible with our finding that the jump was smaller for the more compact (less repetitive) genome of *S. cerevisiae* (Table 1) and absent for the synthetic chromosome that was engineered to be devoid of repeats. The resulting assembly, i.e., an outcome of grappling with the genome's repeats, could be quite sensitive to small changes in conditions such as program/parameter options, input reads and their quality, and the repeat landscape of the genome.

4.2. Recommendations and analyses from other groups

There are two questions that may rightly be asked at this point. First, how do other groups see the problem of choosing the 'right' *k*-mer size? Second, have discontinuities in profiles of assembly metrics or statistics versus *k*-mer choices been observed by other groups?

Since we used Illumina reads, and since we focused more on the details of SOAPdenovo2 than on other programs, we first mention their views regarding *k*-mer choice. A document with brief guidelines written by Illlumina, apparently with *E. coli* and perhaps other bacteria in mind Illumina (2010), states the following opinion:

> The right choice for *k* depends on coverage, read length, and error rates and is hard to determine in advance. Anecdotal recommendations indicate that the size of *k* should not be lower than half of the read length. If time allows, we recommend performing several assemblies over a small range of *k* and choosing the one that yields the best assembly for the desired application.

Here, there is a clear preference not to commit to any overly simplistic recipe for the right choice of *k*, although the document does mention an example that has been considered. We agree with Illumina's own recommendation, namely to perform several assemblies over a range of *k* values before making a choice, if possible. Several *k* values in a small range implies high resolution (closely spaced *k* values), and this seems best, as for example sharp discontinuities can go unnoticed if resolution is low, where users may intuitively assume some form of smooth or monotone interpolation. In this study we consider the possible advantage of a larger range of *k* to get a wider picture, but without reducing the resolution. We feel that with a modern laboratory setup, where a reasonably equipped server will be needed to run even a single assembly, the additional time needed to execute a loop of exploratory runs will, in many cases, be repaid by the benefit of a more informed decision.

Our results for fungal genomes presented here do not, however, suggest extrapolating the anecdotal rule of thumb to a general fungal context. For the reads of 101 bp, if the anecdotal $k \geq r/2$ criterion had been optimal, we would have needed to choose *k* values of at least 51 bp. For the reads of 150 bp, the optimal *k*-mer values according to the anecdotal recommendation would have been at 75 bp or higher. For SOAPdenovo2, then, the $k \geq r/2$ advice would have given us possibly suboptimal assemblies for the read sets we used, i.e., assemblies characterized by many small contigs rich in repeats and, for *P. brasiliensis*, assemblies much longer than the reference assembly, the ABySS assemblies, and the SOAPdenovo2 assemblies for lower *k* (see Fig. 2).

The publication of SOAPdenovo2 (Luo et al., 2012) also had comments on the topic of *k*-mer choice:

> important factor in the success of [de Bruijn graph]-based assembly is *k*-mer size selection. Using a large *k*-mer has the advantage of resolving more repeat regions; whereas, use of small *k*-mers is advantageous for assembling lowcoverage depth and removing sequencing errors.

In that paper's additional file the authors also mention that, where heterozygosity is present in a genome (unlike many fungi), it can be an additional factor that tends to reduce contig lengths, in particular where high *k*-mer sizes are used. Their conclusion: "For a complex genome, it is difficult to determine the optimal *k*-mer size based on theory." This view is also in line with ours, and one way of taking a more practical route is to profile at high resolution, and then try to explain, the behavior of various assembly and/or gene annotation metrics as *k* changes, in order to choose candidate *k*-mer sizes for a final assembly. For example, in our *P. brasiliensis* runs of SOAPdenovo2 for *k* just above 51 bp with our settings, there would be features of the assembly, such as many small contigs often containing repeats, that one might wish to address with additional filtering strategies.

A complementary approach, elected for SOAPdenovo2 (Luo et al., 2012) and by other groups in genomics and transcriptomics contexts (Samanta, 2012; Peng et al., 2012; Wences et al., 2013; Bankevich et al., 2012; Surget-Groba and Montoya-Burgos, 2010; Melicher et al., 2014), is to use *multiple* *k*-mer strategies: the idea is to merge (meta-assemble), and/or iteratively refine, assemblies obtained using different *k*-mers and/or assembly programs. Indeed, the 'one size fits all' assumption that is often made in practical assembly projects can be inherently problematic, as the choice of *k* can dramatically influence quality and contiguity of assemblies, and there may be different optimal *k* values for different regions of the same genome (Wences et al., 2013). Whether the strategy is multi-*k* or single *k*, high-resolution profiling such as we advocate here can assist in recognizing propitious ranges of the *k*-mer size(s).

Other authors have proposed criteria or methods for *k*-mer choice. For example, Chikhi and Medvedev (2014) propose an automatable method, which first generates abundance histograms for putative or candidate values of *k*, then fits a generative model to each histogram in order to estimate how many distinct *k*-mers in the histogram are genomic (i.e. error-free), and then picks the value of *k* which maximizes the number of genomic *k*-mers. They benchmark their method (tool `KmerGenie`) using Genome Assembly Gold-standard Evaluation data or GAGE (Salzberg et al., 2012). GAGE data have been used also in Luo et al. (2012) and, in the specific context of post-contig-assembly scaffolding, in Hunt et al. (2014).

4.3. Discontinuities

We now address the second question mentioned above, that of previous reports of discontinuities in profiles of metrics having *k* as the independent variable. In a search for metrics analyses that include eukaryotic organisms, and use NGS read lengths of $\geq$ 100 bp, we did not find any high-resolution *k*-dependency profiles of eukaryotes employing SOAPdenovo2 that had been obtained in a similar way to ours. The closest data that we could find, at a lower resolution than we used (steps of 10 bp versus 2 bp in our analyses), were in a table of results in Chikhi and Medvedev (2014), obtained via SOAPdenovo2 from 124-bp GAGE reads (Salzberg et al., 2012) for the common eastern bumble bee *Bombus impatiens*. That table profiles N50 and assembly size; assembly sizes were 224.1 Mb for *k*= 41 bp, 229.7 Mb for 51 bp, 230.4 for 61 bp, 226.1 Mb for 71 bp and

Please cite this article in press as: Gallo, J.E., et al., The complex task of choosing a *de novo* assembly: Lessons from fungal genomes. Comput. Biol. Chem. (2014), http://dx.doi.org/10.1016/j.compbiolchem.2014.08.014

G Model
CBAC-6351; No. of Pages 11
ARTICLE IN PRESS
J.E. Gallo et al. / Computational Biology and Chemistry xxx (2014) xxx–xxx

207.1 Mb for 81 bp. Clearly we cannot expect similar results to our unicellular fungi for the bumble bee genome, which is roughly an order of magnitude larger (and could not be successfully assembled with some other programs, including Velvet (Salzberg et al., 2012)). The low resolution also does not allow discontinuities to be seen. However, the rapid rise of assembly size from 41 bp to 51 bp, and subsequent slower rise and then decline, are qualitatively analogous to the shape of the SOAPdenovo2 profile we observed for *P. brasiliensis* and other fungi.

In analyses using other assembly programs/pipelines, we found very few *k*-dependence profiles with a high resolution and/or a suggestion of a discontinuity. An exception was a variable-resolution profiling (Samanta, 2012) for contigs of a ≈ 700 Mb genome assembled from 100 bp Illumina reads using Minia (Chikhi and Rizk (2012); see also Salikhov et al. (2013)): N50 had a jump or steep decrease and assembly size had a jump or steep increase (from 540 to 600 Mb). Both were located between 45, 47 and 49 bp on one side, and 55 bp and higher *k* on the other side; no assemblies were recorded for *k* values of 51 and 53 bp. The interpretation given was that successively larger segments are being assembled as *k* is increased (because large *k*-mers can distinguish better between slightly different small repeats), whereas with increasing *k* the de Bruijn graph gets increasingly fragmented (e.g., in the limit $k = r$ each unique read becomes a separate node unconnected to others). The position of the steep changes around *k*=50 bp was interpreted as half the read size. By contrast, in our different context, with a different assembler, the same position of the jumps remained fixed at *k*=50 (and *k*=100) bp regardless of read size.

Finally, discontinuities sometimes have easily explainable technical causes, and can affect internal metrics; an example that is unrelated to our observations is one of *E. coli* assembly profiles obtained for Minia and two related programs (2-bloom and 4-bloom), in which three internal statistics (a traversal time, a construction time and a structure size expressed in bits/*k*-mer) jumped at *k*=32 bp because of a switch from 64-bit to 128-bit representation of *k*-mers (Salikhov et al., 2013).

In summary, we are still not sure what caused the systematic discontinuities we observed in our fungal profiles. We posted the question for SOAPdenovo on seqanswers.com but no answers were proposed. Three broad categories of explanation might be: known technical reasons, previously undocumented anomalies or artifacts (possibly occurring only for certain program(s), settings, taxa, error levels, etc.), or biological reasons. Although we consider the last category unlikely as a sole determinant of the jump position(s) on the *k* axis, it could contribute.

4.4. Comments on genome complexity and repeats

The term 'complexity' as applied to DNA sequences is often encountered by molecular biologists in contexts such as BLAST searches on a server at NCBI, in which one can decide whether or not to mask out 'low complexity' regions. These regions have often very low or very high GC, so that the 4-letter alphabet of DNA almost collapses to a 2-letter alphabet (A and T, or else G and C), which in turn makes chance repeats or chance matches more probable than in a fully used 4-letter alphabet. Similarly, in textbooks (Lewin, 1999), monographs (Davidson, 1976, p. 189 ff.) and articles since the first DNA reassociation work of Britten and his colleagues in the 1960s (Britten and Kohne, 1968), a genome of a given size is traditionally considered most complex if it contains very few repeats, and less complex if it has more repeats. Thus, repetitiveness can have a connotation of lowering complexity.

If, however, one approaches the issue from another point of view, namely the task of assembling a genome from short reads *de novo*, then the connotation is quite different. A genome without any perfect repeats of a given size on either strand (i.e., a double-stranded de Bruijn sequence) can be generated (Fraenkel and Gillis, 1966; Orenstein and Shamir, 2013), and is the simplest conceivable genome of its length, for the short-read assembly task. From this viewpoint, adding repeats to such a genome can make a typical *de novo* or reference assembly task more difficult, but never easier, so in this sense it *increases* the complexity of the genome (see Results, and examples given in Muñoz et al. (2014)).

Not only genome assemblers can face tasks that become more complex when repeats in the genome are abundant or long. The nucleus or mitochondrion of the eukaryotic cell (or selection) may face more complex tasks of genome integrity/stability maintenance or 'ambiguity management' when its genome is repeat-rich. For example, the repeats can recombine illegitimately, which can lead to deletion mutants such as the petite mitochondrial genomes of baker's yeast, which generate petite (anaerobic-only) colonies that can be distinguished with the naked eye (see Bernardi (2004) and refs. therein). In human, illegitimate recombination can also lead to disease, for example when an important region of a chromosome is looped out in the process and deleted; in some instances inappropriate repetition of sequences could be of much importance for the fitness of the organism (Ahmed et al., 1999; Abrusan and Krambeck, 2006).

4.5. Toward a first-pass evaluation methodology for new releases of assembly programs

Advice for choosing a *k* value sometimes considers assembly programs as straightforward and transparent implementations of the fundamental algorithms that they use, which can be elegant, or simple to describe or portray. Such algorithms include basic steps involving the use of de Bruijn graphs that are shared, in one form or another, by several modern assembly programs currently in use.

In the present study, we have approached the question from a very different point of view, not considering primarily what assembly programs based on de Bruijn graph principles should do, but observing what they do in real practice, when one uses them directly 'out of the box' and applies them as specified, to small genomes of fungi having a moderate amount of complexity (as assessed by a moderate, but notable, content of repetitive DNA; see previous section). A graphical overview, and a first insight into dominant processes of the programs or pipelines, is gained already by visualizing how rough 'presence' metrics of assembly success vary if one varies *k*, and possibly also when one modifies key experimental parameters such as *r* and *l*. The 'presence' metrics we used here roughly estimate the presence of a genome's DNA, its genes, and its expected types of repeats. Also relevant, although not included in this limited study, would be the variation with *k* of dedicated 'accuracy' metrics that gauge the presence of small errors in assembly sequences. Such rare, small errors might, despite their scarcity, occasionally cause frameshift errors where they appear in coding regions, and/or readthrough errors at exon boundaries or stops, which even if infrequent would be important to take into account, if their occurrence were to vary systematically with *k* in the range considered in practice (>20 bp). There might also occasionally exist a collapse of reads from similar coding sequences of two close paralogs onto one. If such rare errors affecting the faithful portrayal of the gene space were to occur preferentially in some intervals along the *k* axis, it would be good to consider them in order to make an informed choice of *k*.

Sometimes it is helpful to conceptually separate the central and often elegant algorithmic core(s) of a bioinformatic program from its other, more peripheral parts, which can occasionally encode relatively *ad hoc* decisions, peripheral algorithms (e.g., in our case possibly algorithms not related to de Bruijn graph methods), or heuristic expedients. Such additional parts might sometimes even

Please cite this article in press as: Gallo, J.E., et al., The complex task of choosing a *de novo* assembly: Lessons from fungal genomes. Comput. Biol. Chem. (2014), http://dx.doi.org/10.1016/j.compbiolchem.2014.08.014

be responsible for superior performance compared to other programs using the same algorithmic core. Some of those parts may be carefully planned; in others, a decision, although reasonably made, may be to some extent arbitrary, or may not be crucial in the original context for which the program was intended. As a program evolves complexity, and the scenarios for which it is used expand, it may be difficult to keep track of such past decisions and their possible effects. In the more general context of re-testing evolving software, (Brooks, 1995, p. 6) writes: "the number of cases [to test] grows combinatorially. It is time-consuming, for subtle bugs [or anomalies, or unexpected effects can] arise from unexpected interactions of debugged components." In some bioinformatic programs, *ad hoc* or arbitrary decisions can also form part of the main algorithm, as in some widely used progressive multiple alignment programs, which include the CLUSTAL series; of such programs, (Felsenstein, 2004, p. 500) notes that "alignments are combined using some rather arbitrary rules", and (Durbin et al., 1998, p. 148) write that "CLUSTALW is unabashedly *ad hoc* in its alignment construction and scoring stage". In this study, we have addressed the profiling of assembly programs' behaviors for a given genome as k is varied, in a way that monitors not only the algorithmic core(s) but also more peripheral parts of the programs.

We envisage a possible future tradition for first-pass performance assessment of genome assembly software, which would not imply extensive testing whenever a major new release is offered. Instead, and in line with benchmarking that we have previously proposed (Muñoz et al., 2014), evaluation (internal or external) could be standardized by using performance plots or curves such as those we sketch here, and using a set of standard test genomes that include eukaryotes; because of the limited size of the eukaryotic genomes of unicellular fungi, we propose that they would be good candidates to include. It could be particularly useful to include fungi having well-characterized genomes that are not as compact (repeat-poor) as the nuclear genome of baker's yeast, of which *P. brasiliensis* is an example.

Acknowledgements

We thank Wentian Li for sharing manuscripts on closely related topics prior to publication, for literature and software references, and for discussions, and Christina Cuomo and the members of the Broad Institute's Genome Sequencing and Analysis Program for discussions and advice, and for the use of unpublished *Emmonsia* read sets obtained in collaboration. We also thank two referees for suggestions and references that helped us to improve the paper, and Dr. Manoj Samanta (Systemix Institute) for a clarification and references. This work was partly supported by Colciencias grant 1222-56934875, "A gene atlas for human pathogenic fungi", and by a sustainability grant from the Universidad de Antioquia, Medellín. Colciencias National Doctorate Program funding supported J.F.M., and the Universidad del Rosario partly supported J.E.G. via a Ph.D. scholarship.

Appendix A. Supplementary Data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.compbiolchem.2014.08.014.

References

Abrusan, G., Krambeck, H.J., 2006. The distribution of L1 and Alu retroelements in relation to GC content on human sex chromosomes is consistent with the ectopic recombination model. J. Mol. Evol. 63, 484–492.
Ahmed, S., Clay, O.K., Schaffner, W., 1999. Proto-oncogenes, unlike 'harmless' genes, tend to be dispersed in the human genome: selection against out-of-register recombination? Biol. Chem. 380, 3–5.
Bankevich, A., Nurk, S., Antipov, D., Gurevich, A., Dvorkin, M., et al., 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 19, 455–477.
Bassett, D.E., Boguski, M.S., Hieter, P., 1996. Yeast genes and human disease. Nature 379, 589–590.
Berge, C., 2001. The Theory of Graphs. Dover, Mineola, NY (reprint of 1962 edition).
Bernardi, G., 2004. Structural and Evolutionary Genomics: Natural selection in genome evolution. Elsevier, Amsterdam, etc.
Bradnam, K.R., Fass, J.N., Alexandrov, A., Baranay, P., Bechner, M., Birol, I., et al., 2013. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. Gigascience 2 (1), 10.
Britten, R.J., Kohne, D.E., 1968. Repeated sequences in DNA. Science 161, 529–540.
Brooks, F.P., 1995. The Mythical Man-Month: Essays on Software Engineering, with four new chapters, Anniversary Edition. Addison-Wesley, Reading, MA.
Carels, N., Barakat, A., Bernardi, G., 1995. The gene distribution of the maize genome. Proc. Natl. Acad. Sci. U. S. A. 92, 11057–11060.
Chikhi, R., Medvedev, P., 2014. Informed and automated k-mer size selection for genome assembly. Bioinformatics 30, 31–37.
Chikhi, R., Rizk, G., 2012. Space-efficient and exact de Bruijn graph representation based on a Bloom filter. In: WABI. Vol. 7534 of Lecture Notes in Computer Science. Springer, Berlin and Heidelberg, pp. 236–248.
Compeau, P.E.C., Pevzner, P.A., Tesler, G., 2011. How to apply de Bruijn graphs to genome assembly. Nat. Biotechnol. 29, 987–991.
Davidson, E.H., 1976. Gene activity in early development, 2nd Edition. Academic Press, New York.
Deng, A., Wu, Y., 2005. De Bruijn digraphs and affine transformations. Eur. J. Comb. 26, 1191–1206.
Desjardins, C., Champion, M., Holder, J., Muszewska, A., Goldberg, J., et al., 2011. Comparative genomic analysis of human fungal pathogens causing paracoccidioidomycosis. PLoS Genet. 7, e1002345.
Drummond, M.F., Sculpher, M.J., Torrance, G.W., O'Brien, B.J., Stoddart, G.L., 2005. Methods for the economic evaluation of health care programmes, 3rd Edition. Oxford University Press, Oxford and New York.
Durbin, R., Eddy, S., Krogh, A., Mitchison, G., 1998. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press, Cambridge.
Felsenstein, J., 2004. Inferring phylogenies. Sinauer, Sunderland, MA.
Foury, F., 1997. Human genetic diseases: a cross-talk between man and yeast. Gene 195, 1–10.
Fraenkel, A., Gillis, J., 1966. Proof that sequences of A, C, G, and T can be assembled to produce chains of ultimate length avoiding repetitions everywhere. Prog. Nucleic Acid. Res. Mol. Biol. 5, 343–348.
Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., et al., 1996. Life with 6000 genes. Science 274, 563–567.
Haznedaroglu, B.Z., Reeves, D., Rismani-Yazdi, H., Peccia, J., 2012. Optimization of de novo transcriptome assembly from high-throughput short read sequencing data improves functional annotation for non-model organisms. BMC Bioinform. 13, 170.
Hunt, M., Newbold, C., Berriman, M., Otto, T., 2014. A comprehensive evaluation of assembly scaffolding tools. Genome Biol. 3 (1), R42.
Illumina, 2010. De novo assembly using Illumina reads, Illumina, http://www.res.illumina.com/documents/products/technotes/technote_denovo_assembly_ecoli.pdf.
Keller, O., Kollmar, M., Stanke, M., Waack, S., 2011. A novel hybrid gene prediction method employing protein multiple sequence alignments. Bioinformatics 27, 757–763.
Lewin, B., 1999. Genes VII. Oxford University Press, Oxford, etc.
Li, W., 1997. The complexity of DNA. Complexity 3, 33–38.
Li, W., Freudenberg, J., Miramontes, P., 2014. Diminishing return for increased mappability with longer sequencing reads: implications of the k-mer distributions in the human genome. BMC Bioinform. 15, 2.
Li, W., Stolovitzky, G., Bernaola-Galván, P., Oliver, J., 1998. Compositional heterogeneity within, and uniformity between, DNA sequences of yeast chromosomes. Genome Res. 8, 916–928.
Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., et al., 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience 1 (1), 18.
MacCallum, I., Przybylski, D., Gnerre, S., Burton, J., Shlyakhter, I., Gnirke, A., Malek, J., McKernan, K., Ranade, S., Shea, T.P., Williams, L., Young, S., Nusbaum, C., Jaffe, D.B., 2009. ALLPATHS 2: small genomes assembled accurately and with high continuity from short paired reads. Genome Biol. 10 (10), R103.
Medvedev, P., Brudno, M., 2009. Maximum likelihood genome assembly. J. Comput. Biol. 16, 1101–1116.
Melicher, D., Torson, A.S., Dworkin, I., Bowsher, J., 2014. A pipeline for the de novo assembly of the Themira biloba (Sepsidae: Diptera) transcriptome using a multiple k-mer length approach. BMC Genomics 15, 188.
Muñoz, J.F., Misas, E., Gallo, J.E., McEwen, J.G., Clay, O.K., 2014. Limits to sequencing and de novo assembly: classic benchmark sequences for optimizing fungal NGS design. Adv. Intell. Syst. Comput. 232, 221–230.
Orenstein, Y., Shamir, R., 2013. Design of shortest double-stranded DNA sequences covering all k-mers with applications to protein-binding microarrays and synthetic enhancers. Bioinformatics 29, i71–i79.
Parra, G., Bradnam, K., Korf, I., 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 23, 1061–1067.
Parra, G., Bradnam, K., Ning, Z., Keane, T., Korf, I., 2009. Assessing the gene space in draft genomes. Nucleic Acids Res. 37, 289–297.

Please cite this article in press as: Gallo, J.E., et al., The complex task of choosing a *de novo* assembly: Lessons from fungal genomes. Comput. Biol. Chem. (2014), http://dx.doi.org/10.1016/j.compbiolchem.2014.08.014

Peng, Y., Leung, H., Yiu, S., Chin, F., 2012. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics 28, 1420–1428.
Pennisi, E., 2014. DNA sequencers still waiting for the nanopore revolution. Science 343, 829–830.
Román-Roldán, R., Bernaola-Galván, P., Oliver, J., 1998. Sequence compositional complexity of DNA through an entropic segmentation algorithm. Phys. Rev. Lett. 80, 1344–1347.
Salikhov, K., Sacomoto, G., Kucherov, G., 2013. Using cascading Bloom filters to improve the memory usage for de Brujin graphs. In: Algorithms in Bioinformatics. Springer, Berlin and Heidelberg, pp. 364–376.
Salzberg, S., Phillippy, A., Zimin, A., Puiu, D., Magoc, T., et al., 2012. GAGE: a critical evaluation of genome assemblies and assembly algorithms. Genome Res. 22, 557–567.
Samanta, M., October 2012. From multiple kmers to multi-kmer de Bruijn graph. Systemix Institute, Redmond, WA http://www.homolog.us/blogs/blog/2012/10/10/multi-kmer-de-bruijn-graphs/
Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J., Birol, I., 2009. ABySS: a parallel assembler for short read sequence data. Genome Res. 19, 1117–1123.
Surget-Groba, Y., Montoya-Burgos, J., 2010. Optimization of de novo transcriptome assembly from next-generation sequencing data. Genome Res. 20, 1432–1440.
Wences, A., Baranay, P., Schatz, M., 2013. De novo genome metassembly. Biology of Genomes. Cold Spring Harbor, May 7–11 2013, schatzlab.cshl.edu/publications/posters/2013.BOG.Metassembly.pdf.
Zerbino, D.R., Birney, E., 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18, 821–829.

Please cite this article in press as: Gallo, J.E., et al., The complex task of choosing a *de novo* assembly: Lessons from fungal genomes. Comput. Biol. Chem. (2014), http://dx.doi.org/10.1016/j.compbiolchem.2014.08.014

Chapter 5

Genome update of the dimorphic human pathogenic fungi causing Paracoccidiodomycosis

OPEN ACCESS Freely available online

PLOS | NEGLECTED TROPICAL DISEASES

Genome Update of the Dimorphic Human Pathogenic Fungi Causing Paracoccidioidomycosis

José F. Muñoz[1,2☯], Juan E. Gallo[1,3☯], Elizabeth Misas[1,2], Margaret Priest[4], Alma Imamovic[4], Sarah Young[4], Qiandong Zeng[4], Oliver K. Clay[1,5], Juan G. McEwen[1,6], Christina A. Cuomo[4]*

1 Cellular and Molecular Biology Unit, Corporación para Investigaciones Biológicas, Medellín, Colombia, **2** Institute of Biology, Universidad de Antioquia, Medellín, Colombia, **3** Doctoral Program in Biomedical Sciences, Universidad del Rosario, Bogotá, Colombia, **4** Broad Institute of MIT and Harvard, Cambridge, Massachusetts, United States of America, **5** School of Medicine and Health Sciences, Universidad del Rosario, Bogotá, Colombia, **6** School of Medicine, Universidad de Antioquia, Medellín, Colombia

Abstract

Paracoccidiodomycosis (PCM) is a clinically important fungal disease that can acquire serious systemic forms and is caused by the thermodimorphic fungal *Paracoccidioides* spp. PCM is a tropical disease that is endemic in Latin America, where up to ten million people are infected; 80% of reported cases occur in Brazil, followed by Colombia and Venezuela. To enable genomic studies and to better characterize the pathogenesis of this dimorphic fungus, two reference strains of *P. brasiliensis* (Pb03, Pb18) and one strain of *P. lutzii* (Pb01) were sequenced [1]. While the initial draft assemblies were accurate in large scale structure and had high overall base quality, the sequences had frequent small scale defects such as poor quality stretches, unknown bases (N's), and artifactual deletions or nucleotide duplications, all of which caused larger scale errors in predicted gene structures. Since assembly consensus errors can now be addressed using next generation sequencing (NGS) in combination with recent methods allowing systematic assembly improvement, we re-sequenced the three reference strains of *Paracoccidioides* spp. using Illumina technology. We utilized the high sequencing depth to re-evaluate and improve the original assemblies generated from Sanger sequence reads, and obtained more complete and accurate reference assemblies. The new assemblies led to improved transcript predictions for the vast majority of genes of these reference strains, and often substantially corrected gene structures. These include several genes that are central to virulence or expressed during the pathogenic yeast stage in *Paracoccidioides* and other fungi, such as *HSP90*, *RYP1-3*, *BAD1*, catalase B, alpha-1,3-glucan synthase and the beta glucan synthase target gene *FKS1*. The improvement and validation of these reference sequences will now allow more accurate genome-based analyses. To our knowledge, this is one of the first reports of a fully automated and quality-assessed upgrade of a genome assembly and annotation for a non-model fungus.

Citation: Muñoz JF, Gallo JE, Misas E, Priest M, Imamovic A, et al. (2014) Genome Update of the Dimorphic Human Pathogenic Fungi Causing Paracoccidioidomycosis. PLoS Negl Trop Dis 8(12): e3348. doi:10.1371/journal.pntd.0003348

Editor: Joseph M. Vinetz, University of California, San Diego, School of Medicine, United States of America

Received August 21, 2014; **Accepted** October 14, 2014; **Published** December 4, 2014

Data Availability: The authors confirm that all data underlying the findings are fully available without restriction. All updated genomes and annotations are available from GenBank (accession numbers: *Paracoccidioides lutzii* Pb01 (ABKH00000000), *Paracoccidioides brasiliensis* Pb03 (ABHV00000000), and *Paracoccidioides brasiliensis* Pb18 (ABKI00000000).

Funding: This project has been funded in whole or in part with Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Contract No.: HHSN272200900018C. This work was partly supported by Colciencias via the grants "Comparative genomics and virulence in the pathogenic fungus *Paracoccidioides brasiliensis*" 2213-48925460 and "A gene atlas for human pathogenic fungi" 1222-56934875, and by the Universidad de Antioquia via a grant "Sostenibilidad 2013/14". Colciencias National Doctorate Program funding supported JFM; Enlaza Mundos partly supported his fellowship. The Universidad del Rosario partly supported JEG. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* Email: cuomo@broadinstitute.org

☯ These authors contributed equally to this work.

Introduction

Paracoccidioides spp. is a thermally dimorphic pathogenic fungus that causes paracoccidioidomycosis (PCM), a neglected health-threatening human systemic mycosis endemic to Latin America where up to ten million people are infected. Disease can progress slowly, with roughly five new cases of disease per million infected individuals per year, with a male to female ratio of 13 to 1. About 80% of PCM cases occur in Brazil, followed by Colombia and Venezuela [2].

Within the *Paracoccidioides* genus, the three characterized phylogenetic lineages of *P. brasiliensis* (PS2, PS3, S1) and the one characterized lineage of *P. lutzii* (Pb01-like) can infect humans, and these groups can vary in virulence and induce different immune responses by the host [3,4]. To better understand the pathogenesis and to enable genomics-based studies, the genomes of *Paracoccidioides* spp. were sequenced, analyzed and made publicly available in 2011 [1]. The Broad Institute of MIT and Harvard in partnership with the *Paracoccidioides* research community selected three reference isolates for sequencing and genomic analysis; assembly size for these strains varied between 29.1 and 32.9 Mb, and between 7,875 and 9,132 genes were identified in each strain [1]. These included two strains of *P. brasiliensis* (Pb18 representing the S1 lineage and Pb03 representing the PS2 lineage) and one strain of *P. lutzii* (Pb01) [1].

Author Summary

The fungal genus *Paracoccidioides* is the causal agent of paracoccidioidomycosis (PCM), a neglected tropical disease that is endemic in several countries of South America. *Paracoccidioides* is a pathogenic dimorphic fungus that is capable of converting to a virulent yeast form after inhalation by the host. Therefore the molecular biology of the switch to the yeast phase is of particular interest for understanding the virulence of this and other human pathogenic fungi, and ultimately for reducing the morbidity and mortality caused by such fungal infections. We here present the strategy and methods we used to update and improve accuracy of three reference genome sequences of *Paracoccidioides* spp. utilizing state-of-the-art Illumina re-sequencing, assembly improvement, re-annotation, and quality assessment. The resulting improved genome resource should be of wide use not solely for advancing research on the genetics and molecular biology of *Paracoccidioides* and the closely related pathogenic species *Histoplasma* and *Blastomyces*, but also for fungal diagnostics based on sequencing or molecular assays, characterizing rapidly changing proteins that may be involved in virulence, SNP-based population analyses and other tasks that require high sequence accuracy. The genome update and underlying strategy and methods also serve as a proof of principle that could encourage similar improvements of other draft genomes.

These sequenced isolates are extensively referenced in molecular biology and experimental mycology laboratories working with *Paracoccidioides* spp. and also other pathogenic fungi, including those working with yeast phase specific genes expressed during host infection. These sequences also serve as a reference to analyze high-throughput data increasingly generated by genomic, metagenomic, transcriptomic and proteomic approaches. Additionally, accurate sequences are critical for evolutionary analyses, e.g., to identify positively selected genes, as well as to provide new targets for the design of diagnostic assays.

The *P. brasiliensis* Pb18 and Pb03 strains and the *P. lutzii* Pb01 strain were sequenced using the sequencing technology and computational methods available at the time, which produced high quality draft assemblies. However, the assemblies included a large number of gaps and uncertain or low quality nucleotides in the final consensus sequences. Also, the annotation pipelines flagged only the most extreme annotation errors for curation and did not address the larger number of smaller scale errors in the gene models and underlying sequence [5]. Correction of such errors requires re-evaluation of the assembly consensus sequence and associated annotation.

Assembly errors that could not be detected in previous data and passed standard quality control criteria at that time can now be corrected using next generation sequencing (NGS) for systematic assembly improvement. These include errors in gene-containing regions of the original genomic assembly affected by poor quality sequence or ambiguities, which can cause incorrect gene structure predictions. Since predicted genes of reference genomes are now frequently used for homology-based inference or confirmation of gene structures in closely related species, errors in the reference sequence may be propagated to other genomes [6,7]. Therefore, systematic improvement of a genome assembly and annotation can impact not just the understanding of that particular species, but also that of other related species for which it is used as a reference for comparison.

Here, we present an update of the three *Paracoccidioides* reference genome sequences achieved using Illumina re-sequencing to correct assembly errors and document the improvements obtained. The improved and updated reference genome assemblies and annotations of this important human fungal pathogen now allow more accurate SNP analyses, genome-wide evolutionary (e.g., selection) analyses that depend on high-quality sequences, phylogenetic footprinting studies of regulatory regions, or primer and probe design for diagnostic assays.

Methods

Paracoccidioides reference strains and previous sequencing

Three reference isolates of *Paracoccidioides* spp. (Pb01, Pb03 and Pb18), representing two species, were previously sequenced. The isolate of *P. lutzii* (Pb01) was a clinical isolate originating from an acute form of paracoccidioidomycosis (PCM) in an adult male. The two *P. brasiliensis* isolates were from individuals presenting chronic PCM; Pb03 represents the PS2 phylogenetic group and Pb18 the S1 group [1,4]. In a partnership between the Broad Institute and the *Paracoccidioides* research community, these genomes were previously sequenced using multiple whole genome shotgun libraries constructed from genomic DNA for each strain; paired-end sequences were generated for each with Sanger technology and assembled using Arachne [1] (assembly v1; S1 Table).

Re-sequencing of *Paracoccidioides* reference strains

The reference strains Pb01 (previously sequenced DNA sample) and Pb03 and Pb18 (newly extracted DNA samples) were re-sequenced using Illumina technology. For library construction, 100 ng of genomic DNA was sheared to ~250 bp using a Covaris LE instrument and prepared for sequencing as previously described [8]. A library for each of the three samples was used to generate 101 base paired-end reads on the Illumina HiSeq2000 platform, producing an average genome coverage of 165X.

Assembly improvement using Pilon

To improve the genome sequence of *Paracoccidioides* spp. strains Pb01, Pb03 and Pb18, Illumina paired-end reads were aligned to the draft reference assemblies (assembly v1) using BWA version 0.5.9 with default settings [9]. The assembly consensus sequence was re-evaluated by providing these alignments as input to the automated assembly improvement program Pilon (version 1.4, default parameters, www.broadinstitute.org/software/pilon/). Pilon uses the Illumina read alignments for multiple classes of assembly correction. First, Pilon scans the read alignments for positions where the sequencing data disagree with the input genome (assembly v1) and corrects small errors such as single nucleotide differences and small insertion/deletion events. Second, Pilon looks for coverage and alignment discrepancies to identify potential mis-assemblies and larger variants. Finally, Pilon uses reads anchored adjacent to discrepant regions and gaps in the input genome to reassemble the region, attempting to fill in the true sequence including large insertions. As output, Pilon provides the sequence of this improved genome assembly (assembly v2; Fig. 1) along with files summarizing the changes and quality measures used in the assessment.

Gene prediction and annotation

Protein-coding genes were predicted in the improved assemblies (assembly v2) using a combination of gene models from the prediction programs Augustus [10], Genemark-ES [11],

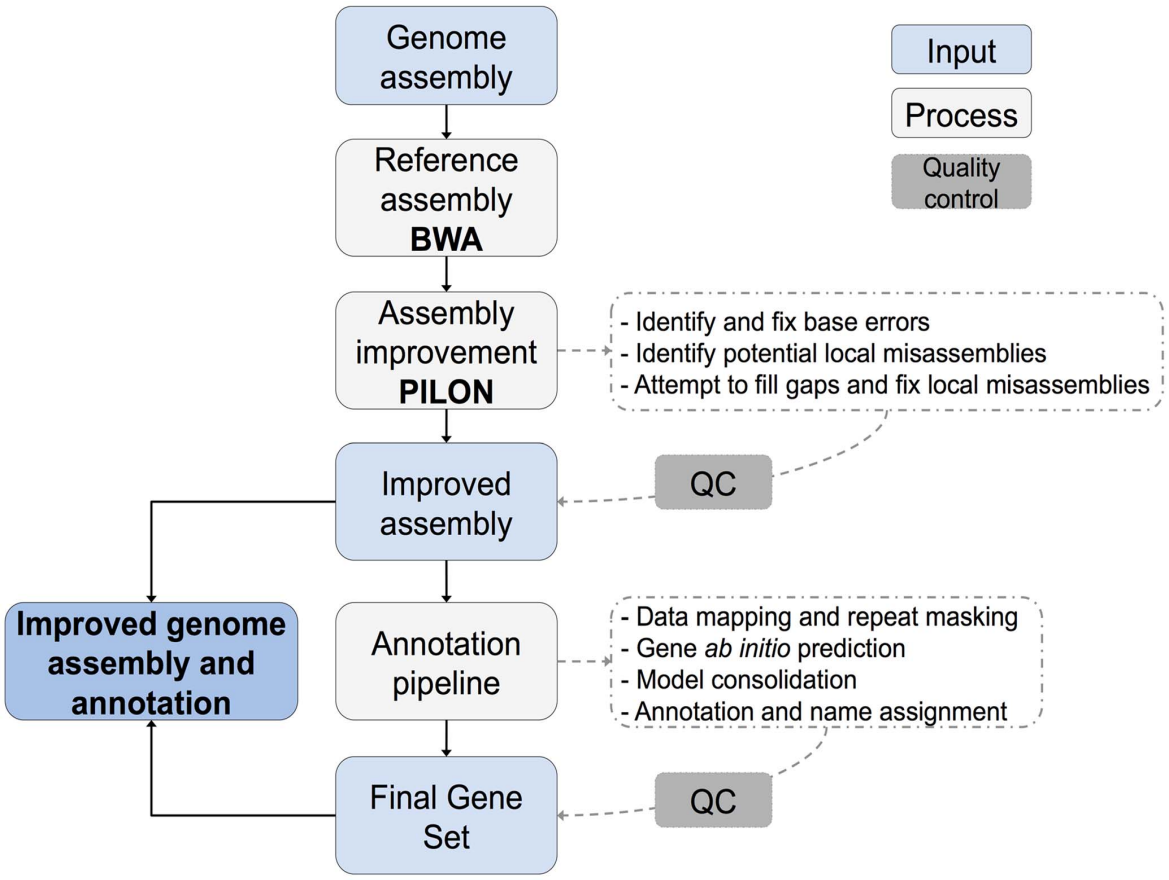


Fig. 1. Overview of genome assembly and annotation improvement process.
doi:10.1371/journal.pntd.0003348.g001

GlimmerHMM [12], Genewise [13], and Snap [14], as well as automated revision based on EST data (e.g., from [15]) and manual gene revision of flagged calls. The predicted gene sets were then provided as input to EvidenceModeler (EVM) [16] to obtain the best consensus model for a given locus. The consistency of the gene models was evaluated by examining alignments of protein orthology groups identified using OrthoMCL [17]. EVMLite was used to rescue orphan genes not captured in EVM; only those genes with additional evidence such as overlap to Genewise or non-repeat HMMER3 PFAM domains were rescued, as well as non-redundant genes overlapping the OrthoMCL genes in clusters containing 2 or more genomes. Lastly spurious gene models matching repetitive or low-complexity sequences were removed.

For each *Paracoccidioides* genome, we compared the original annotation (v1) with the updated annotation (v2) to evaluate the changes in the new gene sets. To precisely characterize the types of changes across the v1 and v2 annotations, we first mapped the corresponding gene between the two assemblies. The v1 and v2 assemblies were aligned using nucmer [18], and the alignment coordinates were used to assign gene correspondence between the initial annotation v1 and the new annotation v2. This mapping also allowed us to preserve locus numbers in the updated gene set. Each annotated gene was assigned a locus number, keeping where appropriate the previous locus number of the form PAAG_##### (Pb01), PABG_##### (Pb03) or PADG_##### (Pb18), which serves as a unique identifier within each genome and across assemblies. New genes, merged genes, and genes with large structure and sequence changes in transcripts were assigned new and unique locus numbers following the last locus number of annotation v1. Locus numbers of deleted genes do not appear in the final gene sets.

Evaluation of gene annotation improvements between *Paracoccidioides* strains

To evaluate whether the changes in gene sequence and structure produced a more accurate gene set, the gene sequences of annotation v1 and of the updated annotation v2 were compared via sequence similarity and orthology analysis. To evaluate the consistency of gene structures for orthologs as well as their conservation between species, OrthoMCL version 1.4 with a Markov inflation index of 1.5 and a maximum e-value of 1e-5 was used to identify orthologous clusters across the six total protein sets corresponding to annotations v1 and v2 of each *Paracoccidioides* strain. For each cluster group representing putative orthologs, we compared the maximum length difference among the three *Paracoccidioides* genes in annotation v2 to that in the annotation v1.

To compare the functional content of the v1 and v2 gene sets, we evaluated both protein domain families (PFAM) and pathway information (KEGG). Using HMMER3 [19], we mapped v27 of the PFAM domain database [20] to both the v1 and v2 gene sets. KEGG domains [21] from release 65 were also mapped to both gene sets using BLAST.

To evaluate changes in the gene structure, the corresponding transcripts from annotation v1 and v2 were identified as described above. We also aligned gene sets v1 and v2 using BLASTn [22] version 2.2.28+ with default parameters, using an in house Perl script to determine the types of modification for each gene, which included changes in gene length, gene coverage and percent nucleotide identity. We manually checked a random gene sample of each type of change (up to 10 genes) from both gene correspondence and BLAST analyses to verify that changes in gene set v2 were actually gene improvements. To evaluate changes in the coding regions of genes of high interest to the community, we selected known specific yeast-phase genes or virulence factors of *Paracoccidioides* spp., as well as other genes that are generally considered relevant for research on *Paracoccidioides* or related dimorphic pathogens, for manual review. The sequences of these genes' coding regions were aligned at the protein level with CLUSTALW [23] version 2.1, using both the v1 and v2 annotations.

Gene annotation improvements using genes from CEGMA and from related dimorphic pathogenic fungi

The coverage of Core Eukaryotic Genes defined by CEGMA [24] was evaluated using the CoreAlyze tool (http://sourceforge.net/projects/corealyze/) to summarize results for all the v1 and v2 gene sets. BLASTp version 2.2.28+ was run with default settings using protein sets from annotations v1 and v2 as the database, with *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe* CEGMA proteins as the query. We also included the protein gene sets of two close relatives of *Paracoccidioides*, the dimorphic fungal pathogens *Blastomyces dermatitidis* and *Histoplasma capsulatum*. In order to obtain a detailed picture of the changes where gene annotations were modified but not completely overridden, we compared protein sequences between the two versions, excluding proteins that were added or deleted from the final gene set v2, as well as proteins for which the new annotation was for a completely different transcript at the same locus. For a hit to be counted, the protein needed to match a protein in the reference set with at least 75% identity for the v1 and v2 annotations. This percent identity cutoff was determined empirically to eliminate spurious low similarity alignments. The percent identity and the bit score between the query protein and each version of the *Paracoccidioides* annotations (v1 and v2) were compared.

Results

Genome resequencing with NGS technology

The strains of the genomes of *Paracoccidioides* spp. previously sequenced [1], Pb18 and Pb03 and Pb01, were re-sequenced using Illumina 101 bp paired-end reads. This sequencing generated 93.6 million reads for Pb18 with an average coverage of 198X, 124.2 million reads for Pb03 with an average coverage of 150X and 110.0 million reads for Pb01 with an average coverage of 148X. This high coverage sequence data was then used to refine the consensus sequence of the original assembly by assessing differences between the new sequence and the previous assemblies. This can target a wide range of improvements, including correcting base calls, resolving ambiguous bases and closing gaps within scaffolds.

Genome assembly improvement using Pilon

Fig. 1 shows a simplified overview of the workflow of genome improvement. The new Illumina data were used to systematically improve the three *Paracoccidioides* spp. assemblies using Pilon (http://www.broadinstitute.org/software/pilon/). Pilon bases its improvement calls on an alignment of the reference genome and the sequenced reads. The aligned bases and depth at each sequenced position provides evidence for the reference base or for an alternative; where changes are supported they can result in single base differences, insertion or deletion of single bases or larger regions, identification of collapsed regions and more complex changes and gap filling based on local reassembly. Reads of each of the genomes of Pb18, Pb03 and Pb01 were aligned to the corresponding reference assembly using BWA [9] and the resulting bam file was used as input for Pilon.

In each of the *Paracoccidioides* assemblies, Pilon identified and fixed base errors in the consensus sequence. The statistical improvements for the assemblies v2 of *Paracoccidioides* spp. are summarized in Table 1. The most frequent class of changes was single base substitutions, identified as single nucleotide polymorphisms (SNPs) between the assembly and reads. Between 3,018 and 3,290 single base errors were corrected in each assembly. Small insertions and deletions were also incorporated into each assembly. The major classes of changes can be attributed to bases added and removed in reassembly fixes, collapsed bases in the new assembly and the closing of gaps (Table 1). Regions of misassembly identified and fixed by Pilon resulted in bases added or removed but no new gaps opened. Across all the assemblies, 20% of all initial gaps were closed by Pilon; the number of gaps closed were 113, 56 and 212, for Pb18, Pb03 and Pb01, respectively. Overall, the assembly improvement process led to an increase of contig N50 for all strains. About ~99% of low quality nucleotides in assemblies v1 were well supported or fixed with high flag coverage in assemblies v2.

Overall, the *P. lutzii* Pb01 genome assembly was most substantially improved, based on comparing statistics for all v1 and v2 assemblies (S1 Table). The contig N50 for Pb01 v2 increased by 29.1 kb; more bases were added and removed after re-assembly fixes and more gaps were closed than in the other two genomes. The genome size and number of scaffolds of Pb18 and Pb03 were essentially unchanged. The Pb01 genome size decreased slightly from 32.94 to 32.93 Mb; the updated assembly contains one scaffold fewer, as two scaffolds were merged by closing the gap between them. The number of contigs was reduced in all three strains, which considering the increase in the contig N50 indicates that the assemblies v2 for Pb18, Pb03 and Pb01 were less fragmented. All these changes indicate that the genome assemblies v2 after Pilon improvements were more contiguous, contained more bases with high quality, and had fewer gaps and errors.

Gene annotation improvement in updated assemblies

The gene annotations of the reference strains Pb18, Pb03 and Pb01 were updated using a pipeline to transfer and revise gene structures (Methods). The implemented annotation pipeline was an updated and improved version of the previous protocol used to annotate *Paracoccidioides* spp. assemblies v1. The current pipeline includes an updated set of gene prediction programs, including the EVM caller used to select the best call for each locus. Databases used for training these gene prediction programs are also more comprehensive, with more sequences available from the dimorphic fungi group for comparison. Also, the databases used for homology inference and functional annotation were updated since the previous annotation. In addition, we identified orthologs to evaluate the gene calls for consistency (see below). The incorporation of these new methods and data improved the evidence supporting gene prediction.

The updated gene sets are more consistent across the three *Paracoccidioides* genomes (S1 Table). The total gene count for the

Table 1. Summary of assembly metrics after Pilon improvement.

Pilon summary metrics	*P. brasiliensis*		*P. lutzii*
	Pb18	**Pb03**	**Pb01**
Read depth of coverage	127	146	148
SNPs	3,290	3,018	3,072
Ambiguous bases	246	222	221
Small insertion bases	957	1,083	1,062
Small deletion bases	725	714	628
Bases added in reassembly fixes	109,312	89,243	118,931
Bases removed in reassembly fixes	37,417	38,906	41,822
Gaps opened	0	0	0
Gaps closed	113	56	212
Collapsed regions	3	1	2
Collapsed bases	64,378	20,918	43,967
Increase in contig N50 (kb)	14.16	4.98	29.17

doi:10.1371/journal.pntd.0003348.t001

two *P. brasiliensis* genomes now only differs by 37 in v2 whereas the v1 gene counts differed by 866; overall the update removed 351 genes from Pb18 and added 552 genes to Pb03. *P. lutzii* (Pb01) also has a more similar gene count, due to 306 fewer genes in the v2 compared to v1. A more detailed view of the gene structure changes by major categories is provided in Table 2; these statistics were calculated by mapping the transcripts from the previous annotation to the corresponding locus on assembly v2. Notably, this analysis helped recover a large number of genes missed by the original annotation in each genome; the total genes newly added to a region was 840 in Pb18, 933 in Pb03 and 936 in Pb01. In addition, dubious genes were removed from each genome; the number of genes no longer present at the same locus was 1187 in Pb18, 490 in Pb03 and 1265 in Pb01. Other changes include extending or truncating transcripts, merging or splitting transcripts, changes to splice sites, and changes to UTRs (Table 2). Only 23% of genes in the v2 annotations were unchanged from v1; the primary transcripts were identical for 1816 genes in Pb18, 2599 in genes Pb03 and 1581 in genes Pb01. Genes with any type of change in their coding sequences represent a smaller subset, in total 5734 (68%) in Pb18, 4895 (58%) in Pb03 and 6309 (71%) in Pb01.

Both sequence addition (gap filling and local reassembly) and small changes in the genome assemblies (single-nucleotide substitutions or insertion/deletion events (indels)), contributed to the improvement of the gene annotation in the update v2. Two

Table 2. Summary of annotation changes in protein coding genes.

Change type	*P. brasiliensis*		*P. lutzii*	Change description
	Pb18	**Pb03**	**Pb01**	
Add	840	933	936	Gene added to a region that previously had none
Splice site	1,124	1,122	1,000	Same start and same stop; internally, a splice site moved
Extended	3	6	2	Splice agreement, new model is longer; upstream start and downstream stop
Start Extended	262	329	307	Splice agreement and same stop; new model is longer, upstream start
Stop Extended	46	38	30	Splice agreement and same start; new model is longer, downstream stop
Shift	15	12	18	Splice agreement; new model has upstream, or downstream, start and stop
Truncated	5	4	2	Splice agreement, new model is shorter; downstream start and upstream stop
Start Truncated	237	208	276	Splice agreement and same stop; new model is shorter, downstream start
Stop Truncated	7	11	6	Splice agreement and same start; new model is shorter, upstream stop
UTR	1,504	802	1,679	Splice agreement and same start and same stop, but differ in UTR
Cluster	12	9	16	Multiple old genes map to multiple new genes; complex change
Merge	118	88	102	Multiple old genes have been merged into one
Split	402	509	495	Single old gene has been split into multiple new genes
Other CDS change	1,999	1,757	2,376	Other model not covered by another category or multiple models
None	1,816	2,599	1,581	Primary transcript is identical
Total	8,390	8,427	8,826	Total genes in current annotation version 2

doi:10.1371/journal.pntd.0003348.t002

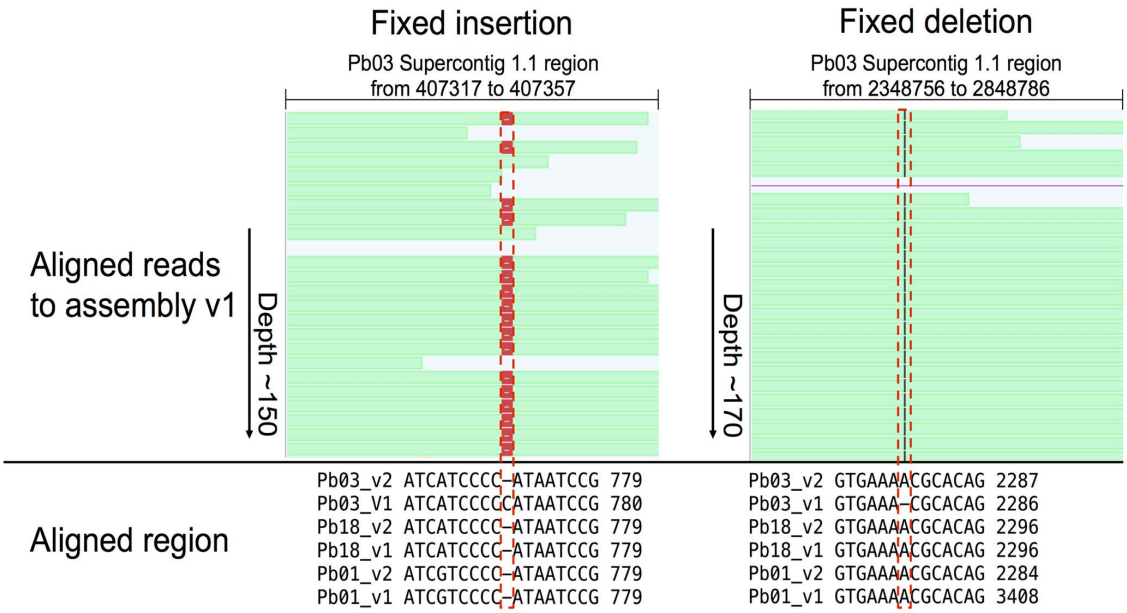


Fig. 2. Examples of an artifactual insertion and an artifactual deletion that were corrected during the update of the *P. brasiliensis* Pb03 genome sequence. Screenshots of Pilon-generated genome browser tracks in GenomeView v1.0 [35] show the evidence used by Pilon to recognize and correct an incorrect insertion in the gene PABG_00120 (left) and an incorrect deletion in the gene PABG_00790 (right). Tracks (top panels) depict paired-end reads (green) aligned to the corresponding region of the reference assembly v1, a subset of the total depth of ~150X or ~170X; these alignments were used by Pilon to refine the consensus sequence, generating the improved Pb03 assembly v2. Positions in the v1 assembly where aligned reads suggest a change due to either a gap (red box) or an insertion (black line) are indicated with dashed red boxes. The changes suggested by Pilon are also supported by conservation of the changed bases in a multiple alignment (bottom panels) with the corresponding region of *P. brasiliensis* Pb18 and *P. lutzii* Pb01.
doi:10.1371/journal.pntd.0003348.g002

examples of how indel correction fixed gene structures are shown in Fig. 2. In the first case (left panel), an extra C was inserted at a polyC tract in PABG_00129 of Pb03; correction of this position resulted in extending the coding DNA sequence (CDS) of this gene by 423 bases. In the second case (right panel), an A was deleted at a polyA tract in PABG_00790; correction of this position also corrected the reading frame, allowing for removal of a false intron that was needed to step over a stop codon and extension of the CDS of this gene by 252 bases. While these are small changes to the underlying assembly, both have had larger impact on correcting these gene structures.

The annotation improvements were also analyzed by comparing the alignments of orthologs for all three *Paracoccidioides* genomes, identified by OrthoMCL (Methods). For orthologs identified either from the v1 or v2 assemblies, maximum and minimum gene length was computed for each ortholog cluster. In comparing these gene lengths (Fig. 3A), the v1 gene annotations (red points in scatterplot) exhibited a higher variation among Pb18, Pb03 and Pb01 orthologs compared to annotation v2 (blue points). The positions that are closer to the diagonal correspond to smaller differences in gene length between orthologs; as expected for an improved annotation, the v2 points are closer to the diagonal than v1. These differences between maximum and minimum length of the genes within each orthologous cluster group were also plotted on a logarithmic scale (Fig. 3B), based on sorting cluster differences from smallest to largest. The v2 annotation differences (blue curve) were lower and well separated from the v1 annotation (red curve), providing additional support of the increased length concordance in the v2 annotation.

Further analysis of gene conservation also supported the greater consistency among the *Paracoccidioides* spp. genomes in the v2 annotation. The number of genes found in all three genomes increased, whereas the number of unique genes specific to only one genome decreased; this has produced a more uniform set of protein coding genes (Figure S2). The improved structural annotation also led to improvements in functional annotation. The v2 annotation had more genes with assigned protein domain families (PFAM) and pathway information (KEGG), using the same version of these databases for the v1 and v2 gene sets (Figure S2). This supports the higher functional content of the revised gene sets, despite the lower total gene counts in two of the genomes.

We also manually reviewed and curated the predicted structures of a number of protein-coding genes that are of importance to the *Paracoccidioides* research community, including well-characterized yeast-phase specific genes and other virulence factors. This introduces changes to the transcript sequence of 27 of these genes (Table 3). The improvements to the assemblies resulted in updated transcript predictions for the vast majority of genes of the three reference strains, with substantially corrected gene structures for several virulence-associated or yeast-phase specific genes of central importance in *Paracoccidioides* or other dimorphic fungi, including *HSP90* [25], Pb*GP43* [26], Pb*P27* [27], *RYP1-3* [28], *BAD1* [29], catalase B, alpha 1,3 glucan synthase and the beta glucan synthase target gene *FKS1* [30,31]. An extreme example is the *HSP90* gene, where corrections were made to the sequence of each of the three *Paracoccidioides* genomes (Fig. 4). This example illustrates the annotation errors in v1 of all *Paracoccidioides* reference strains that were fixed in v2 after Pilon improvement and re-annotation. In this case one or more single-nucleotide errors,

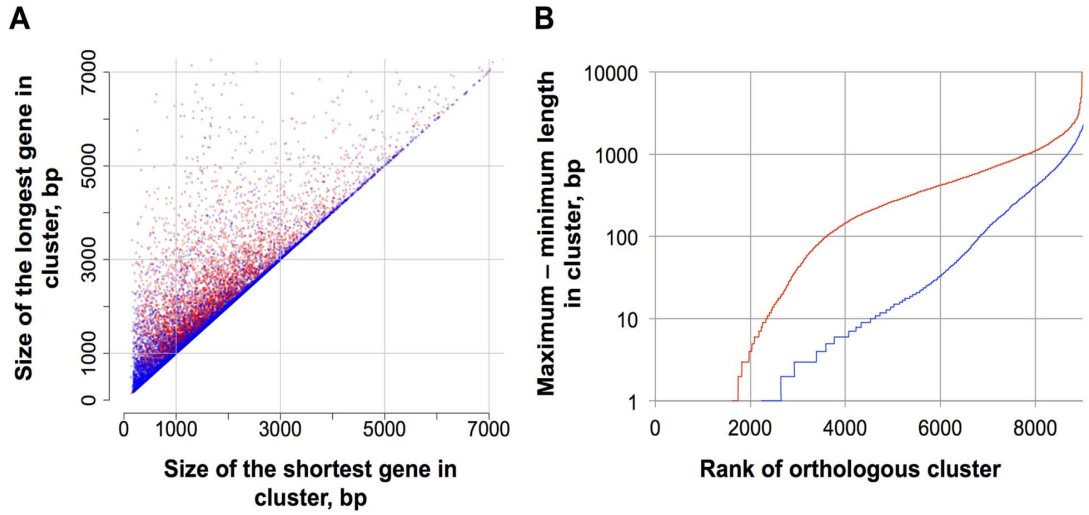


Fig. 3. Improved consistency of gene annotation in v2 genomes. The final predicted gene sets of the three *Paracoccidioides* strains were clustered using OrthoMCL, in v1 and v2. The scatterplots (A) compare, for each clustered group, the maximum length versus the minimum length of the three *Paracoccidioides* genes in the same cluster, for each of the two versions. The scatterplot contrasts the maximum-minimum pairs from annotation v1 (red points) and those from annotation v2 (blue points). The location of blue points closer to the diagonal illustrates that the annotation v2 was more consistent across the three genomes with smaller differences in gene length. In the same sense, the rank plots (B) show the difference between maximum and minimum length for each clustered group, for each of the two versions; again annotation v2 (blue line) showed fewer (later increase) and smaller (more gradual increase) differences, corresponding to the improvement of the genome annotation in v2.
doi:10.1371/journal.pntd.0003348.g003

unknown single nucleotides (N's), and/or single nucleotides that were erroneously reported as absent or duplicated by Sanger sequencing resulted in radically different predicted gene structures (intron/exon and/or gene boundary errors). This is shown in detail for a cluster of errors present at the end of HSP90 in Pb03 (Fig. 4B), which included alteration of the proper stop codon, resulting in premature truncation of this gene. Another example of coding sequence updates to multiple genomes is shown for FKS1, where different regions of the Pb03 and Pb18 proteins were restored in the updated assemblies and annotations (Figure S1).

The improvements in the annotation v2 were also analyzed for completeness by comparing to a set of highly conserved fungal genes defined by CEGMA and to protein sets of related dimorphic human pathogenic fungi. Genes in the v2 annotation showed a higher coverage of both the CEGMA and related dimorphic fungal data sets in comparison with annotation v1, suggesting these v2 genes are more complete (Figure S3). Furthermore, we examined the level of conservation to other fungi, by analyzing the difference of the BLASTp score between the v1 and v2 protein sets compared to those of *B. dermatitidis*, *H. capsulatum* and the CEGMA genes of *S. cerevisiae* and *S. pombe*. We observed that in all cases the v2 annotation had more hits greater than the minimum-similarity cutoff, and that the vast majority of genes of the v2 annotation had higher BLAST score values than their counterparts from v1 annotation (Figure S4).

Discussion

The initial draft genomes of three isolates of *Paracoccidioides* (*P. brasiliensis* isolates Pb03 and Pb18, and *P. lutzii* isolate Pb01) served as the first complete genome references for this fungal species [1]. Although these assemblies were obtained using the best technology available at that time, they included gaps and low quality sequence in genic and intergenic regions, which in turn resulted in a number of suboptimal gene structures, coding sequences and predicted protein sequences. This work has revised these reference genomes, providing the *Paracoccidioides* community with more complete and accurate sequences; this provides a more accurate foundation for future genome-based, molecular biological or genetic research on paracoccidioidomycosis and the fungal strains that cause it. The strategy we have followed will more widely be useful also for other groups wishing to update fungal and other microbial genomes in future.

The updating of a reference genome, in particular of the underlying assembly and annotation, can be thought of as a largely computational form of deep sequence curation. The success of the update we present here shows that next-generation sequencing (NGS) together with publicly available software tools can markedly enhance the quality of a eukaryotic genome resource. Indeed, the availability of affordable NGS sequencing opportunities makes such endeavors accessible to small bioinformatics groups. The massively computer-assisted component of such an update, which can include tabular and graphical views for monitoring improvements and performing quality control, can be complemented by choosing and following a few 'guide genes' to evaluate the process. This focused analysis provides tangible examples of how the update affected predicted properties of important genes, such as gene structure or encoded proteins.

The accuracy of a genome sequence and associated annotations are critically important for many types of analysis; therefore validating and improving the accuracy of the sequence and annotation can have wide impact, especially for methods highly sensitive to sequence errors. One example involves examining a genome sequence for evidence of genes and genic regions likely to be under positive selection. Such genes and genic regions, which are believed to be relatively rare in many eukaryotic genomes (see, e.g., [32]), are sometimes associated in pathogenic organisms with virulence or rapid adaptation to host conditions, including resistance to defense by the host or avoidance of the host immune system. An example of such adaptation has been found for surface

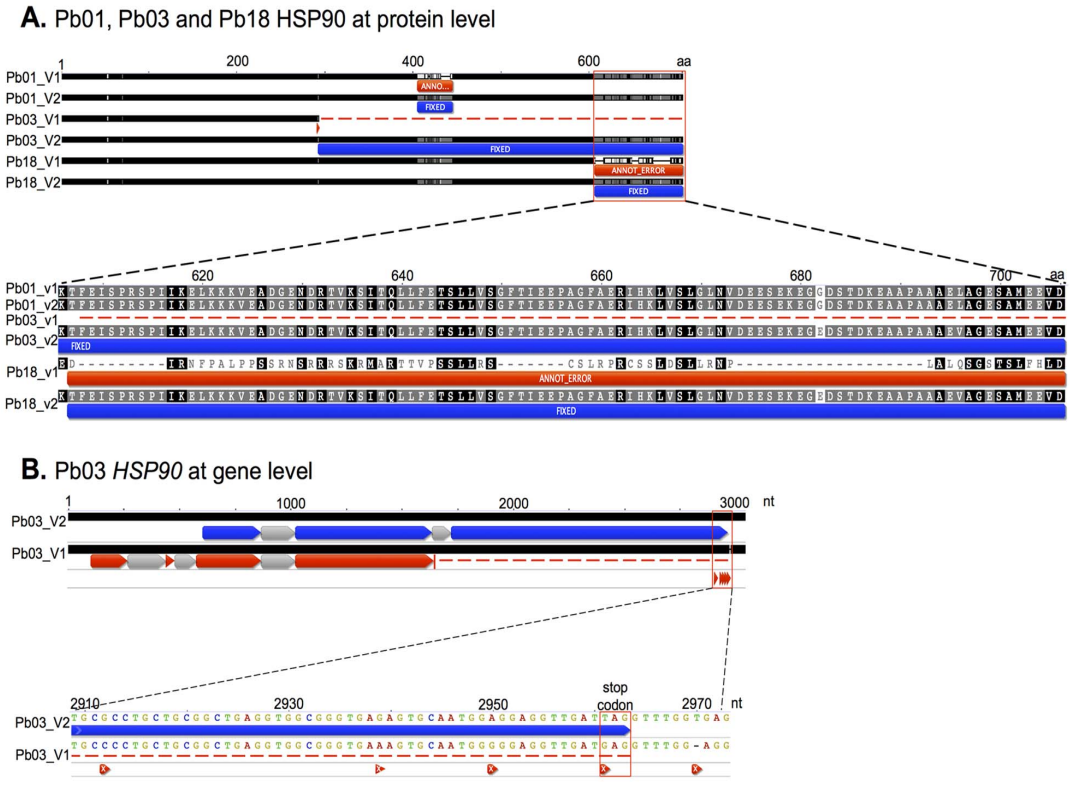


Fig. 4. Diverse error correction for the 90 kDa heat shock protein (HSP90 gene) of *Paracoccidioides* spp. (A) In this example different annotation errors were present in v1 of all three *Paracoccidioides* reference strains, all of which were fixed in v2 after Pilon improvement and re-annotation. The example also illustrates how one or more single-nucleotide errors, unknown single nucleotides (N's), or single nucleotides that were erroneously reported as absent or duplicated by a Sanger sequencer can amplify across annotations, generating radically different gene structure (intron/exon and/or gene boundary) predictions. (B) Five changes are shown at assembly (DNA sequence) level, one of which was a single nucleotide error in a stop codon; as a result, the gene-calling program did not recognize the end of an exon and it was not reported.
doi:10.1371/journal.pntd.0003348.g004

proteins of diverse pathogens [32] and in fungi of the proline-rich antigen gene in *Coccidioides* spp. [33]. Positive selection can also occur in response to antimicrobial drugs, as in chloroquine resistance in *Plasmodium falciparum* [34]. Candidate regions under positive selection are commonly identified as sections of coding regions having unusually high rates of nonsynonymous (amino-acid changing) substitutions. Precisely because such regions are quite rare, a coding region of low sequence quality having several sequencing errors could be categorized as a region under positive selection, and if there are several such regions in a genome, an automated genome-wide screen will report a high percentage of false positives. Conversely, assembly and annotation improvements such as we describe here can effectively evaluate and fix such regions of a genome so that even error-sensitive evolutionary analyses become realistic.

Similar considerations also apply to analyses that have more obvious clinical relevance. For example, improving a DNA sequence's accuracy can bring it closer to being 'clinical grade' or 'diagnostic grade'. Indeed, identification of a clinical sample of a human pathogenic fungus isolated in a hospital using sequence comparison requires certainty that any nucleotide differences (e.g., resulting from single nucleotide polymorphisms/SNPs) observed between the sequenced sample and trusted reference strain(s) or isolate(s) are not simply errors in the reference. For fungi encountered in clinical contexts, only one or a few traditionally used loci are typically represented by reliable reference sequences, which are often from the ribosomal DNA, or from one or two protein-coding genes known beforehand to be diagnostically informative. Reference diagnostics, as well as diagnostic PCR assays (e.g., primer/probe design in real-time PCR assays), depend on such regions that have been reliably characterized at the molecular level in a fair number of related species or strains that could be present in clinical settings. Whole-genome gene sets offer, however, new perspectives; if their sequence quality is high, one could then systematically and exhaustively screen alignments of the full gene sets for diagnostically promising genes and genic regions that are likely to be informative for the identification task at hand. Such genome-wide screens should be able to identify new, candidate target loci, and molecular assays could then be developed for them and validated. Genome sequences also allow for metagenomic or metatranscriptomic analysis, where reference genomes enable identification of the pool random sequence from the population of microbes in a sample. Such wide applications will be better powered by efforts such as this to improve the set of reference genomes that form a fundamental basis of comparison and analysis.

By re-sequencing three reference strains of *Paracoccidioides* spp., using deep sequencing depth of Illumina paired-end reads, we have been able to substantially improve the assemblies and annotations for this important human fungal pathogen. Here we have presented the updated and improved annotated genome sequences, which constitute new references that can be used in diverse future molecular projects by those working in the field of

Table 3. Changes in updated annotations of known yeast-phase specific genes or virulence factors of *Paracoccidioides* and other dimorphic human pathogenic fungi.

Gene name description	Annotation v2 IDs			Type of change			Ref.
	P. lutzii	*P. brasiliensis*		*P. lutzii*	*P. brasiliensis*		
	Pb01	Pb03	Pb18	Pb01	Pb03	Pb18	
	PAAG_	PABG_	PADG_	PAAG_	PABG_	PADG_	
1,3-beta-glucan synthase component GLS1 (*FKS1*)	05071	04524	**11846***	None	Splice site	CDS/ID	[36]
1,3-beta-glucanosyltransferase (*GEL1*)	03782	00831	03286	Splice site	Splice site	Splice site	[31]
4-hydroxyphenylpyruvate dioxygenase (*4-HPPD*)	02615	03102	01636	UTR	Splice site	UTR	[37]
Alpha-1,3-glucan synthase (*AGS1*)	03297	00726	03169	CDS	CDS	Splice site	[38,39]
Alternative oxidase (*AOX*)	01078	01661	03747	UTR	UTR	UTR	[40,41]
bZIP transcription factor (*HAPX*)	04257	04038	07492	UTR	UTR	UTR	[42,43]
Catalase B (*CATB*)	01553	03611	00225	None	None	Stop extended	[44,45]
Catalase peroxixomal (*CATP*)	01454	01943	00324	Stop extended	Start truncated	Start truncated	[46]
Conserved hypothetical protein (*P27*)	08096	07332	08402	UTR	None	UTR	[27]
Cu Zu superoxide dismutase (*SOD1*)	04164	03954	07418	CDS	CDS	CDS	[47,48]
Cu Zu superoxide dismutase (*SOD3*)	02971	00431	02842	UTR	None	None	[36,47]
Dimorphism regulator histidine kinase (*DRK1*)	05810	06372	07579	UTR	CDS	CDS	[49]
Glucan 1,3-beta-glucosidase (*GP43*)	05770	06340	07615	UTR	UTR	UTR	[26]
Glyceraldehyde-3-phosphate dehydrogenase (*GAPDH*)	08468	00022	02411	Splice site	Extended	UTR	[48,50]
HAD-superfamily hydrolase (*HAD32*)	00503	06765	02181	UTR	Splice site	Splice site	[51]
Heat shock protein 60 Kda (*HSP60*)	08059	07300	08369	UTR	UTR	UTR	[36,52]
Heat shock protein 90 Kda (*HSP90*)	05679	06249	07715	Splice site	CDS	CDS	[25]
L-ornithine 5-monooxygenase (*SID1*)	01682	03730	00097	Splice site	None	None	[43,53]
Tubulin beta chain (*TUB*)	03031	00486	02900	CDS	CDS	CDS	[15]
Adhesin WI-1 (*BAD1*)	08980	07814	**12525°**	UTR	None	New gene	[29,54]
cAMP-independent regulatory protein pac2 (*RYP1*)	**11652****	06919	06243	CDS/ID	None	UTR	[28,55]
Ornithine decarboxylase (*ODC*)	03153	00600	03032	None	None	UTR	[56]
Required for yeast phase growth 2 (*RYP2*)	02671	03151	**11302*****	Splice site	Splice site	CDS/ID	[28,57]
Required for yeast phase growth 3 (*RYP3*)	06081	06575	08037	Start extended	CDS	Stop extended	[28,57]
Urease	00954	01291	03871	CDS	Start truncated	CDS	[58]
Urease accessory protein (*UREG*)	06237	05255	07010	UTR	Start truncated	UTR	[59]
Ureidoglycolate hydrolase (*UGH*)	04751	00102	02493	UTR	UTR	UTR	[59]

*IDs from version 1 *PADG_04920; **PAAG_03579; ***PADG_01695.
°New gene that was not reported previously.
doi:10.1371/journal.pntd.0003348.t003

medical mycology. Since the process leading to the new sequences is largely automated using publicly available programs and the NGS technology used is cost-effective, the success of our strategy represents a proof of concept that may stimulate similar updates of other genomes in future.

Supporting Information

Figure S1 Regions of FKS1 protein sequence alignment highlighting changes between genome versions. This example is of a gene relevant to medical and experimental mycology, the 1,3-beta glucan synthase component *FKS1*. The colored regions (blue text for bases, red '-' for gaps in the alignment) illustrate the improvements in the annotation ('CDS' category) that led to improved protein sequences in two of the three *Paracoccidioides* strains.
(TIF)

Figure S2 Comparison of ortholog conservation and annotation in v1 and v2 genomes. The final gene sets of each annotation version were clustered using OrthoMCL. (**A**) Bar chart showing the relative contributions of core, auxiliary and unique genes to the final gene clusters of versions 1 and 2. The clustered groups were categorized as 'core' if present in all strains, 'aux' if present in two strains and 'uniq' if present in one strain. (**B**) Venn diagram showing numbers of shared and unique genes in annotation v2. (**C**) Total numbers of predicted genes, and total numbers of predicted genes that were assigned functional annotation from PFAM and/or KEGG. In all cases, the annotation v2 is more consistent or homogeneous across the three strains than annotation v1: there are more core genes in annotation v2 and also in v2 the two strains Pb03 and Pb18 from the same species (*P. brasiliensis*) give more similar results (bar charts). Furthermore, the new annotation has more genes with assigned functional annotation even using the same version of the databases.
(TIF)

Figure S3 Coverage of Core Eukaryotic Genes (CEGs) in original and updated genomes. More genes in annotation v2 had higher percent of coverage of CEGs in comparison with annotation v1. This analysis was performed and plotted using the CoreAlyze tool (http://sourceforge.net/project/corealyze).
(TIF)

Figure S4 Difference in BLAST scores using the protein sets of *Paracoccidioides* annotations v1 and v2. The references used for the BLAST matching were the protein sets of two dimorphic pathogenic fungi that are closely related to *Paracoccidioides* (top row) and the Core Eukaryotic genes of the fungi in CEGMA (bottom row). The comparison shows that the vast majority of proteins with any change, included in the comparison, have higher BLAST score in annotation v2. Here the graphs depict the results for the Pb03 strain; Pb18 and Pb01 showed the same pattern. The horizontal axis shows the genes numbered in order of increasing score difference (v2 minus v1).
(TIF)

Table S1 Summary of assembly and gene statistics of v1 and the updated v2 of the three reference genomes of *Paracoccidioides* spp.
(XLSX)

Acknowledgments

We thank the members of the Broad Institute's Genome Sequencing and Analysis Program for their assistance in assembly and annotation, for discussions and advice, and for the use of unpublished genomes obtained in collaboration, and Christopher Desjardins for helpful comments on the manuscript.

Author Contributions

Conceived and designed the experiments: OKC JGM CAC. Analyzed the data: JFM JEG EM. Wrote the paper: JFM JEG OKC JGM CAC. Performed the assembly and annotation improvement: MP AI SY QZ.

References

1. Desjardins CA, Champion MD, Holder JW, Muszewska A, Goldberg J, et al. (2011) Comparative genomic analysis of human fungal pathogens causing paracoccidioidomycosis. PLoS Genet 7: e1002345. doi:10.1371/journal.pgen.1002345.
2. Brummer E, Castaneda E, Restrepo A (1993) Paracoccidioidomycosis: an update. Clin Microbiol Rev 6: 89–117.
3. Theodoro RC, Teixeira M de M, Felipe MSS, Paduan KDS, Ribolla PM, et al. (2012) Genus paracoccidioides: Species recognition and biogeographic aspects. PloS One 7: e37694. doi:10.1371/journal.pone.0037694.
4. Matute DR, McEwen JG, Puccia R, Montes BA, San-Blas G, et al. (2006) Cryptic speciation and recombination in the fungus *Paracoccidioides brasiliensis* as revealed by gene genealogies. Mol Biol Evol 23: 65–73. doi:10.1093/molbev/msj008.
5. Haas BJ, Zeng Q, Pearson MD, Cuomo CA, Wortman JR (2011) Approaches to Fungal Genome Annotation. Mycology 2: 118–141. doi:10.1080/21501203.2011.606851.
6. Muñoz JF, Gallo JE, Misas E, McEwen JG, Clay OK (2013) The eukaryotic genome, its reads, and the unfinished assembly. FEBS Lett 587: 2090–2093. doi:10.1016/j.febslet.2013.05.048.
7. Cruveiller S, Jabbari K, Clay O, Bernardi G (2003) Compositional features of eukaryotic genomes for checking predicted genes. Brief Bioinform 4: 43–52.
8. Fisher S, Barry A, Abreu J, Minie B, Nolan J, et al. (2011) A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. Genome Biol 12: R1. doi:10.1186/gb-2011-12-1-r1.
9. Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26: 589–595. doi:10.1093/bioinformatics/btp698.
10. Stanke M, Waack S (2003) Gene prediction with a hidden Markov model and a new intron submodel. Bioinforma Oxf Engl 19 Suppl 2: ii215–ii225.
11. Ter-Hovhannisyan V, Lomsadze A, Chernoff YO, Borodovsky M (2008) Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. Genome Res 18: 1979–1990. doi:10.1101/gr.081612.108.
12. Majoros WH, Pertea M, Salzberg SL (2004) TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. Bioinforma Oxf Engl 20: 2878–2879. doi:10.1093/bioinformatics/bth315.
13. Birney E, Clamp M, Durbin R (2004) GeneWise and Genomewise. Genome Res 14: 988–995. doi:10.1101/gr.1865504.
14. Korf I (2004) Gene finding in novel genomes. BMC Bioinformatics 5: 59. doi:10.1186/1471-2105-5-59.
15. Goldman GH, dos Reis Marques E, Duarte Ribeiro DC, de Souza Bernardes LA, Quiapin AC, et al. (2003) Expressed sequence tag analysis of the human pathogen *Paracoccidioides brasiliensis* yeast phase: identification of putative homologues of *Candida albicans* virulence and pathogenicity genes. Eukaryot Cell 2: 34–48.
16. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, et al. (2008) Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol 9: R7. doi:10.1186/gb-2008-9-1-r7.
17. Li L, Stoeckert CJ, Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13: 2178–2189.
18. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, et al. (2004) Versatile and open software for comparing large genomes. Genome Biol 5: R12.
19. Eddy SR (2011) Accelerated Profile HMM Searches. PLoS Comput Biol 7: e1002195. doi:10.1371/journal.pcbi.1002195.
20. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, et al. (2014) Pfam: the protein families database. Nucleic Acids Res 42: D222–D230. doi:10.1093/nar/gkt1223.
21. Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28: 27–30.
22. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389–3402.
23. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, et al. (2007) Clustal W and Clustal X version 2.0. Bioinforma Oxf Engl 23: 2947–2948. doi:10.1093/bioinformatics/btm404.

24. Parra G, Bradnam K, Korf I (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinforma Oxf Engl 23: 1061–1067. doi:10.1093/bioinformatics/btm071.
25. Tamayo D, Muñoz JF, Torres I, Almeida AJ, Restrepo A, et al. (2013) Involvement of the 90 kDa heat shock protein during adaptation of *Paracoccidioides brasiliensis* to different environmental conditions. Fungal Genet Biol FG B 51: 34–41. doi:10.1016/j.fgb.2012.11.005.
26. Torres I, Hernandez O, Tamayo D, Muñoz JF, Leitão NP, et al. (2013) Inhibition of PbGP43 expression may suggest that gp43 is a virulence factor in *Paracoccidioides brasiliensis*. PloS One 8: e68434. doi:10.1371/journal.pone.0068434.
27. Torres I, Hernandez O, Tamayo D, Muñoz JF, García AM, et al. (2014) *Paracoccidioides brasiliensis* PbP27 gene: knockdown procedures and functional characterization. FEMS Yeast Res 14: 270–280. doi:10.1111/1567-1364.12099.
28. Beyhan S, Gutierrez M, Voorhies M, Sil A (2013) A temperature-responsive network links cell shape and virulence traits in a primary fungal pathogen. PLoS Biol 11: e1001614. doi:10.1371/journal.pbio.1001614.
29. Brandhorst TT, Wüthrich M, Warner T, Klein B (1999) Targeted gene disruption reveals an adhesin indispensable for pathogenicity of *Blastomyces dermatitidis*. J Exp Med 189: 1207–1216.
30. Puccia R, Vallejo MC, Matsuo AL, Longo LVG (2011) The paracoccidioides cell wall: past and present layers toward understanding interaction with the host. Front Microbiol 2: 257. doi:10.3389/fmicb.2011.00257.
31. Rappleye CA, Goldman WE (2006) Defining virulence genes in the dimorphic fungi. Annu Rev Microbiol 60: 281–303. doi:10.1146/annurev.micro.59.030804.121055.
32. Endo T, Ikeo K, Gojobori T (1996) Large-scale search for genes on which positive selection may operate. Mol Biol Evol 13: 685–690.
33. Johannesson H, Vidal P, Guarro J, Herr RA, Cole GT, et al. (2004) Positive directional selection in the proline-rich antigen (PRA) gene among the human pathogenic fungi *Coccidioides immitis*, *C. posadasii* and their closest relatives. Mol Biol Evol 21: 1134–1145. doi:10.1093/molbev/msh124.
34. Wootton JC, Feng X, Ferdig MT, Cooper RA, Mu J, et al. (2002) Genetic diversity and chloroquine selective sweeps in Plasmodium falciparum. Nature 418: 320–323. doi:10.1038/nature00813.
35. Abeel T, Van Parys T, Saeys Y, Galagan J, Van de Peer Y (2012) GenomeView: a next-generation genome browser. Nucleic Acids Res 40: e12. doi:10.1093/nar/gkr995.
36. Tavares AHFP, Silva SS, Dantas A, Campos EG, Andrade RV, et al. (2007) Early transcriptional response of *Paracoccidioides brasiliensis* upon internalization by murine macrophages. Microbes Infect Inst Pasteur 9: 583–590. doi:10.1016/j.micinf.2007.01.024.
37. Nunes LR, Costa de Oliveira R, Leite DB, da Silva VS, dos Reis Marques E, et al. (2005) Transcriptome analysis of *Paracoccidioides brasiliensis* cells undergoing mycelium-to-yeast transition. Eukaryot Cell 4: 2115–2128. doi:10.1128/EC.4.12.2115-2128.2005.
38. Marques ER, Ferreira MES, Drummond RD, Felix JM, Menossi M, et al. (2004) Identification of genes preferentially expressed in the pathogenic yeast phase of *Paracoccidioides brasiliensis*, using suppression subtraction hybridization and differential macroarray analysis. Mol Genet Genomics MGG 271: 667–677. doi:10.1007/s00438-004-1016-6.
39. Rappleye CA, Engle JT, Goldman WE (2004) RNA interference in *Histoplasma capsulatum* demonstrates a role for alpha-(1,3)-glucan in virulence. Mol Microbiol 53: 153–165. doi:10.1111/j.1365-2958.2004.04131.x.
40. Martins VP, Dinamarco TM, Soriani FM, Tudella VG, Oliveira SC, et al. (2011) Involvement of an alternative oxidase in oxidative stress and mycelium-to-yeast differentiation in *Paracoccidioides brasiliensis*. Eukaryot Cell 10: 237–248. doi:10.1128/EC.00194-10.
41. Ruiz OH, Gonzalez A, Almeida AJ, Tamayo D, Garcia AM, et al. (2011) Alternative oxidase mediates pathogen resistance in *Paracoccidioides brasiliensis* infection. PLoS Negl Trop Dis 5: e1353. doi:10.1371/journal.pntd.0001353.
42. Gauthier GM, Sullivan TD, Gallardo SS, Brandhorst TT, Vanden Wymelenberg AJ, et al. (2010) SREB, a GATA transcription factor that directs disparate fates in *Blastomyces dermatitidis* including morphogenesis and siderophore biosynthesis. PLoS Pathog 6: e1000846. doi:10.1371/journal.ppat.1000846.
43. Hwang LH, Mayfield JA, Rine J, Sil A (2008) *Histoplasma* requires SID1, a member of an iron-regulated siderophore gene cluster, for host colonization. PLoS Pathog 4: e1000044. doi:10.1371/journal.ppat.1000044.
44. Holbrook ED, Smolnycki KA, Youseff BH, Rappleye CA (2013) Redundant catalases detoxify phagocyte reactive oxygen and facilitate *Histoplasma capsulatum* pathogenesis. Infect Immun 81: 2334–2346. doi:10.1128/IAI.00173-13.
45. Inglis DO, Voorhies M, Hocking Murray DR, Sil A (2013) Comparative transcriptomics of infectious spores from the fungal pathogen *Histoplasma capsulatum* reveals a core set of transcripts that specify infectious and pathogenic states. Eukaryot Cell 12: 828–852. doi:10.1128/EC.00069-13.
46. Chagas RF, Bailão AM, Pereira M, Winters MS, Smullian AG, et al. (2008) The catalases of *Paracoccidioides brasiliensis* are differentially regulated: protein activity and transcript analysis. Fungal Genet Biol FG B 45: 1470–1478. doi:10.1016/j.fgb.2008.08.007.
47. Youseff BH, Holbrook ED, Smolnycki KA, Rappleye CA (2012) Extracellular superoxide dismutase protects *Histoplasma* yeast cells from host-derived oxidative stress. PLoS Pathog 8: e1002713. doi:10.1371/journal.ppat.1002713.
48. Hernandez O, Garcia AM, Almeida AJ, Tamayo D, Gonzalez A, et al. (2011) Gene expression during activation of *Paracoccidioides brasiliensis* conidia. Yeast Chichester Engl 28: 771–781. doi:10.1002/yea.1902.
49. Nemecek JC, Wuthrich M, Klein BS (2006) Global control of dimorphism and virulence in fungi. Science 312: 583–588. doi:10.1126/science.1124105.
50. Edwards JA, Chen C, Kemski MM, Hu J, Mitchell TK, et al. (2013) *Histoplasma* yeast and mycelial transcriptomes reveal pathogenic-phase and lineage-specific gene expression profiles. BMC Genomics 14: 695. doi:10.1186/1471-2164-14-695.
51. Hernández O, Almeida AJ, Gonzalez A, Garcia AM, Tamayo D, et al. (2010) A 32-kilodalton hydrolase plays an important role in *Paracoccidioides brasiliensis* adherence to host cells and influences pathogenicity. Infect Immun 78: 5280–5286. doi:10.1128/IAI.00692-10.
52. Guimarães AJ, Nakayasu ES, Sobreira TJP, Cordero RJB, Nimrichter L, et al. (2011) *Histoplasma capsulatum* heat-shock 60 orchestrates the adaptation of the fungus to temperature stress. PloS One 6: e14660. doi:10.1371/journal.pone.0014660.
53. Parente AFA, Bailão AM, Borges CL, Parente JA, Magalhães AD, et al. (2011) Proteomic analysis reveals that iron availability alters the metabolic status of the pathogenic fungus *Paracoccidioides brasiliensis*. PloS One 6: e22810. doi:10.1371/journal.pone.0022810.
54. Bohse ML, Woods JP (2007) RNA interference-mediated silencing of the YPS3 gene of *Histoplasma capsulatum* reveals virulence defects. Infect Immun 75: 2811–2817. doi:10.1128/IAI.00304-07.
55. Nguyen VQ, Sil A (2008) Temperature-induced switch to the pathogenic yeast form of *Histoplasma capsulatum* requires Ryp1, a conserved transcriptional regulator. Proc Natl Acad Sci U A 105: 4880–4885. doi:10.1073/pnas.0710448105.
56. Guevara-Olvera L, Hung CY, Yu JJ, Cole GT (2000) Sequence, expression and functional analysis of the *Coccidioides immitis* ODC (ornithine decarboxylase) gene. Gene 242: 437–448.
57. Webster RH, Sil A (2008) Conserved factors Ryp2 and Ryp3 control cell morphology and infectious spore formation in the fungal pathogen *Histoplasma capsulatum*. Proc Natl Acad Sci U A 105: 14573–14578. doi:10.1073/pnas.0806221105.
58. Mirbod-Donovan F, Schaller R, Hung C-Y, Xue J, Reichard U, et al. (2006) Urease produced by *Coccidioides posadasii* contributes to the virulence of this respiratory pathogen. Infect Immun 74: 504–515. doi:10.1128/IAI.74.1.504-515.2006.
59. Whiston E, Zhang Wise H, Sharpton TJ, Jui G, Cole GT, et al. (2012) Comparative transcriptomics of the saprobic and parasitic growth phases in *Coccidioides* spp. PloS One 7: e41034. doi:10.1371/journal.pone.0041034.

Chapter 6

The dynamic genome and transcriptome of the human fungal pathogen *Blastomyces* and close relative *Emmonsia*

PLOS | GENETICS

RESEARCH ARTICLE

The Dynamic Genome and Transcriptome of the Human Fungal Pathogen *Blastomyces* and Close Relative *Emmonsia*

José F. Muñoz[1,2☯], Gregory M. Gauthier[3☯], Christopher A. Desjardins[4☯], Juan E. Gallo[1,5], Jason Holder[4], Thomas D. Sullivan[6], Amber J. Marty[3], John C. Carmen[6¤], Zehua Chen[4], Li Ding[7], Sharvari Gujja[4], Vincent Magrini[7], Elizabeth Misas[1,2], Makedonka Mitreva[7], Margaret Priest[4], Sakina Saif[4], Emily A. Whiston[8], Sarah Young[4], Qiandong Zeng[4], William E. Goldman[9], Elaine R. Mardis[7], John W. Taylor[8], Juan G. McEwen[1,10], Oliver K. Clay[1,11], Bruce S. Klein[3,6,12], Christina A. Cuomo[4]*

1 Cellular and Molecular Biology Unit, Corporación para Investigaciones Biológicas, Medellín, Colombia, **2** Institute of Biology, Universidad de Antioquia, Medellín, Colombia, **3** Department of Medicine, University of Wisconsin, Madison, Madison, Wisconsin, United States of America, **4** Broad Institute of MIT and Harvard, Cambridge, Massachusetts, United States of America, **5** Doctoral Program in Biomedical Sciences, Universidad del Rosario, Bogotá, Colombia, **6** Department of Pediatrics, University of Wisconsin, Madison, Madison, Wisconsin, United States of America, **7** The Genome Institute, Washington University School of Medicine, St. Louis, Missouri, United States of America, **8** Department of Plant and Microbial Biology, University of California, Berkeley, Berkeley, California, United States of America, **9** Department of Microbiology and Immunology, School of Medicine, University of North Carolina, Chapel Hill, Chapel Hill, North Carolina, United States of America, **10** School of Medicine, Universidad de Antioquia, Medellín, Colombia, **11** School of Medicine and Health Sciences, Universidad del Rosario, Bogotá, Colombia, **12** Department of Medical Microbiology & Immunology, University of Wisconsin, Madison, Madison, Wisconsin, United States of America

☯ These authors contributed equally to this work.
¤ Current address: Department of Biological Sciences, Northern Kentucky University, Highland Heights, Kentucky, United States of America
* cuomo@broadinstitute.org

OPEN ACCESS

Citation: Muñoz JF, Gauthier GM, Desjardins CA, Gallo JE, Holder J, Sullivan TD, et al. (2015) The Dynamic Genome and Transcriptome of the Human Fungal Pathogen *Blastomyces* and Close Relative *Emmonsia*. PLoS Genet 11(10): e1005493. doi:10.1371/journal.pgen.1005493

Editor: Sajeet Haridas, DOE Joint Genome Institute, UNITED STATES

Received: May 5, 2015

Accepted: August 11, 2015

Published: October 6, 2015

Data Availability Statement: All sequence data for this project is available at NCBI and is linked to the umbrella BioProject PRJNA289976. All genome assemblies and annotations are available in NCBI under the following BioProjects: *Blastomyces dermatitidis* ATCC18188: PRJNA39265, *Blastomyces dermatitidis* ATCC26199: PRJNA39263, *Blastomyces gilchristii* SLH14081: PRJNA29173 and *Blastomyces dermatitidis* ER-3: PRJNA29171. *Emmonsia crescens* UAMH3008: PRJNA178252, *Emmonsia parva* UAMH139: PRJNA178178. All transcriptome

Abstract

Three closely related thermally dimorphic pathogens are causal agents of major fungal diseases affecting humans in the Americas: blastomycosis, histoplasmosis and paracoccidioidomycosis. Here we report the genome sequence and analysis of four strains of the etiological agent of blastomycosis, *Blastomyces*, and two species of the related genus *Emmonsia*, typically pathogens of small mammals. Compared to related species, *Blastomyces* genomes are highly expanded, with long, often sharply demarcated tracts of low GC-content sequence. These GC-poor isochore-like regions are enriched for *gypsy* elements, are variable in total size between isolates, and are least expanded in the avirulent *B. dermatitidis* strain ER-3 as compared with the virulent *B. gilchristii* strain SLH14081. The lack of similar regions in related species suggests these isochore-like regions originated recently in the ancestor of the *Blastomyces* lineage. While gene content is highly conserved between *Blastomyces* and related fungi, we identified changes in copy number of genes potentially involved in host interaction, including proteases and characterized antigens. In addition, we studied gene expression changes of *B. dermatitidis* during the interaction of the infectious yeast form with macrophages and in a mouse model. Both experiments

sequencing of ATCC26199 is available under PRJNA185598.

Funding: This project has been funded in whole or in part with Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Contract No. HHSN272200900018C. This work was partly supported by Colciencias via the grants "Comparative genomics and virulence in the pathogenic fungus Paracoccidioides brasiliensis" 2213-48925460 and "A gene atlas for human pathogenic fungi" 1222-56934875, and by the Universidad de Antioquia via a grant "Sostenibilidad 2014/15". Colciencias National Doctorate Program funding supported JFM and EM; Enlaza Mundos fellowship partly supported JFM. The Universidad del Rosario partly supported JEG. National Institutes of Health grants 5K08AI071004 and 1R21AI105537 funded GMG and R01 AI035681, AI040996, and AI093553 funded BSK. National Institutes of Health Research Service Award AI55397 and the Hartwell Foundation funded JCC. National Science Foundation grants DEB-125752 and DEB-1046115 partially supported JWT. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

highlight a strong antioxidant defense response in *Blastomyces*, and upregulation of dioxygenases *in vivo* suggests that dioxide produced by antioxidants may be further utilized for amino acid metabolism. We identify a number of functional categories upregulated exclusively *in vivo*, such as secreted proteins, zinc acquisition proteins, and cysteine and tryptophan metabolism, which may include critical virulence factors missed before in *in vitro* studies. Across the dimorphic fungi, loss of certain zinc acquisition genes and differences in amino acid metabolism suggest unique adaptations of *Blastomyces* to its host environment. These results reveal the dynamics of genome evolution and of factors contributing to virulence in *Blastomyces*.

Author Summary

Dimorphic fungal pathogens including *Blastomyces* are the cause of major fungal diseases in North and South America. The genus *Emmonsia* includes species infecting small mammals as well as a newly emerging pathogenic species recently reported in HIV-positive patients in South Africa. Here, we synthesize both genome sequencing of four isolates of *Blastomyces* and two species of *Emmonsia* as well as deep sequencing of *Blastomyces* RNA to draw major new insights into the evolution of this group and the pathogen response to infection. We investigate the trajectory of genome evolution of this group, characterizing the phylogenetic relationships of these species, a remarkable genome expansion that formed large isochore-like regions of low GC content in *Blastomyces*, and variation of gene content, related to host interaction, among the dimorphic fungal pathogens. Using RNA-Seq, we profile the response of *Blastomyces* to macrophage and mouse pulmonary infection, identifying key pathways and novel virulence factors. The identification of key fungal genes involved in adaptation to the host suggests targets for further study and therapeutic intervention in *Blastomyces* and related dimorphic fungal pathogens.

Introduction

Blastomyces is a genus of a thermally dimorphic fungal pathogen, which is the etiological agent of blastomycosis, a lung infection that can become a systemic mycosis. In North America, *Blastomyces* is endemic in the Ohio and Mississippi river valleys, the Great Lakes region, and the St. Lawrence River [1]. Within *Blastomyces*, two lineages of *B. dermatitidis* have been recognized [2], with recent work providing evidence that one lineage is a distinct species, *B. gilchristii* [3]. Both species can infect humans, and vary in morphology, virulence and immune responses by the host. The primary mode of infection is inhalation of conidia and the subsequent conversion of these conidia into parasitic yeast [4,5]. Clinical manifestations range from asymptomatic infection to symptomatic disease and include pneumonia, acute respiratory distress syndrome, and a rapidly progressive dissemination involving multiple organ systems that is often fatal [5,6]. Diagnosis is often complicated by the similarity of symptoms to those of viral or bacterial respiratory infection and by the aforementioned variety of manifestations [7].

As a thermally dimorphic fungus, *Blastomyces* has the remarkable ability to switch between two different morphologies in response to external stimuli, predominantly temperature [5]. At 22–25°C, *Blastomyces* grows as septate hyphae that produce infectious conidia and at 37°C it grows as a budding yeast [8]. *Blastomyces* is part of a larger group of dimorphic fungal pathogens, including *Histoplasma*, *Paracoccidioides*, and *Coccidioides*, all belonging to

the order Onygenales. The dimorphic fungi collectively are the most common cause of invasive fungal disease worldwide and account for several million infections each year [8]. Unlike opportunistic fungi, such as *Candida albicans*, *Cryptococcus neoformans*, or *Aspergillus fumigatus*, the dimorphic fungi can infect immunocompetent and immunocompromised hosts [6,9–11].

Previous work has shown that in *Blastomyces*, the temperature-dependent switch from hyphae to yeast along with upregulation of yeast-phase specific genes is critical for virulence [12–14]. The dimorphism-regulating kinase-1 (*DRK1*) promotes the temperature-dependent conversion from mold to yeast, and its deletion renders *Blastomyces* avirulent during experimental murine pulmonary infection [12]. The upregulation of yeast-phase specific genes, such as the *Blastomyces* yeast-phase specific gene 1 (*BYS1*) [15] and the *Blastomyces* adhesion-1 gene (*BAD1*) [13,14], is also important for the adaptive response of the yeast cells in the host environment. *BAD1* is considered an essential virulence factor in *Blastomyces*, since it binds yeast cells to host tissue and impairs host immune defenses by inhibiting the production of tumor necrosis factor-α and blocking $CD4^+$ T lymphocyte activation [13].

Within the Onygenales, *Blastomyces*, *Histoplasma* and *Paracoccidioides* belong to the family Ajellomycetaceae. Also within Ajellomycetaceae is the genus *Emmonsia*, which includes *E. crescens* and *E. parva*, the etiological agents of adiaspiromycosis, a pulmonary disease of small mammals and occasionally of humans [16]. Recently, a cluster of systemic infections of HIV-positive patients in South Africa were shown to be caused by *Emmonsia* isolates [17]. While *E. crescens* and *E. parva* also undergo a dimorphic shift at high temperature, they transform into large, thick-walled adiaspores rather than yeast cells [18] (S1 Table). Two phylogenetic studies using 18S ribosomal DNA sequences found that *E. parva* was the sister species to *Blastomyces* [19,20]. The positioning of *E. crescens* was less clear; in one analysis it was a sister group to *Paracoccidioides* [19] while in the other analysis it was grouped with *Blastomyces* and *E. parva* [20]. In neither phylogeny was the alternative positioning of *E. crescens* strongly supported.

To further investigate the genomic basis of differences observed among the Ajellomycetaceae in terms of pathogenicity, morphology, and the infection process, we sequenced six genomes of *Blastomyces* and *Emmonsia*, as well as sequencing the *B. dermatitidis* transcriptome during macrophage co-cultivation and *in vivo* pulmonary infection. The newly sequenced genomes included three representative strains of *B. dermatitidis* (ER-3, ATCC18188, and ATCC26199), and one strain of each of *B. gilchristii* (SLH14081), *E. parva* (UAMH139), and *E. crescens* (UAMH3008). *Blastomyces dermatitidis* ER-3 was isolated from a woodpile located in a highly endemic region of Wisconsin and is hypovirulent in mice [21,22]. The ATCC18188 strain is the only current example of the 'a' mating type (*MAT1-1* locus) available for *B. dermatitidis* [23]. ATCC26199 is a clinical isolate from South Carolina that is commonly used for *in vitro* and *in vivo* laboratory assays [14]. *Blastomyces gilchristii* SLH14081 is a human clinical isolate that is highly virulent in a murine model of blastomycosis [22,24]. Both *Emmonsia* strains were isolated from small mammals, *E. parva* from a weasel in Ravelli County, Montana, and *E. crescens* from lungs of a rodent (*Arvicola terrestris*) in Norway.

Utilizing this genomic data, we find that the *Blastomyces* genomes are much larger than those of their close relatives, and are characterized by large, isochore-like GC-poor regions overrun by repetitive elements. Our whole-genome analyses provide further evidence for the phylogenetic relationships between *Blastomyces* and *Emmonsia* and other Onygenales. Finally, we identify novel sets of candidate virulence factors through comparison of the *Blastomyces* transcription during *in vivo* pulmonary infection to growth in co-culture with macrophages or in different media or temperature. This combination of genomic and transcriptomic analysis provides a foundation and new candidate genes to further characterize the underlying

molecular differences that determine the infectious potency of *Blastomyces* strains and give rise to the clinical profiles attributable to blastomycosis.

Results

Expanded genomes of *Blastomyces* species

We sequenced and assembled the genomes of three *Blastomyces dermatitidis* strains and one *B. gilchristii* strain, and representatives of two *Emmonsia* species. The *Blastomyces* strains were sequenced using either Sanger technology or a hybrid of Sanger and 454 technologies. The *Emmonsia* strains were sequenced using Illumina technology, and *de novo* assemblies were generated for each strain (Methods). Comparison of the genomes of four *Blastomyces* strains, SLH14081, ER-3, ATCC18188 and ATCC26199, revealed they were over twice the size of all other Onygenales. The *Blastomyces* assemblies range in size from 66.6 Mb for *B. dermatitidis* strain ER-3 to 75.4 Mb *B. gilchristii* strain SLH14081 (Table 1). These assemblies were over twice as large as those of other dimorphic pathogens in the order Onygenales including the *Emmonsia* species (30.4 Mb), although the use of only short reads from a single library for the two *Emmonsia* may under-represent repetitive sequence (Fig 1). The assemblies of two *Blastomyces* strains, SLH14081 and ER-3, were sequenced to a higher depth than the other two strains, and as a result contain nearly all of the assembled sequence in a relatively small number of scaffolds, 100 and 25 scaffolds respectively. As an independent assessment of genome size and structure, we generated an optical map of the SLH14081 strain (S1 Fig). Consistent with our assembly of this strain, the map had an estimated size of 79.6 Mb, arranged in eighteen linkage groups. In addition, a total of 65.9 Mb of the 75.4 Mb of the SLH14081 assembly was anchored to the optical map (S2 Table).

The total number of predicted genes in *Blastomyces*, *Emmonsia*, and other related fungi was similar despite the large difference in genome size. In *Blastomyces*, the number of predicted genes varied between 9,180 in ATCC26199 to 10,187 in ATCC18188; for *E. parva* and *E. crescens* the counts were similar, 8,563 and 9,444, respectively (Table 1), as were those of other sequenced Onygenales (Fig 1). High representation of core eukaryotic genes in each genome provides evidence that their gene sets are nearly complete; *E. parva* includes 88% of core eukaryotic genes, while the *E. crescens* and *Blastomyces* gene sets include 96–98% (S2 Fig).

Phylogenetic position of *Blastomyces*, *Emmonsia parva* and *E. crescens*

To compare gene content and conservation, we identified orthologous gene clusters in the six genomes sequenced here, 10 additional Onygenales genomes, including three other pathogenic species (*Histoplasma*, *Paracoccidioides*, and *Coccidioides*), and three *Aspergillus* genomes. Using 2,062 single copy core genes present in all strains, we estimated a phylogeny of these

Table 1. Assembly and annotation statistics for *Blastomyces* and *Emmonsia* genomes. *Bd*: *B. dermatitidis*, *Bg*: *B. gilchristii*, *Ep*: *E. parva*, *Ec*: *E. crescens*.

	Total assembly length	Scaffolds	Scaffold N50	GC-content (%)	Genes	Coding (%)	Intergenic length	Repeat (%)
Bg SLH14081	75.35 Mb	100	2.44 Mb	35.8	9,692	16.9	7.2 kb	63.0
Bd ER-3	66.57 Mb	25	5.55 Mb	37.1	9,755	19.2	6.0 kb	60.0
Bd ATCC18188	73.58 Mb	4,159	0.40 Mb	36.7	10,187	17.4	4.2 kb	56.6
Bd ATCC26199	71.52 Mb	3,282	0.29 Mb	36.6	9,180	17.5	4.5 kb	58.5
Ep UAMH139	30.35 Mb	2,682	31.17 kb	44.7	8,563	35.6	1.4 kb	9.9
Ec UAMH3008	30.36 Mb	1,150	95.15 kb	45.4	9,444	41.8	1.4 kb	5.4

doi:10.1371/journal.pgen.1005493.t001

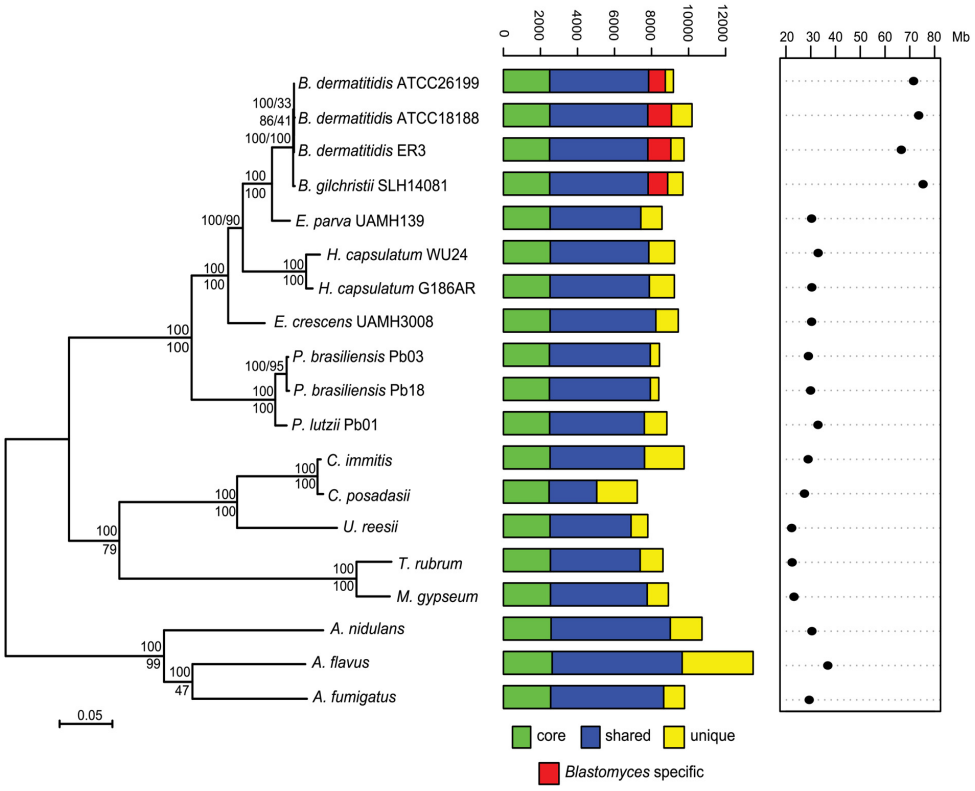


Fig 1. Phylogeny and gene conservation of *Blastomyces* and *Emmonsia* spp. Maximum likelihood tree inferred from concatenated protein alignments of 2,062 core genes based on 1,000 replicates; all bootstrap values (top value for each node) were 100% except for one node within *B. dermatitidis*, which was 88%. Branch order was also well supported by the consensus of individual gene trees (GSF, lower value for each node). A bar plot of orthology classes is shown to the right, where core genes found in all genomes are shown in green, shared genes present in more than one but not all genomes in blue, genes specific to *Blastomyces* genomes in red, and genes that were unique to only one of the 19 genomes in yellow. Finally, genome size is plotted for each genome along the x-axis, which ranges from 20 to 80 Mb.

doi:10.1371/journal.pgen.1005493.g001

organisms using RAxML ([25]; Fig 1). This analysis strongly supports the clustering of *Blastomyces* with *E. parva* (100% of bootstrap replicates and 100% Gene Support Frequency (GSF) [26]) as previously reported [19,20]. In contrast to prior work, *Histoplasma* is strongly supported as sister group to *Blastomyces* and *E. parva* (100% of bootstrap replicates and 90% GSF), with *E. crescens* strongly supported as a sister group to that clade (100% of bootstrap replicates and 100% GSF), and with *Paracoccidioides* in a basal position (Fig 1). The polyphyletic nature of *Emmonsia* suggests that the Ajellomycetaceae have undergone multiple evolutionary transitions allowing the infection of humans and other mammals. Within *Blastomyces*, we found support for strain SLH14081 as an outgroup relative to the other three strains (S3 Fig). This is consistent with the placement of strain SLH14081 within the newly described species *B. gilchristii* [3]; the other three strains sequenced here are still classified as *B. dermatitidis*.

Blastomyces genomes show a bimodal GC distribution

A bimodal distribution of GC-content observed in all *Blastomyces* sequenced, which was less pronounced in *E. parva* and *E. crescens* and absent in other Ajellomycetaceae, suggests that

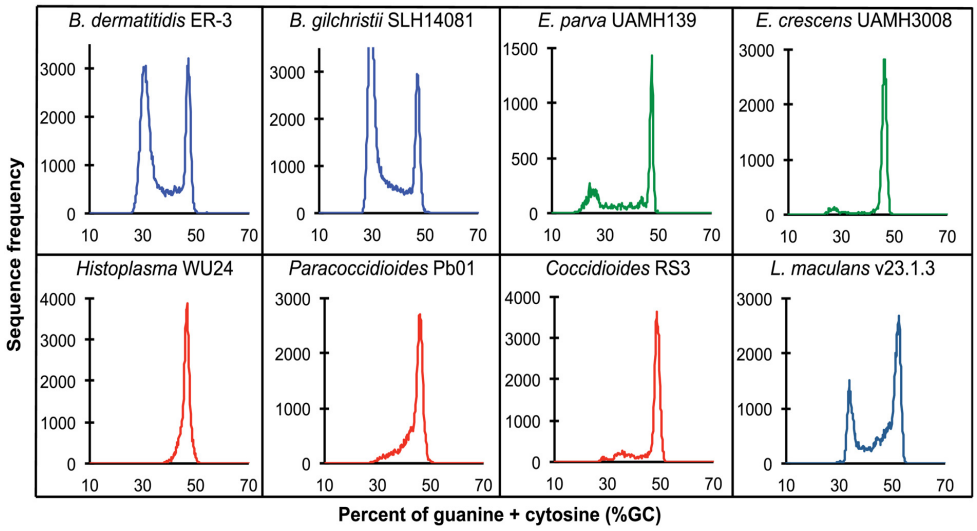


Fig 2. GC frequency distributions (histograms) of overlapping fragments (windows, of 32 kb) of the genome assemblies of *Blastomyces dermatitidis* ER-3, *B. gilchristii* SLH14081, *Emmonsia parva* (UAMH139), *E. crescens* (UAMH3008), *Histoplasma capsulatum* (WU24), *Paracoccidioides lutzii* (Pb01), *Coccidioides immitis* (RS3), and *Leptosphaeria maculans* (v23.1.3). The bin size of the histograms is approximately 0.1% GC. Horizontal axes show GC % and vertical axes show relative frequencies.

doi:10.1371/journal.pgen.1005493.g002

these genomes are organized in large isochore-like regions of high and low GC-content. This finding for nuclear DNA explains the GC-poor fraction of the *Blastomyces* genome initially identified using CsCl gradient analytical ultracentrifugation [27], which the authors hypothesized was due to a large proportion of GC-poor mitochondrial DNA in *Blastomyces* cells. Examining the genome wide GC content revealed a bimodal distribution for all strains of *Blastomyces* including ER-3 and SLH14081, the smallest and largest assembly, respectively (Fig 2), and was observed for all window sizes ranging from 2 kb to 256 kb (S4 Fig). The detection of a bimodal signal in larger windows supports the organization of the genomes in large isochore-like regions, with average GC content of 29.6% and 31.0% in GC-poor regions and 45.9% and 46.6% for the rest of the genome in *B. gilchristii* strain SLH14081 and *B. dermatitidis* strain ER-3, respectively (Table 2). Analysis of the related pathogens *H. capsulatum*, *P. lutzii*, and *C. immitis* showed no evidence for bimodality of GC content, while both *E. parva* and *E. crescens* revealed small peaks of low GC sequence. Read-based analysis and using smaller window sizes (e.g. 128 bp) supported these findings, suggesting they are not due to differences in assembly completeness (S5 Fig).

To further examine the organization of GC-content across the genome, we next defined the boundaries of low GC content regions in *Blastomyces*. In the smallest assembly, of the ER-3 strain, we identified 221 GC-poor tracts with an average size of 186.0 kb, encompassing a total size of 41.1 Mb (Tables 2 and S3). In the largest assembly, of the SLH14081 strain, we identified 350 GC-poor tracts with an average size of 140.2 kb, encompassing a total size 49.1 Mb (Tables 2 and S3). The 8 Mb difference between the total size of GC-tracts in the genomes of *B. dermatitidis* ER-3 and *B. gilchristii* SLH14081 accounts for nearly all of the 8.8 Mb difference in assembly size. Notably, GC-poor tracts in *Blastomyces* can be quite long, and reach maximal lengths of 1.3 Mb. In the assemblies of *E. parva*, *E. crescens* and other Ajellomycetaceae, long GC-poor tracts were rarely observed (e.g., a total of only 4 GC-poor regions larger than 10 kb

Table 2. Gene and repeat features of *Blastomyces* GC-rich and GC-poor regions compared to *Histoplasma*.

	Blastomyces				*Histoplasma*
	GC-poor		GC-rich		
	ER-3	SLH14081	ER-3	SLH14081	WU24
Total size (Mb)	41.1	49.1	25.4	26.2	33
Total genes	1,990	1,858	7,765	7,834	9,251
Gene length (bp)	1,549	1,471	2,737	2,716	1,686
Intergenic distance (bp)	18,523	22,983	1,212	1,400	1,850
GC content (%)	31	29.6	46.6	45.9	46.2
Gene GC content (%)	48	46.9	51.8	51.8	51.2
Coding (%)	7.5	5.6	83.4	80.9	35.9
Syntenic genes (%)	73.4	76.1	99	98.8	NA
Repeat (%)	93.7	95.2	6.2	4.8	15.4

doi:10.1371/journal.pgen.1005493.t002

in *E. parva* were found adjacent to a long GC-rich region in the same scaffold, and just 1 in *E. crescens*), corresponding to the less pronounced bimodal GC distribution of the genome assembly. However, more contiguous assemblies would be needed to reveal the overall extent of long GC-poor tracts. The only other fungal genome noted to have an isochore-like structure, *Leptosphaeria maculans* [28], contains a smaller expansion of GC-poor regions (Fig 2); individual tracts were on average half the size (70.4 kb) of those in *Blastomyces*, and encompassed a smaller fraction (36%) of the *L. maculans* genome [28]. This difference is consistent with the lower fraction of long AT blocks we observe by comparing windows of different sizes in *Blastomyces* and *L. maculans* (S4 Fig).

The GC-poor regions include nearly all the repetitive elements in the genome and consequently have a lower density of predicted genes (e.g., see Fig 3). In ER-3, 93.7% of repetitive sequence is found in GC-poor regions (Table 2). The gypsy elements that dominate repetitive sequence in the *Blastomyces* genomes have low GC-content; on average those in ER-3 and SLH14081 have respective GC-content of 31.0% and 29.9%, matching the overall GC level of the GC-poor regions (Table 2). GC-poor tracts of *Blastomyces* contain only approximately one fifth of the predicted protein-coding gene set, including some notable genes such as 1,3-beta-glucan synthase component (*FKS1*), *Blastomyces* yeast phase-specific gene (*BYS1*), and one of two *BYS1*-like proteins we identified (S6 Fig and S4 Table). By contrast, *BAD1*, which encodes an essential virulence factor involved in host cell interaction and immune evasion [13], is found within a GC-rich region. Intergenic regions are also larger here than for other genes in the genome; the average intergenic region for ER-3 is 18.5 kb in GC-poor regions, a 3-fold expansion compared to the 6.0 kb genome-wide average (Table 2 and Figs 3 and S6).

The GC-poor regions also show lower synteny between the *Blastomyces* genomes compared to other regions with more typical GC content (e.g., see Fig 3). Overall, *B. dermatitidis* strain ER-3 and *B. gilchristii* strain SLH14081 shared 125 syntenic blocks including 93.8% and 94.5% of genes, encompassing only 69.5% and 69.3% of each assembly. These percentages are much smaller than those observed among strains of related species (such as 95% and 93% synteny between strains of *P. brasiliensis* [29]). The lower synteny among *Blastomyces* strains is largely explained by the proportion of genes found in repeat-rich, GC-poor regions (Table 2 and Fig 3). Nearly all (99%) of genes in GC-rich regions are highly syntenic across *Blastomyces* strains, even between *B. dermatitidis* strain ER-3 and *B. gilchristii* strain SLH14081. However, the GC-poor regions have more limited synteny even within strains of *Blastomyces* encompassing 74 to 76% of genes in those regions (Table 2 and Fig 3).

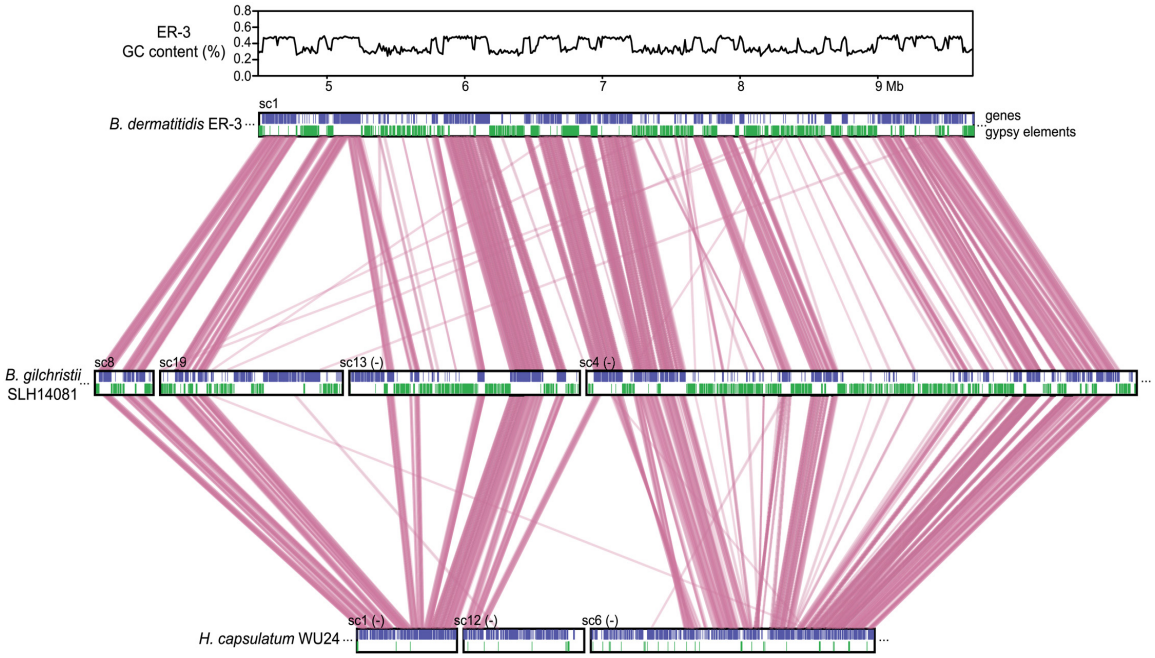


Fig 3. Correspondence of GC content and synteny for *Blastomyces*. Comparison of GC content (top panel) and genome synteny (lower panel) for a 5.2 Mb region of *B. dermatitidis* strain ER-3 (scaffold (sc) 1, coordinates from 4.5 to 9.7 Mb) and corresponding syntenic regions of *B. gilchristii* strain SLH14081 and *Histoplasma capsulatum* strain WU24. Location of genes (blue boxes) and gypsy elements (green boxes) are depicted across each genomic region. Orthologs between genomes are connected in pink, which are organized into syntenic regions that are disrupted by GC-poor regions in both *Blastomyces* genomes.

doi:10.1371/journal.pgen.1005493.g003

Overall, the function, expression, and selective pressure of genes in GC-poor regions appear similar to those genes found elsewhere in the genome. Despite the lower synteny, GC-poor regions are not significantly enriched for *Blastomyces*-specific genes, nor did they show much functional enrichment (S1 Text, S5 Table). Comparing selection pressure on the 7,228 single copy orthologs present in all four *Blastomyces* genomes also did not find a significant difference in the number of genes with high omega values (omega > 1) (Methods). These analyses suggest that despite the dynamic reorganization due to invading gypsy elements, the GC-poor regions do not appear to be fast evolving by these measures. Furthermore, there is no large-scale difference in the expression levels of genes in GC-poor regions. Comparing transcript levels for genes in GC-poor and GC-rich regions, we found that genes in both GC classes show similar expression levels (S7 Fig), again supporting the general similarity of genes found in these two genomic neighborhoods.

Characterization of repetitive sequence expansion

The 2-fold larger size of the *Blastomyces* genomes compared to other dimorphic fungi is due largely to an expansion of repetitive sequence. The proportions of the *Blastomyces* genome assemblies that were covered by repeats ranged from 56.6% (41.6 Mb) for *B. dermatitidis* ATCC18188 to 63.0% (47.5 Mb) for *B. gilchristii* SLH14081. SLH14081 had the highest repeat content and the largest assembly size. The *E. parva* and *E. crescens* assemblies both had a lower repeat content, 9.9% (3.0 Mb) and 5.4% (1.6 Mb), respectively (Table 1). In all genomes, a

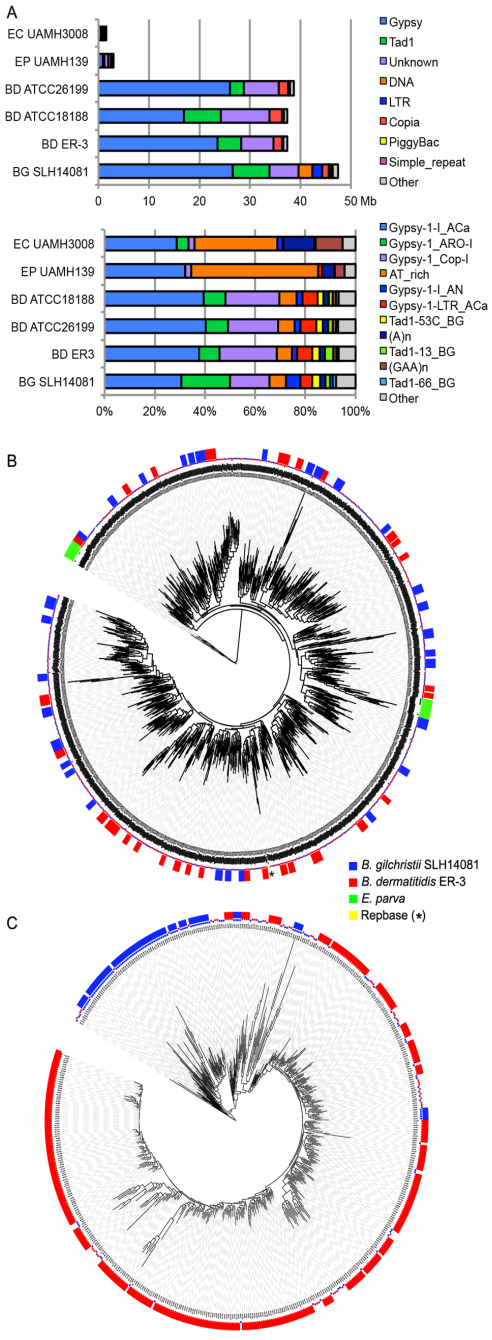


Fig 4. Relative contributions from known repeat categories to *Blastomyces* and *Emmonsia* genomes. **(A)** Repetitive elements were identified in each assembly using a combination of *de novo* classified elements and known elements. The total amount of genome sequence for each element class (top panel) and the relative frequency of known elements (bottom panel) are shown for *B*. *dermatitidis* (BD; ATCC26199, ATCC18188, ER-3), *B*. *gilchristii* (BG; SLH14081), *E*. *crescens* (EC; UAMH3008), and *E*. *parva* (EP;

UAMH139). **(B, C)**. Phylogenetic relationship of two subgroups of gypsy elements was inferred using FastTreeDP from alignments of reverse transcriptase domains. The largest subgroup of 922 sequences **(B)** includes domains from the *Blastomyces* strains ER-3 and SLH14081, *E. parva* strain UAMH139, and the Repbase ACa Gypsy element, whereas the other subgroup of 544 sequences **(C)** is specific to the two *B. dermatitidis* and *B. gilchristii*. The outer ring indicates strain specific duplications of four or more sequences.

doi:10.1371/journal.pgen.1005493.g004

small number of transposable element classes as well as AT-rich simple sequence regions were highly represented (Fig 4A).

More specifically, the genome expansion in *Blastomyces* strains has resulted from a proliferation of gypsy LTR retrotransposons, including both ancestral and lineage-specific proliferation. In the *Blastomyces* genomes, Gypsy elements account for almost all repetitive DNA, with a lower frequency of sequences similar to the non-LTR Tad1 and copia LTR retroelements (Figs 4A and S8). In all *Blastomyces* and *Emmonsia* genomes the most frequent Gypsy element subtype matches the "ACa" (*Ajellomyces* or *Histoplasma capsulatum*) Gypsy element from Repbase [30] (Fig 4A and 4B). Further phylogenetic characterization of 2,331 Gypsy elements identified four subtypes that appear specific to *Blastomyces* (S1 Text and Figs 4 and S9). Some subtypes had diversities that were primarily the result of ancestral duplication, resulting in large numbers of orthologous elements between strains (e.g., Fig 4B), while other subtypes appeared to predominantly contain strain-specific paralogous expansions, consistent with the cryptic speciation in the *Blastomyces* genus (e.g., Fig 4C). Gypsy elements were also detected in the *Emmonsia* and *Histoplasma* assemblies, but in far fewer copies (Figs 3 and 4), consistent with the recent expansion in *Blastomyces*. Gypsy elements are frequently observed in fungal genomes [31], including *Coccidioides* [32] and *Paracoccidioides* [29] but in far fewer copies.

Gene family evolution of *Blastomyces* and other Ajellomycetaceae

To identify gene content that could play a role in the evolution of the dimorphism and pathogenesis within the family Ajellomycetaceae, we searched for expansions or contractions in functionally classified genes compared to the other fungi from the order Onygenales (S6 Table). We identified PFAM domains, KEGG pathways, Gene Ontology (GO) terms, or kinase families that were significantly enriched or depleted. Domains associated with polyketide synthases were depleted in the Ajellomycetaceae, and an independent prediction of secondary metabolite enzymes confirmed that *Blastomyces* and other fungi from the Ajellomycetaceae contain fewer PKS gene clusters than other Onygenales (S7 Table, S1 Text). Other differences between the Ajellomycetaceae and other Onygenales include fewer copies of multiple classes of peptidases (M36, M43, S8) as well as an associated inhibitor (I9, inhibitor of S8 protease), variable copy number of LysM-domain proteins potentially involved in chitin binding, which are most expanded in dermatophytes but at next highest levels among the human pathogens in *Blastomyces*, and a higher number of protein kinases (S6 Table and S10 Fig), including an expansion of the FunK1 family similar to that previously noted in *Paracoccidioides* [29].

We next identified 140 gene clusters conserved in *Blastomyces*, *Emmonsia*, *Histoplasma*, and *Paracoccidioides*, but absent from other Onygenales and *Aspergillus* (S8 Table). These gene clusters include a predicted heme oxygenase (BDBG_02689), which could produce free iron from host heme. In addition to the 140 gene clusters, we also identified conserved genes in subsets of the Ajellomycetaceae including homologs of two previously typed antigens; a gene similar to the 27 kDa antigen of *Paracoccidioides* [33] is present in *Blastomyces* and one *Histoplasma* genome, and a gene cluster specific to *Blastomyces* and *Paracoccidioides* shares similarity with the antigenic *Aspergillus* cell wall mannoprotein [34].

Genes depleted in or absent from *Emmonsia* with possible roles in virulence or phase transitions

To identify potential genetic features of the Ajellomycetaceae pathogenic to immunocompetent humans (*Blastomyces*, *Histoplasma*, and *Paracoccidioides*) relative to *E. parva* and *E. crescens*, we conducted a second enrichment analysis as described above (S9 Table). The primary pathogens showed enrichment of only two PFAM domains, a phosphorylase and endonuclease (S9 Table). The phosphorylase domain over-represented in *Blastomyces* is associated with nucleoside phosphorylases; many of these proteins also contain Ankyrin repeats and NACHT domains. Phosphorylases are involved in nucleotide and amino acid salvage, and could allow pathogens greater metabolic versatility when certain building blocks are unavailable. The absence of any larger pattern of gain or loss of functional classes suggests that smaller changes in gene content, independent gain and loss between the species, or expression differences may account for differences in pathogenesis.

We then identified specific orthologs present in all four strains of *Blastomyces* but absent from both non-pathogenic *Emmonsia* species. Comparing the ortholog set of *Blastomyces* to *E. parva* and *E. crescens*, we found a total of 552 ortholog clusters that were present in all *Blastomyces* strains but absent in both *Emmonsia* genomes (S10 Table). Most of these (393 clusters) were present only in *Blastomyces*, and while most of these proteins (92% in *B. gilchristii* strain SLH14081) had no PFAM domain assignment, the set did include the *Blastomyces* yeast phase-specific protein 1 (*BYS1*). This gene is a marker of the phase transition to and from the yeast phase [15], although it has recently been shown not to be required for virulence in studied strains [24].

While both *E. parva* and *E. crescens* are not reported to be primary human pathogens, phylogenetic analysis suggests that the transition to this lifestyle may have been independent, resulting in differential gene loss. One of the genes absent only in *E. crescens* is the siderophore iron transporter mirB (*MIRB*). Many pathogenic microorganisms have evolved high affinity iron acquisition mechanisms, which include the production and uptake of siderophores. In *B. dermatitidis*, the expression of genes involved in the biosynthesis of siderophores and uptake of siderophores, including two iron transporters (*MIRB* and *MIRC*), are induced under iron-poor conditions [35]. While *MIRB* appears to be absent in *E. crescens*, siderophore uptake may be still enabled by the second transporter, *MIRC*, which is conserved in this species.

Transcriptional profiling of *Blastomyces* in macrophages

To better understand which *Blastomyces* genes play roles in pathogenicity and virulence, we carried out RNA-Seq of *B. dermatitidis* strain ATCC26199 to profile expression under varying temperature, nutrient availability, and infection status. Combining this data allowed us to disambiguate expression variability due solely to differences in temperature and media-specific nutrient availability from those specific to macrophage interactions *in vitro* or host infection *in vivo*. Five conditions were sampled: 37°C with macrophages in RPMI media, 37°C in RPMI media, 37°C in HMM media, 22°C in HMM media, and *in vivo* pulmonary infection with yeast in a mouse model (Fig 5A). For each condition, two biological replicates were performed, and the read counts per transcript were highly correlated between replicates ($R > 0.99$, S11 Fig). Gene expression levels and mapping statistics are presented in S11 and S12 Tables, respectively.

When *B. dermatitidis* yeast cells were co-cultured with human bone marrow derived macrophages, the majority of yeast cells (59%) were internalized by macrophages. Comparison of yeast co-cultured with and without macrophages identified 140 genes differentially expressed between these two conditions, 112 of which were upregulated in the presence of macrophages (S13 Table). This upregulation suggested a small, specific response to macrophages in this

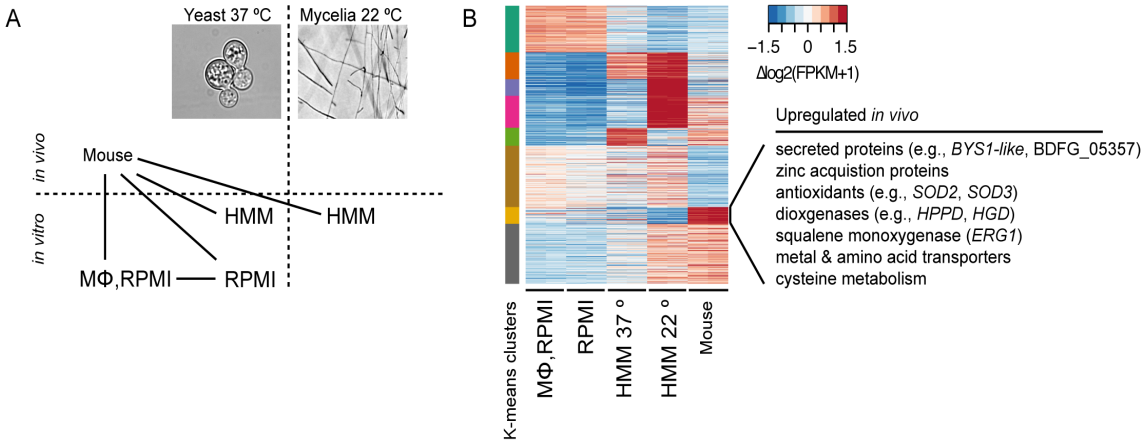


Fig 5. Transcriptional response of *B. dermatitidis* strain ATCC26199 to infection. (A) Schematic of samples compared by RNA-Seq analysis and **(B)** Heatmap of differentially expressed genes, where the cluster of genes specifically induced *in vivo* during mouse infection is highlighted.

doi:10.1371/journal.pgen.1005493.g005

experiment. Examination of this set of genes revealed numerous genes that have the potential to facilitate adaptation to the host environment. The 20 most significantly upregulated genes (Table 3) include a predicted secreted endo-1,3(4)-β-glucanase (BDFG_03060) involved in cell separation after cytokinesis in *C. albicans* [36], transporters, including an ABC transporter (BDFG_05060) homologous to *Aspergillus fumigatus MDR1* and a zinc transporter (BDFG_02462) similar to the vacuolar zinc transporter *ZRT3 in S. cerevisiae*, dehydrogenases

Table 3. *Blastomyces* transcripts most significantly induced during macrophage infection.

Locus	Predicted function	FDR	Fold Change
BDFG_03193	hypothetical protein, secreted	0	3.63
BDFG_03060	beta-1,3-glucanase,secreted	6.40E-196	2.04
BDFG_06058	transcription factor	6.68E-152	1.90
BDFG_05060	ABC transporter	5.07E-131	1.86
BDFG_04440	transcription factor	2.07E-118	1.54
BDFG_04186	hypothetical protein	1.92E-104	1.58
BDFG_06466	succinate dehydrogenase, iron-sulfur subunit	5.27E-101	1.23
BDFG_06207	vacuolar iron transporter	1.71E-92	1.98
BDFG_04494	succinate dehydrogenase, cytochrome b560 subunit	4.51E-90	1.20
BDFG_01343	electron-transferring-flavoprotein dehydrogenase	5.35E-87	1.17
BDFG_02965	catalase, CATP	4.69E-86	1.30
BDFG_00760	pyruvate decarboxylase	1.26E-84	1.66
BDFG_04269	Eukaryotic cytochrome b561	2.68E-78	1.09
BDFG_04739	hypothetical protein	8.97E-76	3.62
BDFG_02901	cytochrome P450 alkane hydroxylase	1.69E-75	1.05
BDFG_09499	cytochrome c	1.98E-73	1.15
BDFG_04995	hypothetical protein, secreted	3.69E-73	1.54
BDFG_01204	Cu-Zn superoxide dismutase, SOD3, secreted	1.51E-66	1.03
BDFG_02462	Metal ion transporter, ZRT3-like	5.43E-66	1.26
BDFG_04916	hypothetical protein	8.63E-66	1.14

doi:10.1371/journal.pgen.1005493.t003

involved in amino acid catabolism, and antioxidants peroxisomal catalase (*CATP*, BDFG_02965) and superoxide dismutases (*SOD3*, BDFG_01204; *SOD2*, BDFG_07895), which may protect against reactive oxygen species (ROS). The induction of endo-1,3(4)-β-glucanase and *CATP* in the presence of macrophages was also confirmed by qRT-PCR (S12 Fig and Methods).

Transcriptional profiling of *Blastomyces* in a mouse model

We also identified gene expression changes specific to *in vivo* murine pulmonary infection from our transcriptomic data of *B. dermatitidis* strain ATCC26199. By k-means clustering of expression values, we identified a set of 72 genes that are upregulated *in vivo* during mouse infection relative to all other conditions, regardless of temperature, media, and *in vitro* macrophage interactions (Fig 5B and S14 Table). Using all conditions for this comparison helped eliminate from consideration differences observed, for example, between the yeast samples cultured in different media. Genes in this set with greater than 2-fold upregulation *in vivo* are highlighted in Table 4, and primarily fell into five functional categories: 1) secreted proteins, 2) zinc acquisition, 3) antioxidants and oxygenases, 4) amino acid metabolism, and 5) transporters.

The most highly expressed gene *in vivo* was *BAD1* (BDFG_03370; S11 Table), which encodes a yeast-phase specific virulence factor that facilitates adhesion to host cells and evasion of host immune defenses [13]. *BAD1* also had the highest transcript abundance for yeast co-cultured with macrophages and yeast without macrophages at 37°C (S13 Table). Thus, *BAD1* was not identified among the set of 72 differentially expressed genes because the transcription of *BAD1* is influenced by temperature [37]. The effect of temperature during the mold to yeast transition on the transcriptome of dimorphic fungal pathogens has been the topic of previous studies [38–41] and was therefore not further evaluated here.

A total of nine secreted proteins were identified in this set of 72, including five of the ten most highly upregulated genes by fold change, potentially playing roles in host-pathogen interactions. Another highly up-regulated secreted protein (BDFG_00717) contains a predicted CFEM domain as well as a GPI-anchor; these features, as well as small size (236 amino acids), are shared with member of the haemoglobin-receptor gene family in *C. albicans* [42]. The most highly upregulated gene, BDFG_05357, encodes a HRXXH domain-containing secreted protein that may function as a zinc scavenging protein (Tables 4 and S14). This gene is present in the genomes of *Blastomyces* and *Coccidioides*, but absent from those of *Emmonsia*, *Histoplasma* and *Paracoccidioides*. BDFG_05357 is a homolog of *C. albicans PRA1* (pH-regulated antigen-1) [43] and *S. cerevisiae ZPS1* (zinc-pH-regulated protein). In *C. albicans*, the transcription of *PRA1* and *ZPS1* is induced under alkaline pH and zinc-deplete conditions [44,45], and *PRA1* is co-regulated with its upstream gene, *ZRT1*, which encodes a high-affinity zinc transporter that interacts with zinc-bound *PRA1* [45]. Similarly, the *B. dermatitidis* homolog of *ZRT1*, BDFG_09159, is highly expressed *in vivo*; the induced expression of both *PRA1* and *ZRT1* was confirmed by qRT-PCR (S12 Fig). However unlike in *C. albicans*, *ZRT1* is not adjacent to *PRA1* in the *B. dermatitidis* genome. While *PRA1* is conserved in all four *Blastomyces* genomes, there is no copy of this gene in *Histoplasma* as previously noted [45], nor in *Emmonsia* or *Paracoccidioides*, suggesting differences in how zinc is acquired within the Ajellomycetaceae.

In addition to *PRA1*/*ZPS1* and *ZRT1*, a larger module of genes that regulate zinc acquisition is co-regulated in *Blastomyces*. The transcript abundance of BDFG_07269, which encodes a low-affinity zinc transporter (*ZRT2*), is also significantly upregulated in the mouse model. In *S. cerevisiae*, the zinc-responsive transcription factor *ZAP1* regulates expression of *ZRT1* and *ZRT2*, along with *ZPS1*. We identified the ortholog of *ZAP1* in strain ATCC26199 as

Table 4. *Blastomyces* genes induced during mouse infection*

Locus	Predicted function	Functional categories	Mouse vs Mac	Mouse vs NoMac	Mouse vs Yeast	Mouse vs Mold	Average fold change
BDFG_02039	cysteine synthase	cysteine	0	0	0	0	9.03
BDFG_05357	HRXXH domain protein	secreted, zinc	0	0	0	0	8.11
BDFG_09329	secreted hypothetical protein	secreted	0	0	0	0	7.80
BDFG_02038	MFS transporter	transport	0	0	0	7.28E-304	6.30
BDFG_06873	secreted hypothetical protein	secreted	0	0	0	2.18E-164	5.96
BDFG_09159	zinc transporter, ZRT1	zinc, transport	0	0	3.27E-248	2.57E-312	5.12
BDFG_08433	glycerate kinase		5.28E-208	6.66E-275	1.46E-196	1.30E-83	4.60
BDFG_01204	superoxide dismutase, SOD3	zinc, redox	1.75E-216	0	0	4.67E-223	4.13
BDFG_07895	superoxide dismutase, SOD2	redox	1.99E-168	0		1.82E-316	4.13
BDFG_04319	oxidoreductase	redox	0	0	7.70E-205	1.31E-156	3.66
BDFG_01073	MaoC-like dehydratase		8.90E-77	1.11E-30	0	0	3.34
BDFG_04176	BYS1-like	secreted	6.92E-18	8.39E-53	1.14E-191	0	3.30
BDFG_09115	short chain dehydrogenase	redox	4.76E-90	7.17E-103	3.50E-79	3.98E-126	3.22
BDFG_07137	RNA ligase-like domain protein		8.80E-54	1.60E-44	3.29E-156	1.18E-219	2.82
BDFG_08059	cysteine dioxygenase	redox, cysteine	1.75E-52	5.95E-118	8.41E-294	4.36E-48	2.72
BDFG_05427	cation/proton antiporter	transport	3.58E-82	9.67E-115		0	2.70
BDFG_08334	secreted hypothetical protein	secreted	5.02E-76	1.31E-191	3.21E-31	1.73E-104	2.62
BDFG_00028	2-oxoisovalerate dehydrogenase E1 component, alpha subunit	redox	4.20E-56	2.69E-200	5.14E-285	2.66E-74	2.59
BDFG_02611	acetyl-coenzyme A synthetase		2.18E-108	4.00E-139	4.48E-116	4.64E-120	2.42
BDFG_07269	zinc transporter, ZRT2	zinc, transport	2.20E-162	6.94E-210	9.85E-44	5.74E-117	2.36
BDFG_00760	pyruvate decarboxylase		3.01E-15	1.22E-124	1.11E-272	9.32E-55	2.26
BDFG_05654	2-oxoisovalerate dehydrogenase E2 component	redox	1.63E-35	3.81E-135	7.75E-167	1.18E-87	2.24
BDFG_06615	Sodium:neurotransmitter symporter family	transport	4.85E-62	4.13E-80	3.43E-56	3.56E-119	2.18
BDFG_04184	phenylpyruvate dioxygenase	redox	1.08E-12	9.55E-99	9.00E-211	6.53E-70	2.16
BDFG_00717	CFEM domain protein	secreted	7.48E-15	1.60E-38	1.86E-106	1.27E-32	2.13
BDFG_01386	methionine sulfoxide reductase	redox	2.69E-140	5.79E-137	9.73E-218	2.50E-71	2.13
BDFG_06042	MFS transporter	transport	2.72E-56	2.47E-81	1.90E-31	6.36E-35	2.12
BDFG_03316	MFS transporter	transport	3.98E-44	4.94E-59	7.45E-99	4.48E-84	2.11
BDFG_03902	2-oxoisovalerate dehydrogenase, E1 component, beta subunit	redox	8.15E-37	2.56E-123	2.70E-135	1.48E-105	2.06
BDFG_05401	BTB/POZ-domain protein		9.74E-64	2.92E-88	2.59E-38	9.34E-62	2.02

*Genes with predicted PFAM domains or secretion signals, and greater than 2-fold higher expression during mouse infection are listed; full list of all significant genes in S14 Table.

doi:10.1371/journal.pgen.1005493.t004

BDFG_07048, which was also significantly upregulated *in vivo* relative to all other conditions (S14 Table) despite not being identified by k-means clustering. These results suggest that zinc acquisition and homeostasis may play a critical role for survival of *B. dermatitidis in vivo*.

Genes that convert reactive oxygen species to dioxygen and dioxygen to metabolites were also highly upregulated *in vivo*. These include two superoxide dismutases (*SOD3*: BDFG_01204 and *SOD2*: BDFG_07895), which were even more upregulated *in vivo* than in macrophages. Four dioxygenases (BDFG_04184, BDFG_04185, BDFG_08059, BDFG_06504) were also upregulated *in vivo*, representing almost half of the dioxygenases found in the genome, which utilize dioxygen to drive amino acid catabolism. This set includes

4-hydroxyphenylpyruvate dioxygenase, (4-HPPD; BDFG_04184) and homogentisate 1,2-dioxygenase (BDFG_04185), which are involved with tyrosine catabolism [46]. Other upregulated oxygenases include indoleamine 2,3-dioxygenase (BDFG_06504) and squalene monooxygenase (*ERG1*—BDFG_07857), which are involved with tryptophan catabolism and ergosterol biosynthesis respectively. *ERG1* is a target of current antifungal drugs, including terbinafine. High *in vivo* expression of this gene may suggest that drugs targeting it may be more effective *in vivo* than *in vitro*.

Genes involved in cysteine biosynthesis and catabolism were highly upregulated during infection including cysteine synthase A (BDFG_02039) and cysteine dioxygenase (BDFG_08059). Cysteine synthase A may provide a large pool of synthesized cysteine for *B. dermatitidis* metabolism; the induced expression during infection was confirmed by qRT-PCR (S12 Fig). Based on orthology analysis, cysteine synthase A is absent from the genome of *H. capsulatum*, and previous studies have shown that *Histoplasma* yeast are auxotrophic for cysteine [47,48]. Cysteine dioxygenase catabolizes cysteine to L-cysteinesulfinic acid, an intermediate that can be used for taurine biosynthesis or metabolized to sulfite and pyruvate. A homolog of *C. albicans SSU1* (BDFG_06814), which encodes a sulfite efflux pump and is co-regulated with cysteine dioxygenase in *C. albicans* [49], was also upregulated in *B. dermatitidis*.

Transporters in fungi have been associated with an enhanced ability to remove harmful products as well as to survive on diverse nutrient sources, both of which could contribute to virulence and pathogenicity. Of the 72 genes upregulated *in vivo* during mouse infection, 11 are predicted transporters. These included the major facilitator type (MFS; BDFG_06068 –unknown function, BDFG_06042 –glycose transport, BDFG_02038 –unknown function), amino acid transporters (BDFG_02310, BDFG_07447) and metal transporters (zinc/iron transporters discussed above, BDFG_09159, BDFG_07269, and *NIC1* nickel transporter, BDFG_02449; S14 Table). This upregulation potentially reflects differences in metabolite and cofactor availability in the host relative to *in vitro* conditions.

Discussion

Phylogenetic position of *Blastomyces* spp. and *Emmonsia parva* and *E. crescens*

Our whole-genome based phylogenetic analysis recovered a sister-group relationship between *Blastomyces* spp. and *Emmonsia parva*, as previously reported from ribosomal DNA sequences [19,20]. However, our study placed *Histoplasma* as the next most basal species, and uniquely placed *E. crescens* between *Histoplasma* and the basal *Paracoccidioides* with strong bootstrap support. This more external position of *Paracoccidioides* compared to *Histoplasma* agrees with an earlier rDNA tree without *Emmonsia* [50]. Furthermore, gene support frequencies (GSF) were relatively high, and increased when we subsampled only well-supported genes, providing additional support for the topology presented here.

The polyphyletic nature of the non-human pathogen *Emmonsia* suggests substantial plasticity in regard to human pathogenesis in this group. Ancestral variation in the ability of these species to infect other mammals could then be associated with exaptation to human hosts. Additional diversity of *Emmonsia*, including the third described species, *E. pasteuriana* [51,52] and other closely related isolates [17] suggests that the full breadth of the *Emmonsia* genus may not be captured by the two isolates sequenced here. Interestingly, both *E. pasteuriana* and related isolates produce yeast cells at high temperature, rather than the adiaspores produced by *E. parva* and *E. crescens*. Further sequencing of *Emmonsia* species and other related strains may reveal additional patterns and trends in the evolution of the dimorphic fungi.

Genome expansion and segmentation: GC-poor isochore-like regions

The mosaicism observed here in the genome of *Blastomyces* differs substantially from that observed in other fungi and larger eukaryote genomes. While isochore-like GC-poor regions are unprecedented at this scale in fungal genomes described to date, there are parallels to the organization of *L. maculans*, though GC-poor regions occupy a smaller fraction of that genome [28]. Longer GC-poor isochores of more than 300 kb are commonly found in mammals and other vertebrates [53–55]. GC-poor isochores in vertebrates are often more stable over long evolutionary times [55,56] and have other properties such as lower gene expression [55] that do not appear to be shared by the GC-poor tracts of *B. dermatitidis* and *B. gilchristii* (S1 Text).

Characterization of repetitive sequence in GC-poor regions suggests these originated with a dramatic expansion of elements of the LTR/Gypsy category. Phylogenetic analysis suggests these elements swept through a lineage leading to the present-day *B. dermatitidis* and *B. gilchristii*, and to a lesser extent *Emmonsia parva*, and have further expanded during the diversification of *Blastomyces*. While *H. capsulatum* does not have such an expanded genome, or a sizable GC-poor component, and so appears less affected by gypsy expansion, *Histoplasma* may alternatively have more robust defense against repetitive elements or be less able to accommodate large amounts of repeats in its genome.

While GC-poor tracts have been particularly dynamic areas due to Gypsy element insertions during the recent evolution of *Blastomyces*, these regions appear typical in gene content and expression. Perhaps due to their recent origin, the GC-poor regions do not appear to have sequestered particular classes of genes such as secreted proteins or have other hallmarks of rapidly evolving gene content. The long GC-poor regions also include some well characterized genes involved in phase transitions and pathogenesis, including the *Blastomyces* yeast-specific gene *BYS1*, a marker of the phase transition to and from the yeast phase [15,24]. Reduced levels of synteny in the GC poor regions between *B. dermatitidis* and *B. gilchristii* could prevent effective meiotic recombination between the two lineages, further supporting their designation as separate species.

Functional diversity of gene content in *Blastomyces* and the other Ajellomycetaceae

Despite the large increase in genome size in *Blastomyces*, the total number of protein coding genes is only modestly expanded. *Blastomyces* and other sequenced species from the Ajellomycetaceae family, including the human primary pathogens *Histoplasma* and *Paracoccidioides*, have similar gene content with only a few gene loss or gain events that map to common functional classes. This stability suggests that more modest differences in gene content and sequence, as well as potential divergence of gene regulation, contribute to phenotypic differences between the species. Larger differences exist between the Ajellomycetaceae and other more divergent members of the Onygenales. There is no expansion of degradative proteases as previously noted for *Coccidioides* [57]; in fact, three peptidase families (M36, M43, and S8) are present at lower copy number in *Blastomyces* and the other Ajellomycetaceae. While *Blastomyces* contains a higher number of LysM proteins than the dimorphic Onygenales, the number is small relative to that found in Dermatophytes [58]. This analysis also identified candidate genes involved in host interaction, including proteins homologous to antigens in related fungi and a heme oxygenase that could release iron from host heme.

Features of *Blastomyces* gene expression in macrophages and *in vivo*

For yeast co-cultured with macrophages and yeast *in vivo*, some aspects of the transcriptional response were shared including response to oxidative stress and amino acid catabolism. Yeast co-cultured with macrophages showed upregulation of numerous genes involved in oxidative

reduction, which may play a major role in protecting *Blastomyces* from ROS. The macrophage phagosome is poor in glucose and amino acids, but rich in ROS [59,60]. *Blastomyces* is relatively resistant to ROS and virulence correlates with the ability to minimize ROS generation in innate immune cells [61,62]. The upregulation of superoxide dismutases (*SOD3*, *SOD2*) and catalase P may protect *B. dermatitidis* yeast against oxidative stress. In *H. capsulatum*, extracellular *SOD3* and intracellular catalase P, contribute to survival within macrophages [63,64]. Moreover, *SOD3* promotes *H. capsulatum* virulence in a murine model of pulmonary infection [63]. The upregulation of 4-HPPD, which is involved with pyomelanin biosynthesis, contributes to antioxidant defense and intracellular survival of *Penicillium marneffei* [65]. Inhibition of 4-HPPD in *P. brasiliensis* and *P. marneffei* blocks the phase transition to yeast at 37°C [65,66]. Furthermore, *in vivo* numerous dioxygenases were upregulated, suggesting that dioxide produced in response to ROS may be utilized for amino acid metabolism.

Changes in amino acid metabolism were prevalent in both the macrophage co-cultured and *in vivo Blastomyces*, suggesting the recycling of amino acids as an energy source (Results, S1 Text). In particular, the very specific increase in cysteine catabolism (cysteine dioxygenase) and biosynthesis (cysteine synthase A) during *in vivo* infection suggests that cysteine may be critical to virulence. In mice, deletion of cysteine dioxygenase (*CDG1*) in *C. albicans* results in attenuated virulence [49]. Furthermore, upregulation of sulfite efflux pump (*SSU1*), which is co-regulated with *CDG1* in *C. albicans*, could play a role in *B. dermatitidis* virulence during *in vivo* infection. Exported sulfite can destabilize host proteins by reducing disulfide bonds and facilitates the growth of dermatophytes on keratinized tissue [67]. How breakdown of tryptophan by indoleamine 2,3-dioxygenase (IDO), which can supply *de novo* nicotinamide adenine dinucleotide (NAD+), alters the fungal-host interaction is unknown. In cancer, tumor cells with increased expression of IDO may facilitate tryptophan depletion in the microenvironment to suppress the host immune response [68]. Infection with *H. capsulatum*, *P. brasiliensis*, and *C. albicans* upregulates host IDO activity, reduces fungal growth, impairs Th17 T-cell differentiation, and blunts excessive tissue inflammation [69–71].

The specific *in vivo* upregulation of genes that encode secreted proteins as well as those involved with transmembrane transport (e.g., amino acids, glucose), amino acid metabolism (e.g., cysteine), and metal acquisition (e.g., zinc, nickel) highlights virulence factors potentially being missed by *in vitro* studies and the importance of understanding nutrient and co-factor availability in any study system. Uptake of zinc and nickel, which serve as enzyme co-factors, contribute to virulence in *C. albicans* and *Cryptococcus neoformans* respectively [45,72]. *PRA1* encodes a secreted "zincophore" under alkaline and zinc-poor conditions that acts in concert with *ZRT1* to promote zinc acquisition and facilitate endothelial cell damage by *C. albicans* [45]. NIC1-mediated nickel uptake is critical for urease activity, which contributes to *C. neoformans* invasion of the central nervous system [72]. In *C. posadasii*, urease induces host tissue damage [73]. While genes involved with the acquisition of zinc (e.g., *ZRT1*, *ZRT2*, *ZAP1* homologs) and nickel (e.g., *NIC1* homolog) are largely conserved with other fungi, the absence of *PRA1* in *Histoplasma*, *Paracoccidioides*, *and Emmonsia* highlights recent evolutionary changes in zinc acquisition mechanisms in the family Ajellomycetaceae. This, in conjunction with differences in cysteine metabolism between *Blastomyces* and *Histoplasma*, suggest that despite the many common elements of dimorphism and pathogenesis, each genus of dimorphic fungi likely has unique nutritional requirements.

Methods

Selection of isolates for sequencing

Four strains of *Blastomyces* were sequenced: SLH14081 representing the new species *B. gilchristii*, and ER-3, ATCC18188 and ATCC26199 representing *B. dermatitidis*. The SLH14081 strain

is a highly virulent, clinical isolate that can cause disease in immunocompetent persons, while ER-3 was isolated from a woodpile and is hypovirulent in mice [21,22]. The remaining two strains are strain ATCC18188, a representative MAT 'alpha' isolate, and ATCC26199, a commonly used laboratory isolate.

Two species that are closely related to *Blastomyces*, that can cause pulmonary disease in rodents (adiaspiromycosis), were also sequenced: *Emmonsia parva* UAMH139 and *Emmonsia crescens* UAMH3008. These isolates were chosen for comparison as these species are not typically human pathogens, yet they are closely related to the three pathogenic dimorphic genera *Blastomyces*, *Histoplasma* and *Paracoccidioides*, with which they form a clade that is nested within the order Onygenales and represents the Ajellomycetaceae family [20].

Sequencing of *Blastomyces*, *E. parva* and *E. crescens*

Genomic DNA for sequencing was prepared from mycelial or yeast culture, using phenol/chloroform extraction. For the *Blastomyces* SLH14081 and ER-3 strains, whole genome shotgun sequence was obtained using Sanger technology on an ABI 3730 from a Fosmid (epiFOS) and two plasmid (pJAN and pOT) libraries. For *B. dermatitidis* ATCC18188, whole genome shotgun sequence was obtained from two small insert libraries (fragment and 1.5 kb) using Roche 454 technology and from a Fosmid library using Sanger technology. For *B. dermatitidis* ATCC26199 20X of sequence was generated using 454 technology from a small insert fragment library. In addition, a plasmid (pOT) and Fosmid (epiFOS) library were constructed and sequenced using Sanger technology by the Washington University Genome Center, generating a total of roughly 3.6X coverage.

For each *Emmonsia* species, a single library was used to generate 101 bp paired-end reads using Illumina technology on a Genome Analyzer II generating a total of 116X coverage for *E. parva* UAMH139 and 163X coverage for *E. crescens* UAMH3009. Libraries of average insert size of 639 bp for *E. parva* and of 686 bp for *E. crescens* were chosen based on the electropherograms obtained from Bioanalyzer. Sequencing of both *Emmonsia* genomes was performed at the Genomic Sequencing Laboratory, University of California, Berkeley.

Genome assemblies

Blastomyces strains SLH14081 and ER-3 were assembled with Arachne [74] (Assemblez Build 20080911). For *B. dermatitidis* ATCC18188, a hybrid assembly was generated with Newbler version 2.3. For *B. dermatitidis* ATCC26199, a hybrid assembly of the Sanger and 454 data was generated with Newbler version "MapAsmResearch-03/15/2010" with options-rip and -scaffold.

For the *Emmonsia* genomes, assemblies were generated using multiple programs, including the SOAPdenovo / GapCloser package [75], ABYSS [76] and Velvet [77]. SOAPdenovo assemblies were selected based on quality metrics. While assemblies with high *k* values increased the fraction of GC-poor regions represented in the assembly, optimal assembly of gene sequences were achieved using lower *k* values, based on comparing each assembly to gene sets of *Blastomyces* and other related dimorphic fungi using TBlastN. The assemblies for the *Emmonsia* genomes ($k = 27$ for *E. parva* and $k = 29$ for *E. crescens*) were processed using the program GAEMR (http://www.broadinstitute.org/software/gaemr/), where overall assembly metrics were used to select the best assembly considering both continuity and completeness, though these measures were lower than for genomes assembled from multiple libraries.

Optical mapping of *Blastomyces*

To validate the assembly of strain SLH14081 and anchor it onto chromosomes, we constructed an optical map, a single-molecule based ordered restriction map. The map of *B. gilchristii* strain

SLH14081 was constructed by OpGen using the restriction enzyme BsiWI (C^GTACG). The optical map consists of 16 linkage groups, with size ranging from 9.728 Mb to 730 kb. The total size of the map was estimated as 79.64 Mb in size, slightly larger than the 75.35 Mb genome assembly, likely due to repetitive element sequence missing from the assembly. A total of 36 assembly scaffolds covering 68.9 Mb were mapped based on shared restriction sites to the optical linkage groups (S2 Table).

RNA-Seq of ATCC26199 from yeast, mold, and infection conditions

To enable more accurate gene prediction and analyze gene expression, RNA was prepared and deeply sequenced from five conditions (yeast with or without macrophages in RPMI media, *in vivo* during murine pulmonary infection, and *in vitro* yeast and mold in *Histoplasma* macrophage media (HMM)) with two biological replicates per condition.

ATCC26199 yeast cells were co-cultured with bone marrow derived murine macrophages from C57BL/6 mice in RPMI media with 10% heat inactivated FBS and supplemented with penicillin (100 U) and streptomycin (100 ug) or incubated in this media alone. Yeast and macrophages were co-cultured using a ratio of one yeast for every two macrophages (MOI 0.5). The use of alveolar macrophages was precluded due to the large numbers of mice that would be needed to harvest these cells. Following inoculation of cell culture flasks with *B. dermatitidis* yeast, the co-cultures were incubated at 37°C for 24 hrs. The majority of the yeast were either single cells or cells with one bud (average 89%). The extent of macrophage internalization of yeast was measured using Uvitex staining to differentiate between extracellular and intracellular yeast. A total of 1,976 cells were counted across seven individual fields of view, with an average of 59% Uvitex negative (intracellular) and 41% Uvitex positive (extracellular). The majority of *B. dermatitidis* cells exhibited yeast morphology (> 96%); pseudohyphal growth occurred in 2.4% of co-cultured yeast and 3.7% of yeast grown in RPMI media without macrophages. Harvested yeast cells were flash frozen in liquid nitrogen, ground with a mortar and pestle into a fine powder, and RNA extracted using the phenol-guanidium thiocyanate-1-bromo-3-chloropropane extraction method [78].

For *in vivo* transcriptional profiling, C57BL/6 mice were infected with 2 x 10^3 *B. dermatitidis* ATCC26199 yeast cells intratracheally and monitored for signs and symptoms of infection [79]. Mice with euthanized by carbon dioxide at 17 days post infection and yeast were isolated from murine lung tissue using the technique developed by Marty et al. [80]. Briefly, excised lungs were homogenized in pre-chilled (4°C) double-distilled, sterile water (ddH_2O) supplemented with DNase I 10 μg/ml (Roche) using an Omni TH tissue homogenizer (Omni International, Kennesaw, GA), passed through a 40 μm cell strainer (ThermoFisher Scientific, Waltham, MA), and centrifuged at 770g for 5 minutes at 4°C. The supernatant and interface were removed using a serologic pipette and yeast pellet was washed with ice-cold ddH_2O and rapidly frozen in liquid nitrogen for RNA extraction. Time *ex vivo* was less than 30 minutes and samples were near-freezing (4°C) during all isolation steps. Quality control analyses using qRT-PCR demonstrated that the short *ex vivo* time (< 30 minutes) at 4°C minimized changes in transcript abundance that would have occurred if the samples were processed at higher temperatures or for a longer duration [80]. Total RNA isolated from *B. dermatitidis* yeast during pulmonary infection was divided into 2 pools of 5 mice each (pool #1 and pool #2).

In vitro yeast were incubated in liquid *Histoplasma* macrophage media (HMM) at 37°C while shaking [81]. The majority of cells had yeast morphology; less than 3.25% of cells grew as pseudohyphae. To generate mycelia, yeast cells were incubated in liquid HMM for 14 days at 22°C while shaking. Harvested yeast and mycelial cells were flash frozen in liquid nitrogen, ground with a mortar and pestle into a fine powder, and RNA extracted using the phenol-guanidium thiocyanate-1-bromo-3-chloropropane extraction method [78].

Total *B. dermatitidis* RNA from all samples (*in vivo*, *in vitro*, co-cultures) was treated with TurboDNase (Bio-Rad, Hurcules, CA) and cleaned using an RNeasy column (Qiagen). RNA integrity and quality was assessed using Nanodrop spectrophotometry, 0.8% agarose gel electrophoresis, and an Agilent Bioanalyzer (Agilent Technologies, Santa Clara, CA). RNA integrity numbers (RIN) for *in vivo* samples were > 7.5 (7.6 for pool #1, 7.8 for pool #2). RIN values for *in vitro* and co-cultures (including yeast only RPMI) were ≥ 8.7.

For RNA-Seq, poly-A mRNA was purified for each total RNA sample and strand-specific libraries prepared as previously described [82,83]; each library was sequenced using Illumina Technology, generating an average of 65,174,908 101 bp reads per sample. RNA-Seq was incorporated into gene prediction and used to detect differentially expressed transcripts as described below.

Genome annotation

For initial gene sets, a total of 38,405 ESTs generated from yeast and mycelial samples of ATCC26199 (Washington University) and from a normalized cDNA library of SLH14081 (Broad Institute) were used for gene prediction. To achieve better transcript coverage, strand-specific RNA-Seq data from 10 samples representing the above yeast, mold, and infection stages was assembled with the Inchworm component of Trinity [84] and processed with PASA [85] to generate a set of transcripts for gene prediction. Gene sets were generated by using EvidenceModeler (EVM) [85] to select the best gene call for a given locus from the gene prediction programs SNAP, Augustus, Geneid, and Genewise and from PASA RNA-Seq transcripts as previously described [85,86].

Project numbers and locus tag prefixes assigned to gene sets are as follows: *B. gilchristii* SLH14081 (PRJNA41099, locus tag prefix BDBG), *B. dermatitidis* ER-3 (PRJNA29171, prefix BDCG), ATCC18188 (PRJNA39265, prefix BDDG), and ATCC26199 (PRJNA39263, prefix BDFG); the *E. parva* UAMH139 (PRJNA178178, prefix EMPG) and *E. crescens* UAMH3008 (PRJNA178252, EMCG).

Expression profiling

RNA-Seq reads were aligned to the transcript sequences of *B. dermatitidis* strain ATCC26199 using Bowtie [87]. Transcript abundance was estimated using RSEM [88], TMM-normalized FPKM for each transcript were calculated, and differentially expressed transcripts were identified using edgeR [89], all as implemented in the Trinity package version r2013-2-25 [90]. To identify GO term enrichment of differentially expressed genes, we classified transcripts using Blast2GO [91] and then performed comparisons with Fisher's exact test. A 2-fold difference in FPKM values and a false discovery rate below 0.05 were used as a criteria for significant differential expression. To identify possible functions of the gene products of significantly differentially expressed parasitic-phase genes, protein homologs were assigned based on BLAST, Gene Ontology (GO) terms and protein family domains (PFAM, TIGRFAM).

Quantitative real-time PCR (qRT-PCR)

Total RNA was extracted from *B. dermatitidis* yeast as described above. One microgram of DNase-treated total RNA was converted to cDNA using iScript cDNA synthesis kit (Bio-Rad). qRT-PCR was performed with SsOFast EvaGreen Supermix (Bio-Rad) using a MyiQ real-time PCR detection system (Bio-Rad). Reactions were performed in triplicate using the following conditions: 1 cycle 95°C x 30 sec, followed by 40 cycles at 95°C for 5 sec, 60°C for 10 sec. Transcript abundance for genes of interest were normalized relative to the transcript abundance of GAPDH. Relative expression (RE) was calculated as $RE = 2^{-\Delta Ct}$, $\Delta Ct = Ct_{\text{gene of interest}} - Ct_{GAPDH}$ [92].

Primer sequences used were as follows: AATCCTTTGACAGTGAAAC (forward) and CCATAAATCTGCTACAACAG (reverse) for BDFG_03060, ACTGTCGGTGGAGAGAAG (forward) and ACTGGGGTGTTGTTGAAG (reverse) for BDFG 02965, GACTATCCCATC CACAAC (forward) and TACAGAGCGGAATCTTTG (reverse) for BDFG 05357, TTTGGCACTGGAGTTATG (forward) and TGCTTCGTAGTCTAAAGTC (reverse) for BDFG 09159, GTGCTACAACGGAGATAC (forward) and GATAACCACCACGAACAC (reverse) for BDFG 02039, ACCCCCGCTCCTCCATCTTC (forward) and GAGTAGCCC CACTCGTTGTCATACC (reverse) for BDBG_07959 (GAPDH).

Segmentation and identification of genes and repeats located in GC-poor tracts

We used the IsoFinder GC segmentation program (http://bioinfo2.ugr.es/oliver/isofinder; [93]) to segment all ER-3 and SLH14081 scaffolds into long homogeneous genomic regions (LHGRs). The option p2 (parametric/student *t*-test with different variances), a window size of 5 kb and a *p* value cutoff of 0.01 (*P* parameter 0.99) were chosen after evaluating *P* cutoffs between 0.95 and 0.99, and window sizes ranging between 3 and 5 kb. The final settings were chosen as they accommodated gene synteny between ER-3 and SLH14081 in the GC-poor segments, obviating the need to manually remove narrow GC peaks caused by short genic regions.

To identify the coordinates of the longer GC-poor and GC-rich tracts on the assemblies of *Blastomyces* strains ER-3 and SLH14081, we set the boundary between GC-poor and GC-rich at 38% GC, a value that is in the deep valley between the two components of these genomes' bimodal GC distribution. The deep valley is robust and persists over a wide range of window/segment sizes ranging up to > 60 kb (S4 Fig). We then grouped adjacent segments located between successive transitions (regime switches) across the 38% GC border. Islands of N's in the interior of the GC-poor tracts were retained, but those at the tract fringes (i.e., next to a jump across the 38% GC threshold) were not. This procedure yields a large-scale segmentation of all scaffolds into strictly alternating "GC-poor" and "GC-rich" tracts. The GC-poor tracts and genes in those regions are listed in S3 and S4 Tables, respectively; GC-rich tracts form the remainder of the assemblies. We performed MySQL joins to identify the genes and repeats (GFF files produced by RepeatMasker of elements from RepeatModeler) located entirely or partly in the GC-poor tracts.

Synteny analyses

DAGchainer [94] was used to identify syntenic blocks with a minimum of 6 genes, which were required to be in the same order and orientations in the compared genomes. Synteny plots were generated using a custom perl script, using the GDgraph library; code is available at https://github.com/gustavo11/syntenia. Geneious Pro was used to visualize smaller-scale syntenies within and among genome assemblies.

Recognition and characterization of repeats

De novo repetitive sequence in each assembly was identified using RepeatModeler version open-1.0.7 (www.repeatmasker.org/RepeatModeler.html). Copies of *de novo* repeats and fungal sequences from RepBase [95] were mapped in each assembly using RepeatMasker version open-3.2.8 (www.repeatmasker.org/). For phylogenetic analysis of gypsy elements, reverse transcriptase domains were identified from each element; matches to the PFAM RVT_1 domain were identified with HMMER (version 3.1b1) [96] for 6-frame translations of each element generated by EMBOSS transeq (version 6.5.7 with parameters-frame 6-clean Y) [97]. The best domain match for each element was selected, requiring 50% alignment coverage and c-

Evalue < 1e-5. The domains identified in *Blastomyces* SLH14081 (991 total) and ER-3 (1,296 total), *E. parva* (40 total), and similar Repbase gypsy elements (4 total) were aligned with MAFFT (version 6.717) [98], and a phylogeny estimated using FastTreeDP (version 2.1.8) [99]. Four large subgroups were identified and visualized using iTOL [100].

Identification and analysis of orthologs and phylogenetic analysis

A total of 16 genomes from the Onygenales order and three *Aspergillus* genomes were chosen for comparative analyses (S15 Table). These include the four *Blastomyces* (SLH14081, ATCC26199, ATCC18188, ER-3) and two *Emmonsia* species (UAMH139, UAMH3008) as well as the following: *Histoplasma capsulatum* WU24 (AAJI01000000), *H. capsulatum* G186AR (ABBS01000000), *Paracoccidioides lutzii* Pb01 (ABKH02000000), *P. brasiliensis* Pb03 (ABHV02000000), and *P. brasiliensis* Pb18 (ABKI02000000), *Coccidioides immitis* RS (AAEC00000000), *C. posadasii* C735 delta SOWgp (ACFW00000000), *Uncinocarpus reesii* 1704 (AAIW00000000), *Microsporum gypseum* CBS118893 (ABQE00000000), *Trichophyton rubrum* CBS118892 (ACPH01000000), *Aspergillus nidulans* FGSC A4 (AACD00000000), *A. flavus* NRRL3357 (AAIH00000000), *A. fumigatus* Af293 (AAHF01000000). OrthoMCL was used to cluster the protein-coding genes of the 19 chosen genomes by similarity.

To estimate the species phylogeny, a total of 2,062 orthologs present in a single copy in all of the 19 genomes were identified. Protein sequences of the 2,062 genes were aligned using MUSCLE, and a phylogeny was estimated from the concatenated alignments using RAxML v7.7.8 with model PROTCATWAG. To more closely examine the relationship of the *Blastomyces* isolates, single copy orthologs were identified in all four strains of *Blastomyces* and *E. parva*; the protein sequences of a total of 6,605 single copy orthologs were aligned using MUSCLE, and the resulting sequences replaced with the corresponding codons. A phylogeny was estimated from this nucleotide alignment using RAxML v7.3.3 with model GTRCAT. A total of 1,000 bootstrap replicates were used for each analysis. The level of support for the best RAxML tree was also evaluated using individual gene trees, by calculating the gene support frequency (GSF, [26]). A phylogeny was estimated and bootstrapped using the same parameters as for the concatenated sequence matrix, and gene trees with high bootstrap support at all nodes were then selected. A total of 162 gene trees were supported by at least 70% of bootstrap replicates at all nodes; the percent of gene trees supporting the RAxML best tree was calculated using RAxML and is shown in Fig 1. We also evaluated larger subsets of trees including those with 60% bootstrap support at all nodes, 50% bootstrap support, or all trees regardless of support, and found lower support respectively in each subset for our best tree.

To examine selective pressure on genes in GC-poor regions, we identified 7228 genes that were single copy in the four *Blastomyces* genomes from the OrthoMCL run. d_N/d_S values for each gene were computed on codon-based nucleotide alignments with the codeml module of PAML [101], using the one-ratio (M0) model.

Gene family and protein domain analysis

Genes were functionally annotated by assigning PFAM domains, GO terms, and KEGG classification. HMMER3 [96] was used to identify PFAM domains using release 27. GO terms were assigned using Blast2GO [91], with a minimum e-value of $1x10^{-10}$. Protein kinases were identified using Kinannote [102] and divergent FunK1 kinases were further identified using HMMER3. Secondary metabolite gene clusters were predicted with antiSMASH version 2.0.2 [103]. Genes were clustered using OrthoMCL [104] with a Markov inflation index of 1.5 and a maximum e-value of $1x10^{-5}$.

To identify functional enrichments in *Blastomyces* and other subsets of the 19 compared genomes, we used four gene classifications: OrthoMCL similarity clusters (see above), PFAM domains, KEGG pathways, and Gene Ontology (GO), including different hierarchy levels. A gene was considered to be a member of a given gene class when, respectively, the gene (a) belonged to the given OrthoMCL cluster, (b) contained at least one instance of the given PFAM domain in the encoded protein, (c) belonged to the given KEGG pathway, or (d) was tagged by the given GO label. Using a matrix of gene class counts for each classification type, we identified enrichment comparing two subsets of queried genomes using Fisher's exact test. Fisher's exact test was used to detect enrichment of PFAM, KEGG, or GO terms between groups of interest, and p-values were corrected for multiple comparisons [105]. Significant (corrected p-value < 0.05) PFAM and GO terms expansion or depletion was examined for three comparisons: Ajellomycetaceae compared to other Onygenales (S6 Table), pathogenic compared to non-pathogenic from Ajellomycetaceae (S9 Table), and *Blastomyces* compared to other Ajellomycetaceae; the only terms found to be expanded only in *Blastomyces* included nucleosome and zinc ion binding. No significant enrichment in KEGG terms was detected for these comparisons.

Supporting Information

S1 Fig. Optical map of *Blastomyces gilchristii* strain SLH14081.
(PNG)

S2 Fig. Conservation of core eukaryotic gene (CEG) set across *Blastomyces*, *Emmonsia*, and other compared genomes. The percent coverage of genes with significant Blast similarity is shown for alignments above and below the recommended 70% coverage threshold; matches with less than 70% coverage suggest these are partial gene structures.
(PDF)

S3 Fig. Phylogeny of *Blastomyces* and *Emmonsia parva*. Maximum likelihood tree of the four *Blastomyces* strains (ATCC26199, ATCC18188, ER-3, SLH14081) and *E. parva* (UAMH139) was inferred using RAxML based on the concatenated nucleotide sequence alignment 6,605 genes.
(PDF)

S4 Fig. GC frequency distributions (histograms) of overlapping fragments (windows, subsequences) of the genome assembly of *Blastomyces gilchristii* (SLH14081) *B. dermatitidis* (ER-3) and *Leptosphaeria maculans* (v23.1.3). Window sizes included 2, 8, 32, 64, 128 and 256 kb. Step size was 1/128 of the window size. The bin size of the histograms is approximately 0.1% GC. Horizontal axes show GC percent and vertical axes show relative frequencies.
(PDF)

S5 Fig. GC frequency distributions (histograms) of small overlapping fragments (windows, of 128 bp) of the genome assembly of *Blastomyces dermatitidis* (ER-3), *B. gilchristii* (SLH14081), *Emmonsia parva* (UAMH139), *E. crescens* (UAMH3008), *Histoplasma capsulatum* (WU24), *Paracoccidioides lutzii* (Pb01), *Coccidioides immitis* (RS3) and *Leptosphaeria maculans* (v23.1.3).
(PDF)

S6 Fig. Comparison of GC-poor insertions in an otherwise syntenic region of *Emmonsia parva* (UAHM139) and the four sequenced strains of *Blastomyces*. The example illustrates the intraspecific variability in presence/absence of GC-poor segments or 'inserts' and, even where their presence and location are conserved, the variability in their lengths. In **(A)** the

dotplot of one complete scaffold of *E. parva* aligned to *B. gilchristii* strain SLH14081 (top) and *B. dermatitidis* strain ER-3 (bottom). In **(B)** the corresponding location of the inserts and the length; only insertion sites that were >15 kb for at least one strain are shown. This 265 kb region of the *E. parva* genome, lacks intermediate-sized (>15 kb) or long inserts, allowing its use as a simple reference for marking positions.
(PDF)

S7 Fig. Comparison of the expression of GC-poor genes vs. GC-rich genes. (A) Box plot of the gene expression (log2(FPKM+1)) of the genes located in GC-rich regions (blue) and genes located in GC-poor regions (green) in all five conditions of the RNA-Seq experiment of *B. dermatitidis* strain ATCC26199. Histograms in **(B)** show in the *x*-axis the distribution of the gene expression (log2(FPKM+1)) of those genes according their location during mouse infection. Similar distribution was observed in the other four conditions.
(PDF)

S8 Fig. Distribution of known repeats families in GC-poor regions as compared with known repeats families in GC-rich regions in *Blastomyces* (ER-3). The list in the left box represent the first 20 LTR/Gypsy representing approximately 90% of the LTR/Gypsy family in the GC-poor regions.
(PDF)

S9 Fig. Phylogenetic characterization of Gypsy elements in *Blastomyces*. Four divergent clades of gypsy elements (**A**, **B**, Fig 2B and 2C) were identified from a phylogeny inferred using FastTreeDP from alignments of reverse transcriptase domains identified in gypsy elements of *B. dermatitidis* (ER-3), *B. gilchristii* (SLH14081) and *E. parva* (UAMH139). Each of the four clades is shown separately; **A**. Subgroup of 220 sequences includes non-ACa Repbase elements. **B**. Subgroup of 554 sequences specific to *Blastomyces*. The outer circle indicates strain specific duplications of four or more sequences.
(PDF)

S10 Fig. Eukaryotic protein kinase superfamily members (kinomes). The kinomes of *Blastomyces gilchristii* (Bg; SLH14081) and *B. dermatitidis* (Bd; ER-3, ATCC26199 and ATCC18188) were compared with *Emmonsia parva* (Ep; UAMH139), *E. crescens* (Ec; UAMH3008), *Paracoccidioides brasiliensis* (Pb; Pb18) and *Coccidioides immitis* (Ci; RS3). Kinases are classified into major groups shown as colored blocks. Abbreviations: AGC, protein kinases A; CAMK, calcium/calmodulin-dependent kinases; CK1, casein kinase 1; CMGC, cyclin-dependent kinases (CDK), mitogen-activated, glycogen-synthase, and CDK-like kinases; STE, sterile phenotype kinases; FunK1, fungal-specific kinase 1; PKL, protein kinase subdomain-containing proteins; STK, serine/threonine protein kinase; STE, sterile phenotype kinases; TKL, tyrosine kinases.
(PDF)

S11 Fig. Correlation coefficients of FPKM values between samples. Two biological replicates for each condition of the RNA-Seq of *Blastomyces dermatitidis* (ATCC26199).
(PDF)

S12 Fig. Quantitative real-time PCR (qRT-PCR) analysis. (A) qRT-PCR analysis of endo-1,3 (4)-β-glucanase (BDFG_03060) and catalase P (*CATP*; BDFG_02965) from *B. dermatitidis* ATCC26199 yeast cells co-cultured with macrophages (Macrophage) and yeast cells grown in the absence of macrophages (No Macrophage) at 37°C in RPMI. **(B)** qRT-PCR analysis of genes encoding a zinc-scavenging protein (*PRA1*; BDFG_05357), zinc transporter (*ZRT1*; BDFG_09159), and cysteine synthase A (*CSA*; BDFG_02039) from *B. dermatitidis* ATCC26199 yeast cells isolated during murine pulmonary infection (*in vivo*) and yeast cells

co-cultured with macrophages (Macrophage) in RPMI at 37°C. qRT-PCR data are from 2 experiments. Relative expression (RE) for the target gene was compared to GAPDH: RE = $2^{-\Delta Ct}$ = $2^{-(\text{gene of interest})-(\text{GAPDH})}$.
(EPS)

S1 Table. Phenotypic differences observed among *B. dermatitidis*, *E. parva* and *E. crescens*.
(DOCX)

S2 Table. Optical map information for *B. gilchristii* strain SLH14081.
(DOCX)

S3 Table. Coordinates of GC-poor tracts in *B. dermatitidis* ER-3 and *B. gilchristii* SLH14081.
(XLSX)

S4 Table. Genes in GC-poor tracts in *B. dermatitidis* ER-3 and *B. gilchristii* SLH14081.
(XLSX)

S5 Table. Significant PFAM and GO enrichments comparing genes in GC-poor regions.
(XLSX)

S6 Table. Significant gene class enrichments in Ajellomycetaceae compared to other Onygenales.
(XLSX)

S7 Table. Secondary metabolite gene clusters.
(DOCX)

S8 Table. Gene clusters conserved in *Blastomyces*, *Emmonsia*, *Histoplasma*, and *Paracoccidioides*.
(XLSX)

S9 Table. Significant gene class in human pathogenic from Ajellomycetaceae (Blastomyces/*Paracoccidioides*/*Histoplasma* vs *E. crescens*/*E. parva*).
(XLSX)

S10 Table. *Blastomyces* genes absent in *E. parva* and *E. crescens*.
(XLSX)

S11 Table. Gene expression levels.
(XLSX)

S12 Table. RNA-Seq mapping statistics.
(DOCX)

S13 Table. List of significantly upregulated genes in yeast-Macrophages as compared with yeast-RPMI.
(XLSX)

S14 Table. List of significantly upregulated genes *in vivo*.
(XLSX)

S15 Table. List of genomes of Onygenales and *Aspergillus* species compared in this study.
(DOCX)

S1 Text. Supplementary Notes. Possible biological meaning of the GC-poor tracts; Functional enrichment of genes in GC-poor tracts; The GATA transcription factor *SREB* and

siderophore use; Secondary metabolite biosynthesis clusters; Characterization of gypsy element expansion; Gene expression changes in amino acid metabolism.
(DOCX)

Acknowledgments

We thank the Broad Institute Genomics Platform for generating DNA and RNA sequence described here, Christina Raymond and Sinéad Chapman for coordinating the sequencing, and Li-Jun Ma and Matthew Henn for coordinating the white paper that included initial sequencing of *Blastomyces*. We thank Kevin Haub, Bob Fulton, Lucinda Fulton, and Pat Mix for generating sequence of the *Blastomyces* strain 26199. We also thank Gustavo Cerqueira for assistance with preparing the synteny image in Fig 4.

Author Contributions

Conceived and designed the experiments: GMG BSK JGM OKC WEG ERM CAC. Performed the experiments: GMG TDS AJM JCC. Analyzed the data: JFM CAD JEG JH LD VM MM OKC CAC SS SY MP QZ ZC. Contributed reagents/materials/analysis tools: EAW JWT. Wrote the paper: CAC JFM CAD JGM OKC GMG. Assembled Blastomyces genomes: SS SY. Assembled Emmonsia genomes: JEG JFM EM. Annotated Blastomyces genomes: MP QZ ZC. Annotated Emmonsia genomes: JEG JFM SG.

References

1. Pfaller MA, Diekema DJ. Epidemiology of invasive mycoses in North America. Crit Rev Microbiol. 2010; 36: 1–53. doi: 10.3109/10408410903241444 PMID: 20088682
2. Meece JK, Anderson JL, Fisher MC, Henk DA, Sloss BL, Reed KD. Population Genetic Structure of Clinical and Environmental Isolates of *Blastomyces dermatitidis*, Based on 27 Polymorphic Microsatellite Markers. Appl Environ Microbiol. 2011; 77: 5123–5131. doi: 10.1128/AEM.00258-11 PMID: 21705544
3. Brown EM, McTaggart LR, Zhang SX, Low DE, Stevens DA, Richardson SE. Phylogenetic Analysis Reveals a Cryptic Species *Blastomyces gilchristii*, sp. nov. within the Human Pathogenic Fungus *Blastomyces dermatitidis*. PLoS ONE. 2013; 8: e59237. doi: 10.1371/journal.pone.0059237 PMID: 23533607
4. Schwarz J, Baum GL. Blastomycosis. Am J Clin Pathol. 1951; 21: 999–1029. PMID: 14885118
5. Gauthier G, Klein BS. Insights into Fungal Morphogenesis and Immune Evasion: Fungal conidia, when situated in mammalian lungs, may switch from mold to pathogenic yeasts or spore-forming spherules. Microbe Wash DC. 2008; 3: 416–423. PMID: 20628478
6. Gauthier GM, Safdar N, Klein BS, Andes DR. Blastomycosis in solid organ transplant recipients. Transpl Infect Dis Off J Transplant Soc. 2007; 9: 310–317. doi: 10.1111/j.1399-3062.2007.00227.x
7. Pfister JR, Archer JR, Hersil S, Boers T, Reed KD, Meece JK, et al. Non-rural point source blastomycosis outbreak near a yard waste collection site. Clin Med Res. 2011; 9: 57–65. doi: 10.3121/cmr.2010.958 PMID: 20974888
8. Klein BS, Tebbets B. Dimorphism and virulence in fungi. Curr Opin Microbiol. 2007; 10: 314–9. doi: 10.1016/j.mib.2007.04.002 PMID: 17719267
9. Marr KA, Patterson T, Denning D. Aspergillosis: Pathogenesis, clinical manifestations, and therapy. Infect Dis Clin North Am. 2002; 16: 875–894. PMID: 12512185
10. Wisplinghoff H, Bischoff T, Tallent SM, Seifert H, Wenzel RP, Edmond MB. Nosocomial Bloodstream Infections in US Hospitals: Analysis of 24,179 Cases from a Prospective Nationwide Surveillance Study. Clin Infect Dis. 2004; 39: 309–317. doi: 10.1086/421946 PMID: 15306996
11. Park BJ, Wannemuehler KA, Marston BJ, Govender N, Pappas PG, Chiller TM. Estimation of the current global burden of cryptococcal meningitis among persons living with HIV/AIDS. AIDS Lond Engl. 2009; 23: 525–530. doi: 10.1097/QAD.0b013e328322ffac
12. Nemecek JC, Wuthrich M, Klein BS. Global control of dimorphism and virulence in fungi. Science. 2006; 312: 583–8. doi: 10.1126/science.1124105 PMID: 16645097

13. Brandhorst TT, Roy R, Wüthrich M, Nanjappa S, Filutowicz H, Galles K, et al. Structure and function of a fungal adhesin that binds heparin and mimics thrombospondin-1 by blocking T cell activation and effector function. PLoS Pathog. 2013; 9: e1003464. doi: 10.1371/journal.ppat.1003464 PMID: 23853587

14. Brandhorst TT, Wüthrich M, Warner T, Klein B. Targeted gene disruption reveals an adhesin indispensable for pathogenicity of *Blastomyces dermatitidis*. J Exp Med. 1999; 189: 1207–1216. PMID: 10209038

15. Burg EF, Smith LH. Cloning and characterization of bys1, a temperature-dependent cDNA specific to the yeast phase of the pathogenic dimorphic fungus *Blastomyces dermatitidis*. Infect Immun. 1994; 62: 2521–2528. PMID: 8188377

16. England DM, Hochholzer L. Adiaspiromycosis: an unusual fungal infection of the lung. Report of 11 cases. Am J Surg Pathol. 1993; 17: 876–886. PMID: 8352373

17. Kenyon C, Bonorchis K, Corcoran C, Meintjes G, Locketz M, Lehloenya R, et al. A Dimorphic Fungus Causing Disseminated Infection in South Africa. N Engl J Med. 2013; 369: 1416–1424. doi: 10.1056/NEJMoa1215460 PMID: 24106934

18. Anstead GM, Sutton DA, Graybill JR. Adiaspiromycosis causing respiratory failure and a review of human infections due to *Emmonsia* and *Chrysosporium* spp. J Clin Microbiol. 2012; 50: 1346–1354. doi: 10.1128/JCM.00226-11 PMID: 22259200

19. Peterson SW, Sigler L. Molecular genetic variation in *Emmonsia crescens* and *Emmonsia parva*, etiologic agents of adiaspiromycosis, and their phylogenetic relationship to *Blastomyces dermatitidis* (*Ajellomyces dermatitidis*) and other systemic fungal pathogens. J Clin Microbiol. 1998; 36: 2918–2925. PMID: 9738044

20. Untereiner WA, Scott JA, Naveau FA, Sigler L, Bachewich J, Angus A. The Ajellomycetaceae, a new family of vertebrate-associated Onygenales. Mycologia. 2004; 96: 812–821. PMID: 21148901

21. Baumgardner DJ, Paretsky DP. The *in vitro* isolation of *Blastomyces dermatitidis* from a woodpile in north central Wisconsin, USA. Med Mycol. 1999; 37: 163–168. PMID: 10421847

22. Sullivan TD, Rooney PJ, Klein BS. *Agrobacterium tumefaciens* integrates transfer DNA into single chromosomal sites of dimorphic fungi and yields homokaryotic progeny from multinucleate yeast. Eukaryot Cell. 2002; 1: 895–905. PMID: 12477790

23. Li W, Sullivan TD, Walton E, Averette AF, Sakthikumar S, Cuomo CA, et al. Identification of the mating-type (MAT) locus that controls sexual reproduction of *Blastomyces dermatitidis*. Eukaryot Cell. 2013; 12: 109–117. doi: 10.1128/EC.00249-12 PMID: 23143684

24. Krajaejun T, Wüthrich M, Gauthier GM, Warner TF, Sullivan TD, Klein BS. Discordant influence of *Blastomyces dermatitidis* yeast-phase-specific gene BYS1 on morphogenesis and virulence. Infect Immun. 2010; 78: 2522–2528. doi: 10.1128/IAI.01328-09 PMID: 20368350

25. Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 2006; 22: 2688–90. doi: 10.1093/bioinformatics/btl446 PMID: 16928733

26. Salichos L, Rokas A. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature. 2013; 497: 327–331. doi: 10.1038/nature12130 PMID: 23657258

27. Bawdon RE, Garrison RG, Fina LR. Deoxyribonucleic acid base composition of the yeastlike and mycelial phases of *Histoplasma capsulatum* and *Blastomyces dermatitidis*. J Bacteriol. 1972; 111: 593–596. PMID: 5053471

28. Rouxel T, Grandaubert J, Hane JK, Hoede C, van de Wouw AP, Couloux A, et al. Effector diversification within compartments of the *Leptosphaeria maculans* genome affected by Repeat-Induced Point mutations. Nat Commun. 2011; 2: 202. doi: 10.1038/ncomms1189 PMID: 21326234

29. Desjardins CA, Champion MD, Holder JW, Muszewska A, Goldberg J, Bailao AM, et al. Comparative genomic analysis of human fungal pathogens causing paracoccidioidomycosis. PLoS Genet. 2011; 7: e1002345. doi: 10.1371/journal.pgen.1002345 PMID: 22046142

30. Kapitonov VV, Jurka J. A universal classification of eukaryotic transposable elements implemented in Repbase. Nat Rev Genet. 2008; 9: 411–2; author reply 414. doi: 10.1038/nrg2165-c1 PMID: 18421312

31. Daboussi MJ, Capy P. Transposable elements in filamentous fungi. Annu Rev Microbiol. 2003; 57: 275–99. PMID: 14527280

32. Muszewska A, Hoffman-Sommer M, Grynberg M. LTR Retrotransposons in Fungi. PLoS ONE. 2011; 6: e29425. doi: 10.1371/journal.pone.0029425 PMID: 22242120

33. McEwen JG, Ortiz BL, García AM, Florez AM, Botero S, Restrepo A. Molecular cloning, nucleotide sequencing, and characterization of a 27-kDa antigenic protein from *Paracoccidioides brasiliensis*. Fungal Genet Biol FG B. 1996; 20: 125–131. PMID: 8810517

34. Yuen KY, Chan CM, Chan KM, Woo PC, Che XY, Leung AS, et al. Characterization of AFMP1: a novel target for serodiagnosis of aspergillosis. J Clin Microbiol. 2001; 39: 3830–3837. PMID: 11682494

35. Gauthier GM, Sullivan TD, Gallardo SS, Brandhorst TT, Vanden Wymelenberg AJ, Cuomo CA, et al. SREB, a GATA transcription factor that directs disparate fates in *Blastomyces dermatitidis* including morphogenesis and siderophore biosynthesis. PLoS Pathog. 2010; 6: e1000846. doi: 10.1371/journal.ppat.1000846 PMID: 20368971

36. Esteban PF, Ríos I, García R, Dueñas E, Plá J, Sánchez M, et al. Characterization of the CaENG1 gene encoding an endo-1,3-beta-glucanase involved in cell separation in *Candida albicans*. Curr Microbiol. 2005; 51: 385–392. PMID: 16328626

37. Rooney PJ, Sullivan TD, Klein BS. Selective expression of the virulence factor BAD1 upon morphogenesis to the pathogenic yeast form of *Blastomyces dermatitidis*: evidence for transcriptional regulation by a conserved mechanism. Mol Microbiol. 2001; 39: 875–889. PMID: 11251809

38. Beyhan S, Gutierrez M, Voorhies M, Sil A. A temperature-responsive network links cell shape and virulence traits in a primary fungal pathogen. PLoS Biol. 2013; 11: e1001614. doi: 10.1371/journal.pbio.1001614 PMID: 23935449

39. Nguyen VQ, Sil A. Temperature-induced switch to the pathogenic yeast form of *Histoplasma capsulatum* requires Ryp1, a conserved transcriptional regulator. Proc Natl Acad Sci U A. 2008; 105: 4880–5. doi: 10.1073/pnas.0710448105

40. Webster RH, Sil A. Conserved factors Ryp2 and Ryp3 control cell morphology and infectious spore formation in the fungal pathogen *Histoplasma capsulatum*. Proc Natl Acad Sci U A. 2008; 105: 14573–8. doi: 10.1073/pnas.0806221105

41. Boyce KJ, Schreider L, Kirszenblat L, Andrianopoulos A. The two-component histidine kinases DrkA and SlnA are required for in vivo growth in the human pathogen *Penicillium marneffei*. Mol Microbiol. 2011; 82: 1164–1184. doi: 10.1111/j.1365-2958.2011.07878.x PMID: 22059885

42. Weissman Z, Kornitzer D. A family of *Candida* cell surface haem-binding proteins involved in haemin and haemoglobin-iron utilization. Mol Microbiol. 2004; 53: 1209–1220. doi: 10.1111/j.1365-2958.2004.04199.x PMID: 15306022

43. Sentandreu M, Elorza MV, Sentandreu R, Fonzi WA. Cloning and characterization of PRA1, a gene encoding a novel pH-regulated antigen of *Candida albicans*. J Bacteriol. 1998; 180: 282–289. PMID: 9440517

44. Lamb TM, Xu W, Diamond A, Mitchell AP. Alkaline response genes of *Saccharomyces cerevisiae* and their relationship to the RIM101 pathway. J Biol Chem. 2001; 276: 1850–1856. PMID: 11050096

45. Citiulo F, Jacobsen ID, Miramón P, Schild L, Brunke S, Zipfel P, et al. *Candida albicans* scavenges host zinc via Pra1 during endothelial invasion. PLoS Pathog. 2012; 8: e1002777. doi: 10.1371/journal.ppat.1002777 PMID: 22761575

46. Schmaler-Ripcke J, Sugareva V, Gebhardt P, Winkler R, Kniemeyer O, Heinekamp T, et al. Production of pyomelanin, a second type of melanin, via the tyrosine degradation pathway in *Aspergillus fumigatus*. Appl Environ Microbiol. 2009; 75: 493–503. doi: 10.1128/AEM.02077-08 PMID: 19028908

47. Boguslawski G, Akagi JM, Ward LG. Possible role for cysteine biosynthesis in conversion from mycelial to yeast form of *Histoplasma capsulatum*. Nature. 1976; 261: 336–338. PMID: 1272413

48. Howard DH, Dabrowa N, Otto V, Rhodes J. Cysteine transport and sulfite reductase activity in a germination-defective mutant of *Histoplasma capsulatum*. J Bacteriol. 1980; 141: 417–421. PMID: 7354005

49. Hennicke F, Grumbt M, Lermann U, Ueberschaar N, Palige K, Böttcher B, et al. Factors supporting cysteine tolerance and sulfite production in *Candida albicans*. Eukaryot Cell. 2013; 12: 604–613. doi: 10.1128/EC.00336-12 PMID: 23417561

50. Herr RA, Tarcha EJ, Taborda PR, Taylor JW, Ajello L, Mendoza L. Phylogenetic Analysis of *Lacazia loboi* Places This Previously Uncharacterized Pathogen within the Dimorphic Onygenales. J Clin Microbiol. 2001; 39: 309–314. PMID: 11136789

51. Gori S, Drouhet E, Gueho E, Huerre M, Lofaro A, Parenti M, et al. Cutaneous disseminated mycosis in a patient with AIDS due to a new dimorphic fungus. J Mycol Med. 1998; 8: 57–63.

52. Drouhet E, Gu Ho E, Gori S, Huerre M, Provost F, Borgers M, et al. Mycological, Ultrastructural and Experimental Aspects of a New Dimorphic Fungus *Emmonsia Pasteuriana* sp. nov. Isolated From a Cutaneous Disseminated Mycosis in AIDS. J Mycol Médicale. 2008; 8: 64–77.

53. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001; 409: 860–921. PMID: 11237011

54. Costantini M, Clay O, Auletta F, Bernardi G. An isochore map of human chromosomes. Genome Res. 2006; 16: 536–541. doi: 10.1101/gr.4910606 PMID: 16597586

55. Bernardi G. The neoselectionist theory of genome evolution. Proc Natl Acad Sci. 2007; 104: 8385–8390. doi: 10.1073/pnas.0701652104 PMID: 17494746

56. Pavlícek A, Paces J, Clay O, Bernardi G. A compact view of isochores in the draft human genome sequence. FEBS Lett. 2002; 511: 165–169. PMID: 11821069

57. Sharpton TJ, Stajich JE, Rounsley SD, Gardner MJ, Wortman JR, Jordar VS, et al. Comparative genomic analyses of the human fungal pathogens *Coccidioides* and their relatives. Genome Res. 2009; 19: 1722–31. doi: 10.1101/gr.087551.108 PMID: 19717792

58. Martinez DA, Oliver BG, Gräser Y, Goldberg JM, Li W, Martinez-Rossi NM, et al. Comparative genome analysis of *Trichophyton rubrum* and related dermatophytes reveals candidate genes involved in infection. mBio. 2012; 3: e00259–00212. doi: 10.1128/mBio.00259-12 PMID: 22951933

59. Tavares AHFP, Silva SS, Dantas A, Campos EG, Andrade RV, Maranhão AQ, et al. Early transcriptional response of *Paracoccidioides brasiliensis* upon internalization by murine macrophages. Microbes Infect Inst Pasteur. 2007; 9: 583–590. doi: 10.1016/j.micinf.2007.01.024

60. Isaac DT, Coady A, Van Prooyen N, Sil A. The 3-hydroxy-methylglutaryl coenzyme A lyase HCL1 is required for macrophage colonization by human fungal pathogen *Histoplasma capsulatum*. Infect Immun. 2013; 81: 411–420. doi: 10.1128/IAI.00833-12 PMID: 23184522

61. Sugar AM, Chahal RS, Brummer E, Stevens DA. Susceptibility of *Blastomyces dermatitidis* strains to products of oxidative metabolism. Infect Immun. 1983; 41: 908–912. PMID: 6885169

62. Morrison CJ, Stevens DA. Mechanisms of fungal pathogenicity: correlation of virulence in vivo, susceptibility to killing by polymorphonuclear neutrophils in vitro, and neutrophil superoxide anion induction among *Blastomyces dermatitidis* isolates. Infect Immun. 1991; 59: 2744–2749. PMID: 1649799

63. Youseff BH, Holbrook ED, Smolnycki KA, Rappleye CA. Extracellular superoxide dismutase protects *Histoplasma* yeast cells from host-derived oxidative stress. PLoS Pathog. 2012; 8: e1002713. doi: 10.1371/journal.ppat.1002713 PMID: 22615571

64. Holbrook ED, Smolnycki KA, Youseff BH, Rappleye CA. Redundant catalases detoxify phagocyte reactive oxygen and facilitate *Histoplasma capsulatum* pathogenesis. Infect Immun. 2013; 81: 2334–2346. doi: 10.1128/IAI.00173-13 PMID: 23589579

65. Boyce KJ, McLauchlan A, Schreider L, Andrianopoulos A. Intracellular Growth Is Dependent on Tyrosine Catabolism in the Dimorphic Fungal Pathogen *Penicillium marneffei*. PLoS Pathog. 2015; 11: e1004790. doi: 10.1371/journal.ppat.1004790 PMID: 25812137

66. Nunes LR, Costa de Oliveira R, Leite DB, da Silva VS, dos Reis Marques E, da Silva Ferreira ME, et al. Transcriptome analysis of *Paracoccidioides brasiliensis* cells undergoing mycelium-to-yeast transition. Eukaryot Cell. 2005; 4: 2115–28. PMID: 16339729

67. Grumbt M, Monod M, Yamada T, Hertweck C, Kunert J, Staib P. Keratin degradation by dermatophytes relies on cysteine dioxygenase and a sulfite efflux pump. J Invest Dermatol. 2013; 133: 1550–1555. doi: 10.1038/jid.2013.41 PMID: 23353986

68. Uyttenhove C, Pilotte L, Théate I, Stroobant V, Colau D, Parmentier N, et al. Evidence for a tumoral immune resistance mechanism based on tryptophan degradation by indoleamine 2,3-dioxygenase. Nat Med. 2003; 9: 1269–1274. PMID: 14502282

69. De Luca A, Carvalho A, Cunha C, Iannitti RG, Pitzurra L, Giovannini G, et al. IL-22 and IDO1 affect immunity and tolerance to murine and human vaginal candidiasis. PLoS Pathog. 2013; 9: e1003486. doi: 10.1371/journal.ppat.1003486 PMID: 23853597

70. Hage CA, Horan DJ, Durkin M, Connolly P, Desta Z, Skaar TC, et al. *Histoplasma capsulatum* preferentially induces IDO in the lung. Med Mycol. 2013; 51: 270–279. doi: 10.3109/13693786.2012.710857 PMID: 23181600

71. Araújo EF, Loures FV, Bazan SB, Feriotti C, Pina A, Schanoski AS, et al. Indoleamine 2,3-dioxygenase controls fungal loads and immunity in Paracoccidioidomicosis but is more important to susceptible than resistant hosts. PLoS Negl Trop Dis. 2014; 8: e3330. doi: 10.1371/journal.pntd.0003330 PMID: 25411790

72. Singh A, Panting RJ, Varma A, Saijo T, Waldron KJ, Jong A, et al. Factors required for activation of urease as a virulence determinant in *Cryptococcus neoformans*. mBio. 2013; 4: e00220–00213. doi: 10.1128/mBio.00220-13 PMID: 23653445

73. Mirbod-Donovan F, Schaller R, Hung C-Y, Xue J, Reichard U, Cole GT. Urease produced by *Coccidioides posadasii* contributes to the virulence of this respiratory pathogen. Infect Immun. 2006; 74: 504–515. PMID: 16369007

74. Jaffe DB, Butler J, Gnerre S, Mauceli E, Lindblad-Toh K, Mesirov JP, et al. Whole-genome sequence assembly for mammalian genomes: Arachne 2. Genome Res. 2003; 13: 91–6. PMID: 12529310

Chapter 7

From NGS assembly challenges to instability of fungal mitochondrial genomes: A case study in genome complexity

Computational Biology and Chemistry 61 (2016) 258–269

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Research Article

From NGS assembly challenges to instability of fungal mitochondrial genomes: A case study in genome complexity

Elizabeth Misas[a,b], José Fernando Muñoz[a,b], Juan Esteban Gallo[a,c], Juan Guillermo McEwen[a,d], Oliver Keatinge Clay[a,e,*]

[a] Cellular & Molecular Biology Unit, Corporación para Investigaciones Biológicas, Medellín, Colombia
[b] Institute of Biology, Universidad de Antioquia, Medellín, Colombia
[c] Doctoral Program in Biomedical Sciences, Universidad del Rosario, Bogotá, Colombia
[d] School of Medicine, Universidad de Antioquia, Medellín, Colombia
[e] School of Medicine and Health Sciences, Universidad del Rosario, Bogotá, Colombia

ARTICLE INFO

Article history:
Received 31 July 2015
Received in revised form 3 February 2016
Accepted 16 February 2016
Available online 21 February 2016

Keywords:
Next generation sequencing
Repetitive DNA
Fungal mitochondria
Genome assembly
DNA secondary structure

ABSTRACT

The presence of repetitive or non-unique DNA persisting over sizable regions of a eukaryotic genome can hinder the genome's successful *de novo* assembly from short reads: ambiguities in assigning genome locations to the non-unique subsequences can result in premature termination of contigs and thus over-fragmented assemblies. Fungal mitochondrial (mtDNA) genomes are compact (typically less than 100 kb), yet often contain short non-unique sequences that can be shown to impede their successful *de novo* assembly *in silico*. Such repeats can also confuse processes in the cell *in vivo*. A well-studied example is ectopic (out-of-register, illegitimate) recombination associated with repeat pairs, which can lead to deletion of functionally important genes that are located between the repeats. Repeats that remain conserved over micro- or macroevolutionary timescales despite such risks may indicate functionally or structurally (e.g., for replication) important regions. This principle could form the basis of a mining strategy for accelerating discovery of function in genome sequences. We present here our screening of a sample of 11 fully sequenced fungal mitochondrial genomes by observing where exact *k*-mer repeats occurred several times; initial analyses motivated us to focus on 17-mers occurring more than three times. Based on the diverse repeats we observe, we propose that such screening may serve as an efficient expedient for gaining a rapid but representative first insight into the repeat landscapes of sparsely characterized mitochondrial chromosomes. Our matching of the flagged repeats to previously reported regions of interest supports the idea that systems of persisting, non-trivial repeats in genomes can often highlight features meriting further attention.

1. Introduction

Repeats along a chromosome of a eukaryotic genome can severely inhibit the success of *de novo* or reference assemblies of the chromosome's sequence from short reads *in silico*, i.e., the stability of the assembly process. This problem is especially pronounced in next-generation sequencing (NGS) pipelines when the short (currently ≤300 bp) reads obtained are then piped through a de Bruijn graph-based *de novo* assembly program (Compeau et al., 2011). Such programs consider in a first instance the NGS reads' subsequences of a fixed length *k* (*k*-mers, words of length *k*), and use them to build up and analyze a quasi-flow on a de Bruijn or de Bruijn-like graph. Concurrently or in a second instance, such programs may then also integrate the information of the full read sequences *per se* and/or, where the reads are paired-end, the pairing information.

A good example of a genome or chromosome that illustrates the repeat assembly problem is a familiar and long-studied mitochondrial genome of a unicellular fungus, the S288C strain of baker's yeast, *Saccharomyces cerevisiae*. In a previous study (Muñoz et al., 2014), we fed subsequences of this completely sequenced mitochondrial genome, i.e., ideal, short, paired *in silico*-generated subsequences of read length 100 bp, without any sequencing errors and without any competing nuclear genomic DNA, to a widely used NGS assembly program. We confirmed what is perhaps already obvious on fundamental grounds: over a large range of *k* value

Abbreviations: NGS, next generation sequencing; mt, mitochondrial; mtDNA, mitochondrial DNA; nt, nucleotide; bp, base pair; kb, kilobase pair; Mb, megabase pair; *ori*, origin of replication; *orf*, open reading frame.

* Corresponding author.
E-mail address: oliver.clay@gmail.com (O.K. Clay).

http://dx.doi.org/10.1016/j.compbiolchem.2016.02.016
1476-9271/

choices ranging up to 63 bp, repeated sequences from the eight *ori* (origin of replication) regions consistently prevented full assembly: the *ori*'s were precisely the positions where the contigs or scaffolds could no longer be extended. In addition to fundamental (i.e., assembler-independent) limits that repeats can impose on the assembly process, there are also assembler-specific problems or complexities that can cause unexpected profiles of the assembly variation as one changes the *k* value (Gallo et al., 2014).

Persistent repeats preventing full *de novo* assembly, i.e., presenting limits to the contiguity of scaffolds/contigs that can be achieved using typical NGS pipelines, can be interpreted as compromising the *stability of the assembly process in silico*. The same repeats that can prevent a satisfactory assembly, such as the *ori* repeats in the mitochondrial genome of baker's yeast, can also compromise the *stability of the genome itself in vivo* or *in vitro*, i.e., they can destabilize processes within the cell and affect the structural stability of the genome (Bernardi, 2005). Just as a *de novo* short-read assembly program can get confused if it encounters a sequence of moderate length that is repeated (i.e., not unique), and may then terminate contig extension because it is not clear where the sequence continues, so a process of the cell such as recombination can apparently get confused at the same places in a mitochondrial genome because of ambiguity, and then create a deletion mutant in which DNA located between the two repeats is lost (such as in the well-studied petite-colony mutants in baker's yeast; Marotta et al. (1982)). In the human mitochondrial genome, a particularly common block deletion that entails clinical consequences, called the common deletion, deletes almost 5 kb. This may be a result of two copies of a 13-bp sequence at the ends of the deleted segment (Samuels et al., 2004), or the cause may be secondary structures co-localizing with those two copies rather than the copies' sequences themselves, as has been proposed based on deletional spectra (Guo et al., 2010).

In the present study, we focused on mitochondrial genomes of 11 unicellular fungi that appear to have been completely and reliably sequenced. We used a simple strategy, namely a search for recurring oligonucleotides (*w*-mers) of a fixed length, to screen for presence of possibly longer sequences that are repeated many times within the genome; an ultimate goal, towards which this study can offer only a first start, would be to understand the repeats and repeat systems in those and other mitochondrial genomes in the light of their possible function and evolution.

The working hypothesis motivating our study was that precisely the risks taken by a mitochondrion that continues to maintain sizable repeats, persisting in numerous copies that are interspersed around its relatively small genome, could often be a clue signaling functionally and/or structurally important sequence features at or flanking the repeats' genomic locations. Indeed, in the absence of any benefit, one might expect that natural selection will tend to eliminate variants exhibiting risky repetition, and to favor variants exhibiting no or only benign repetition. Such a principle, if consolidated by testing, could then be applied to efficiently 'mine for meaning', in a high throughput fashion, across mitochondrial or even nuclear genomes. As a start, we went through the steps of systematically screening our 11 chosen fungal mitochondrial genomes.

2. Materials and methods

2.1. Definition of mitochondrial genome

In this study we use the term "mitochondrial genome" as synonymous with "mitochondrial DNA genome" or "mtDNA genome". Most genome reports use this definition. However, some authors include also nuclear-encoded mitochondrial genes in the definition. Thus, Wallace (2013) defines the mitochondrial genome via the role, not the physical location, of the DNA and its genes, so that the "mitochondrial genome consists of thousands of copies of the maternally inherited mtDNA plus between 1000 and 2000 nDNA genes".

2.2. Species and strains

An objective of the present study was to gain a first insight into the presence and types of repeats in fungal mitochondrial genomes that can be highlighted using a simple strategy, namely a search for persistently recurring oligonucleotides of a fixed length. For reasons given below, we focused primarily on 17-mers. Since repeats can be the most difficult sequences to assemble from reads, we took care to select only fungi represented by a complete mitochondrial genome sequence that appeared reliable, with no obvious gaps that might correspond to unsequenced repeats. We therefore did not include partially sequenced/assembled genomes, in order to avoid mistaking 'absence of evidence' for 'evidence of absence' of repeats.

We sampled transversally across fungi associated with animals and well-studied model fungi, but did not include, for example, non-model fungi associated exclusively with plants, and our list of 11 mtDNA sequences does not include all complete mitochondrial genome sequences of such fungi now available (see, e.g., van de Sande, 2012; Joardar et al., 2012).

BLASTN searches showed that a short segment of the mitochondrial genome sequence of *Paracoccidioides brasiliensis* Pb18 was of non-fungal origin, and it was removed.

2.3. '17 × 4' criterion for persistent repeats

For the study conducted here, unless otherwise mentioned, we chose a pilot criterion that considers *persistent* repeats to be those sequences of length at least 17 bp that are repeated at least 4 times in a given mitochondrial genome. Fourfold repeats of sequences longer than 17 bp always correspond to two or more overlapping 17-bp repeats and are therefore also represented.

We justify our choice of 17 bp by our interest in (a) risk of genome-destabilizing ectopic recombination,[1] (b) risk of assembly-destabilizing repetition, and (c) secondary structures possibly associated with repeats. In these contexts, we note that (a) experimental, quantitative recombination studies consistently suggest that a perfectly repeated 17-mer should generally pose a risk of ectopic recombination, when the repeats occur within short distances similar to those found within a circular fungal mitochondrial genome; (b) some NGS short read assembly programs, such as SOAPdenovo2, offer *k*-mer size choices starting at around 17 bp; and (c) 17 bp is approximately the size of a conceivable, minimal compact stemloop/hairpin with flanks (e.g., flank 2 bp + stem 5 bp + loop 3 bp + stem 5 bp + flank 2 bp; see also Forsdyke (1995) and related oligonucleotide symmetry searches in (Bultrini et al., 2003; Baisnee et al., 2002)). We also note that a recent analysis of frequently repeated words in the human nuclear genome focused on words of a similar length, 15–16 bp (Zahradnik et al., 2014).

[1] We note that the mitochondrial processes that typically interest us here do not involve routine crossovers between two homologous chromosomes during meiosis, so it is not clear if borrowed terms such as 'ectopic', 'illegitimate', or 'out-of-register' recombination really make sense for intra-chromosomal recombination of (haploid) mitochondrial DNA. It is not evident what a corresponding 'non-ectopic' or 'in-register' recombination would be in such a context, as there is no expected, natural, routine or programmed 'in-register' mtDNA recombination event in the mitochondrion's life cycle to which one could refer as an obvious standard. We will however still borrow traditional metaphors such as 'ectopic' to refer to intra- or inter-chromosomal recombination between non-overlapping regions that have high sequence similarity.

Our cutoff at 4 copies was chosen to pick up only clear signals. We were interested where the 17-nt repeats' perseverance is unlikely to be transient, to be retained just by chance, or to have arisen via independent convergent mutations, as might conceivably be the case if only two copies of a 17-mer are present in the mitochondrial genome. In other words, our pilot criterion for fungi was chosen to screen for 'provocative' repetition in fungal mitochondrial genomes. It avoids false positives at the expense of missing repeats that could be relevant, e..g., our criterion misses the clinically important 13-bp repeat pair associated with the 'common' deletion in human (Samuels et al., 2004; Chen et al., 2011).

When counting repeats, we did not distinguish among possible relative orientations, e.g., between direct and inverted repeats.

2.4. Restriction to exact repeats

We deliberately focused almost exclusively on exact (100% identical) repetition of short (e.g., ≤40 bp) sequences, and not on less stringent (<100% identity) repeats as is customary in some other contexts. This was in order to keep the criteria and the analyses simple, and to avoid 'curses of dimensionality' when one simultaneously allows several freely variable parameters (number of mismatches allowed, gap creation and extension penalties, levels of similarity for mismatches, etc.). Other contexts require more relaxed stringency, e.g., when one explores distributions of (possibly degraded) transposons or other families of interspersed repeats in nuclear genomes.

In the initial pre-scanning of genomes we considered different possible minimum lengths w for the exact repeats. Afterwards, we used $w = 17$ bp throughout, for our 'proof of principle'. We did not explore the question of optimizing w for a given purpose in a given genome. We refer to (Li and Freudenberg, 2014b) for quantitative studies of the variation of w (or D) for the human nuclear genome, and a (possibly fundamental) duality between them and copy-number (C) variation analyses such as we present here.

2.5. Programs for counting k-mer repeats

Following Li et al. (2014), we used DSK (Rizk et al., 2013), and as an independent technical check also RepeatScout's pre-processing module `build_lmer_table` with the `-tandem 0` option (Price et al., 2005). These programs were used to extract, from each of the chosen mitochondrial genome sequences, all 17-mers (types) together with the number of times each of them were found in the sequence (number of tokens). We verified that both programs correctly manage reverse complements, i.e., that they correctly identify a token such as `ACGTACGTACGTACGTA` with its reverse complement `TACGTACGTACGTACGT` and count the two sequences together as belonging to *one* double-stranded 17-mer type (labelled by one of the two strands' sequences).

In contrast to RepeatScout's pre-processing module `build_lmer_table`, which tabulates k-mers and their frequencies and always gave correct results, the main part of RepeatScout uses additional criteria to then extend overlapping repeats. This second part of the RepeatScout program may not have been intended or optimized for the kind of maximal extension tasks (≳30 bp) and short initial repeat lengths (≳17 bp) that interested us. Indeed, RepeatScout often did not assemble the overlapping 17-mers to their actual maximal repeat sequence or extension. We therefore complemented exploratory use of RepeatScout with extension by eye and re-running of the `build_lmer_table` module for larger repeat sizes (from 18 bp to over 30 bp).

3. Results

3.1. Selection of whole mitochondrial genome sequences

An aim of this study was to explore a very simple word-size structured, exact-repeat based approach for flagging regions of fungal mitochondrial genomes that are likely to be enriched for potential functional or structural significance.

We focused on a small number of fungal mitochondrial genomes that were represented by relatively high quality, *bona fide* complete, annotated and/or curated mitochondrial genome sequences, i.e., genomes for which we could expect exhaustive k-mer lists and 'final' repeat statistics. In NGS short-read *de novo* assemblies, for example, precisely those regions with substantial repeats can be preferentially missing, so a partial NGS-derived fungal mitochondrial genome would give an unreliable picture of the actual genome's repeat structures (see Section 2.2).

In line with our laboratory's interest, we restricted our search to human pathogenic fungi and model fungi; we chose only one genome/strain per genus, with the exception of *Candida* where we chose two. We found that many fungi having sequenced nuclear genomes do not have complete high-quality mitochondrial genome sequences. (For comments on the more general phenomenon of neglected mitochondrial genomes, see Pesole et al. (2012), Picardi and Pesole (2012)).

Table 1 presents the 11 fungal mitochondrial genome sequences we chose, which range from 19.4 kb for *Schizosaccharomyces pombe* to 100.3 kb for *Podospora anserina*, and cover a representative range for fungi. For comparison, we also included in our analyses the mitochondrial reference genome of human Bandelt et al. (2014, and refs. therein; length 16.6 kb), to serve as an outgroup external to fungi.

3.2. Selection of pilot repeat criterion

To find a working criterion for our pilot study, we pre-scanned the selected 11 fungal mitochondrial genomes for perfect repeats of a fixed length. We looked for values w of that length that were neither too small, in the sense that repetition is inevitable, nor too large, in the sense that there is no repetition at all.

The pre-scanning suggested that a pilot criterion including only ≥17-nt repeats, occurring ≥4 times within the same mitochondrial genome, might be an appropriately poised choice for the mining task we consider. As we describe below, the criterion succeeds in picking up documented, main repeat types previously observed and/or characterized in our chosen mitochondrial genomes, and suggests additional ones that have not yet been characterized. The repeat length $w = 17$ nt is within the range of k-mer or 'window' sizes (de Bruijn graph dimensions) currently used by short-read based *de novo* assembly programs; furthermore, it is within the range of repeat lengths for which an illegitimate recombination event (possibly leading to deletion of functionally important DNA and genome instability) becomes a risk (see Section 2.3).

When we then applied this pilot criterion to scan the 11 selected fungal mitochondrial genomes, we observed a wide variability: from no repeats in some species, through very simple repeats, to rich and complex vocabularies or systems of highly non-trivial repeats, sometimes with clear propensity to form stable secondary structures as in baker's yeast. We therefore chose to report results for the '17 × 4' criterion in this study.

3.3. Basic repeat statistics for pilot '17 × 4' criterion

Table 2 reports some basic repeat statistics for each of the chosen mitochondrial genomes, using the '17 × 4' pilot criterion for persistent repeats. We included also two traditional, kinetic DNA complexity measures in the statistics for these genomes, calculated

Table 1
Mitochondrial genomes analyzed for 17-mer repeats in this study.

Abbr.	Species	Strain	Class	Accession	Length, kb
Afum	*Aspergillus fumigatus*	AF293	Eurotiomycetes	JQ346808	31.8
Calb	*Candida albicans*	SC5314	Saccharomycetes	NC_002653	40.4
Cgla	*Candida glabrata*	CBS138	Saccharomycetes	NC_004691	20.1
Ncra	*Neurospora crassa*	OR74A	Sordariomycetes	Broad-NC12	64.8
Pans	*Podospora anserina*	Race s	Sordariomycetes	NC_001329	100.3
Pbra	*Paracoccidioides brasiliensis*	Pb18	Eurotiomycetes	NC_007935	71.3
Scer	*Saccharomyces cerevisiae*	S288C	Saccharomycetes	NC_001224	85.8
Spom	*Schizosaccharomyces pombe*	ad7-50h-.	Schizosaccharomycetes	NC_001326	19.4
Tmar	*Talaromyces marneffei*[a]	MP1	Eurotiomycetes	NC_005256	35.4
TreS	*Trichoderma reesei*[b]	QM 9414	Sordariomycetes	NC_003388	42.1
Wcan	*Wickerhamomyces canadensis*[c]	21 *ade,his*	Saccharomycetes	D31785	27.7
Hsap	*Homo sapiens*	–	Mammalia	NC_012920	16.6

[a] Other name: *Penicillium marneffei*; *T. marneffei* is now the agreed name, also in clinical studies (Jorgensen et al., 2015).
[b] Other name: *Hypocrea jecorina.*
[c] Other names: *Hansenula wingei*, *Pichia canadensis.*

Table 2
Repetitiveness statistics of 17-mers in the selected mitochondrial genomes, using two early genome complexity measures[a] and simple counts (underlined entries indicate the most repetitive genomes).

Abbr.	Length, kb	Traditional complexities, kb		Complement, %		Max. frequency of a 17-mer	Persistent 17-mer types repeated ≥4×
		BK1968	D1976	BK1968	D1976		
Afum	31.8	31.3	30.9	1.5	2.8	4	21
Calb	40.4	33.2	26.6	17.9	34.2	8	227
Cgla	20.1	19.3	18.7	3.7	6.5	24	35
Ncra	64.8	60.4	58.3	6.8	10.0	124	324
Pans	100.3	99.4	98.6	0.9	1.7	8	25
Pbra	71.3	67.4	65.6	5.2	7.7	165	289
Scer	85.8	71.9	65.7	16.1	23.4	99	1,148
Spom	19.4	19.4	19.4	0.1	0.1	2	0
Tmar	35.4	35.3	35.2	0.4	0.7	18	2
TreS	42.1	42.0	41.8	0.3	0.6	3	0
Wcan	27.7	27.3	27.0	1.4	2.5	16	18
Hsap	16.6	16.5	16.5	0	0	0	0

[a] Two traditional kinetic complexity measures for DNA sequences, BK1968 (Britten and Kohne, 1968) and D1976 (Davidson, 1976), are adapted here for word size $w = 17$ bp and no mismatches, and reported together with their complements (i.e., 1 − complexity/length) expressed as percentages of the entire sequence. These assign high values to sequences with no repeats, and would give Scer one of the lowest complexities per Mb, among the fungal species shown here.

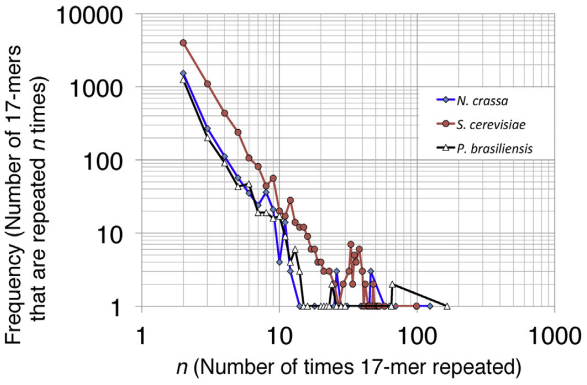


Fig. 1. Frequency plots of three fungal species (*Neurospora crassa*, *Saccharomyces cerevisiae* and *Paracoccidioides brasiliensis*) that exhibit persistent repetition at word size $w = 17$ bp, on double-logarithmic axes. It can be seen by eye that the slope of the almost 'linear' decrease on log-log scale, i.e., the exponent of the approximate power-law, is close to −3 and persists over about one decade. In this plot, values of n having zero frequency are skipped, and adjacent values with non-zero frequencies are directly joined by lines.

at word size 17 bp. These and some other complexity measures essentially indicate levels of repetitiveness: for a given genome size, less repetitive genomes are assigned higher traditional complexities. Early reassociation studies reported kinetic complexities for essentially $w \approx 400$ bp, allowing some mismatches (Britten and Kohne, 1968); in Table 2 we apply the original formulae with $w = 17$ bp, allowing no mismatches. At a repeat or word length of 17 bp, these and other repetitiveness measures calculated for Table 2 agree that 4 of the 12 genomes we examined, Calb, Ncra, Pbra, and Scer, have distinctly more repeats than the others. Indeed, the rightmost column of the table also reports, for each strain studied, the number of 'persistent' 17-mer *types* (including overlapping 17-mer types) that were represented four or more times in the strain's mitochondrial genome. All persistent 17-mer types that are present in a genome in 4 or more copies (i.e., 4 or more *tokens*) are listed in Supplementary Table S1, grouped first by species/strain (pages) and then by number of occurrences, i.e., tokens (columns).

Fig. 1 shows frequency plots of the three mitochondrial genomes having the highest yields of persistently repeated 17-mers: Scer (mtDNA genome size 85.8 kb), Pbra (size 71.3 kb) and Ncra (size 64.8 kb). For the double-logarithmic scales used in the figure, these genomes exhibit a linear decrease with a slope close to −3, which agrees well with slopes observed in the human nuclear genome (see Supplementary Material, section A4).

3.4. Alphabet usage

The three mitochondrial genomes that had no 4-fold persistent repeats at the 17-bp level were Spom, TreS and the non-fungal 'negative control' Hsap. Spom is of interest as a model example, as it was observed already decades ago that the nuclear genome of

Table 3
Examples of non-trivial repeats ('flora and fauna' of repeat landscapes) in selected mitochondrial genomes that were flagged by persistent 17-mers.[a]

Species	Examples of persistent repeats
Afum	37-mer sequence, repeated exactly 4 times: TATTAGTTAAGGTTACCTTACTTAAAGGGTTGAACAC
Pans	GCGCTATATATAGCGCTATATATAGCGC and its variants GCGCTATATATAGCGCAAGCTCCTC and GGCGCAAGCTCCTCCTC (so-called *Hha–Alu* 11-mer); TGTAAAAGAGCAACGAGTAGACGG; AGGATCCTCAGAGACTACACG; TCGACATACTTCGTCAGTA
Calb	Stemloops CCCGAGATCGTAGATCTCGGG (8×), CCCGTAGGGTGTTTCACACCCTACGGGGTT (6×)
Ncra	CCCTGCAGTACTGCAGGG (120×), 17-mers with cores GCAAGCTTGC or ACGGCT
Pbra	Many perfect or imperfect 3-, 4- and 5-nt repeats; GC-rich polypurine or polypyrimidine runs such as AGGAGGGGGAGGAGGGG / CCCCTCCTCCCCCTCCT
Scer	36-mer sequence of surrogate *ori* (*ori*^{S}), exactly repeated 26 times: aATAGTTCCGGGGCCCGGCCACGGGAGCCGGAACCCCgaAAGGAGa; conserved 21- and 18-mer blocks in alignment of the 8 *oris* (*ori*1,..., *ori*8): ATCACCCACCCCCTCCCCCTATt, GGGGTCCCAATTATTATT; uncharacterized sequences/GC clusters, especially: ACTCCTTCGGGGTCCGCCCCGC, CCCCGCGGGGCGGACCCC, GAAGGAGTGAGGGACCCC, AATAGTCCGGCCCGCCCCC, GGGAGGGGGACCGAACCCC, CCCCGAAAGGAGAAATA.

[a] The examples show consensus sequences of exact 4-fold repeats, or minor extensions of those consensus sequences that were supported at least 2-fold. Underlined regions and lower-case letters within the repeats are explained in the text.

used the NCBI MegaBLAST server (Graphics link) and Geneious to locate and visualize where each of the extended sequences had matches to the full Pans mt sequence. We also used the same protocol for other fungal species (below).

The 4-fold repeated (persistent) 17-mers included just one AT-only 17-mer, A_9T_8, but a variety of 17-mer types with high alphabet usage, including several that were subsequences of previously reported palindromes of length reaching up to 34 nt. Table 3 lists some of the extensions found; references and more details are given in Supplementary Material, subsection A6.2.

3.5.3. *Candida albicans* (Calb): most 17-mer repeats have no subrepeats

Fig. 2 showed that *C. albicans* (Calb) stands out compared to the other fungi studied here: almost all persistent 17-mers in its mitochondrial genome make use of the full 4-letter alphabet of DNA, i.e., Calb has essentially only the component of high alphabet usage (only three persistent 17-mers consisted entirely of AT). It is interesting that in this respect it is the opposite of the other *Candida* species we examined, *C. glabrata* (Cgla): the mitochondrial 17-mers of Cgla have only a low alphabet usage component and no high alphabet usage component.

The Calb mitochondrial genome landscape, as viewed *in silico*, has been described (Gerhold et al., 2010) as containing "potentially hairpin-forming sequences, GC clusters and nonanucleotide promoter-like sequences that might serve as sites for RNA priming", which are scattered along the mtDNA sequence rather than clustering into *ori/rep* or other structural loci as in Scer (cf. subsection 3.5.6). For matching of individual repetitive 17-mer types to some of these features, and a possible ambiguity in structure prediction, see Supplementary Material, subsection A6.3; two of the extensions of repeated 17-mers (of lengths 21 and 30 nt) that corresponded to previously noted likely hairpins are given in Table 3.

3.5.4. *Neurospora crassa* (Ncra): an amply repeated 18-mer, predicted to form the loop of much longer stemloops

The mitochondrial genome of Ncra is characterized by a rich repeat landscape, some of which was characterized and discussed already in the 1980's. For example, our most frequent 17-mer (repeated 124 times in the genome) defines a palindromic 18-mer motif, CCCTGCAGTACTGCAGGG (repeated 120 times). Longer GC-rich palindromic sequences, containing the 18-mer as their loop region, appear in the Ncra mtDNA and correspond to often impressively long stemloops (perfect or with few bulges) or other predicted secondary structures having possible roles in replication. The impressively long predicted structures were pointed out in 1981 Yin et al. (1981, see esp. Figs. 3 and 5), and were observed to flank tRNA genes; more results were obtained by Nargang et al. (1983). The positional distribution of the highly repeated 18-mer in the Ncra mtDNA (Fig. 3) shows that its presence, although clustered, is not restricted to few sector(s) of the genome.

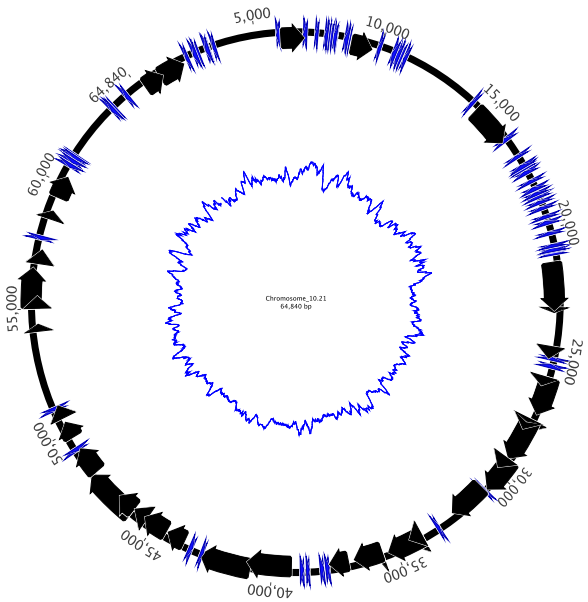


Fig. 3. Mitochondrial genome locations (shown as thin blue bars) of the full palindromic, 120-fold repeated, 18-mer motif CCCTGCAGTACTGCAGGG, that dominates the repeat landscape of *Neurospora crassa*. The 39 CDS segments of this genome's annotation are shown as black arrows; the inner ring shows GC fluctuations (GC increases outwards). (For interpretation of reference to color in this figure legend, the reader is referred to the web version of this article.)

When we selected the forty 17-mers that were repeated at least 10 times, stripped their ends of any perfect or nearly perfect single-nucleotide runs, and looked at the remaining subsequences of the 17-mers, we saw that they always formed part of one of three distinct consensus sequences: the 18-bp palindrome motif mentioned above (notably with no mismatches/variation), the 6-mer ACGGCT/AGCCGT, and the palindromic 12-mer GCAAGCTTGC.

3.5.5. Paracoccidioides brasiliensis (Pbra): almost all persistent 17-mers consist of short tandem subrepeats

Given the Pbra repeats' promising metrics, robust power-law trend and high alphabet usage (Figures 1 and 2), its persistent repeat motifs of length ≤17 bp were disappointingly simplistic. As in several other mt genomes, the frequently repeated 17-mers were mostly AT-only sequences. Thus, the 17-mers occurring 11–165 times were AT-only or, in very few cases, were imperfect AT-only with one C or G, or consisted entirely of $(AGG)_n/(CCT)_n$ triplet repeats. However, 17-mers occurring 4–10 times also showed the same tendencies, or were formed of other perfect or imperfect 3-, 4- or 5-nt repeats such as $(AAGG)_n/(CCTT)_n$, $(AAGT)_n/(ACTT)_n$, $(AGG)_n/(CCT)_n$, $(ACT)_n/(AGT)_n$, $(TCA)_n/(TGA)_n$, $(AGTA)_n/(TACT)_n$, $(CGAT)_n$ or $(AATTC)_n$, which in some cases were interrupted by single-nucleotide runs. There were also a few GC-rich sequences, which were however consistently purine-only (A's and G's) / pyrimidine-only (C's and T's).

3.5.6. Saccharomyces cerevisiae (Scer): persistent 17-mers are mainly in well-studied ori *repeats and GC clusters*

Since structural mutants, repeats, and/or potential secondary structures of the mtDNA genome of the model fungus *S. cerevisiae* have been given much attention over several decades, it is not surprising that many of the particularly abundant 17-mers we found corresponded to previously studied regions of this genome (see colored 17-mers in Supplementary Table S1; an unedited alignment of the genome's documented origin of replication or *ori* regions, which accounted for a large portion of the more persistent 17-mers, is shown in Supplementary Figure S1).

We applied a similar 'distilling' procedure to our repeated 17-mers as we had used for some other mitochondrial genomes: selectively removing 17-mers with exclusive AT usage and then screening the remaining yield of 17-mers using various criteria, such as having low ΔG, having high repetitiveness, and/or belonging to longer repeated supersequences that could be assembled exclusively from the obtained yield of 17-mers. Low (i.e., highly negative) ΔG flags those 17-mers that have higher propensity to alone form secondary structures, but cannot capture those contributing, e.g., to stem-halves of long stemloop structures.

A scatterplot illustrating screenings of the yield of 17-mers is shown in Fig. 4. The Figure shows a few examples from the most energetically favorable structures, and also how the tokens (instances) of the most highly repetitive (≥36-fold) 17-mers from the scatterplot are distributed around the 'master' chromosome: the strongest clustering is observed at or close to the *ori*'s (located as specified in the GenBank sequence).

Longer sequences reconstructed from the persistent 17-mers included classic *ori* sequences, surrogate *ori* sequences and classic GC cluster sequences (Bernardi, 2004) as well as a few sequences that to our knowledge have not been studied before; examples are shown in Table 3.

4. Discussion

4.1. Characterizing the large diversity of 11 fungal mtDNAs' repeat landscapes

We have analyzed repeated DNA in 11 fungal mitochondrial genome sequences that are considered complete, by focusing on persistent repeats of approximately 17 bp that appear at least four times in the same genome. The results we obtained are compatible with the idea that such repeated *k*-mers appear frequently enough, but not too frequently, in order for them to be useful as markers or flags of structurally and/or functionally interesting features. We have illustrated this for a number of fungi, where the regions they flagged often corresponded to regions that happened to have received particular attention by other authors in the past, in some cases largely irrespective of the regions' repeated occurrence. The fungi we chose for study include human pathogens (Afum, Calb, Cgla, the dermatophyte TreS and the thermally dimorphic pathogens Pbra and Tmar) and model organisms (Ncra, Pans, Scer and Spom).

Compared to fungi, mammalian mitochondrial genomes tend to have shorter or infrequently repeated non-unique sequences Lakshmanan et al. (2012, Fig. 1A). Repeats that have attracted interest in the model fly *Drosophila* occurred in tandem (Lewis et al., 1994; Kann et al., 1998; Rand, 2001). In angiosperm plants the sometimes large (e.g., > 500 kb) sizes and large noncoding DNA content of mtDNAs has provided opportunities for exploring their repeats, but perhaps because of size and task complexity, sequencing of entire mtDNA genomes has been slow (Kubo and Newton, 2008; Negruk, 2013), and exhaustive analysis of the mtDNAs' noncoding regions can be less straightforward than for fungi. In summary, fungi currently appear to be a good choice of taxonomic group in which to address repeat-related mtDNA questions.

We proceeded in our analyses of fungal mtDNA genomes and their repeats via stepwise elimination or 'shelling' of repeat types, which we grouped into 'shells' using a number of criteria that the frequency data suggested (or, in some species, dictated). The first shell to be removed was that of repeats that consist entirely, or with very few exceptions, of A and T; the abundance of these repeats was an mtDNA trait that differed among species, even within the same genus (Calb *vs.* Cgla), but in general such individual AT-only repeats or repeat stretches were not sufficiently rare or species-specific to appear useful as flags of structural or functional features, and we did not analyze them further. Subsequent shelling depended on the genome of interest, and its shellability: one criterion we used for shelling was decomposability of the 17-mers into smaller repeats; another expedient variable that revealed distinct groups of repeat types in some genomes was the magnitude of the free energy ($-\Delta G$) of the 17-mer's own folding. The inner shells often coincided with salient or species-specific features of previously noted structural or functional interest. Some of the 17-mers observed in these inner shells and their repetitiveness are understood; others still await follow-up analyses and explanation.

The procedure we followed in this study, and the diversity of the repeat 'flora and fauna' it allowed us to observe (viewed as linear sequences and/or as their likely folding patterns or shapes), led us to revisit notions of complexity that are often used for describing whole-genome and other eukaryotic sequences. We propose that a concept of complexity is needed here that goes beyond the most commonly used complexity measures, which typically penalize deviations from uniform word frequencies, or categorically consider that repetition should lower complexity. In a nutshell: repeats, if not present in overwhelming quantities, can serve roles that are able to facilitate or catalyze the genesis or growth of semantically complex systems. We propose the term 'repeat-associated complexities' for complexity notions that reflect this point of view. A discussion with examples is given in Supplementary Material, section A7.

4.2. Repeat-associated assembly instability may be a transient problem, but repeat-associated genome instability is not

While we were writing this manuscript, Illumina increased the maximum length of paired-end reads offered by its sequencing machines, from around 100–150 bp to 250 –300 bp. As Illumina and other NGS technologies advance, offering reads that are longer and/or cost less, the repeat-associated instability of genome assemblies caused by short reads could soon become a problem of the

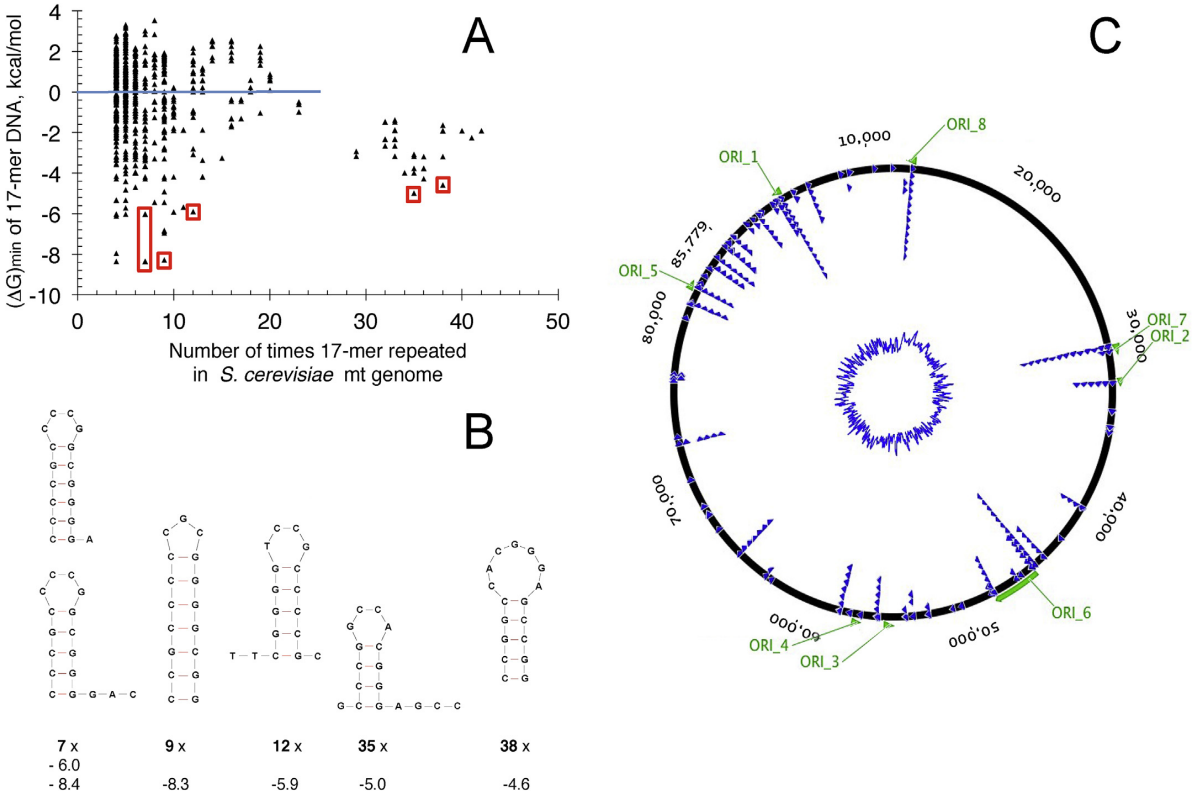


Fig. 4. Predicted secondary structures and chromosomal locations of some highly repeated 17-mers in the mitochondrial genome of *S. cerevisiae* strain S288C (`Scer`). (A) Scatterplot of minimum predicted Gibbs free energies ΔG_{min} of structures of persistent (at least 4-fold repeated) 17-mers of DNA of different repetition levels (token counts) and not including 17-mers consisting solely of A's and T's, as estimated by the `quikfold` program (Zuker, 2003). (B) Examples of some energetically favored predicted structures of individual 17-mers (boxed in panel A), with their ΔG estimates. (C) Locations of the most repeated 17-mers that did not consist only of A's and T's (repeated ≥36-fold; 457 tokens), showing their frequent presence at or near the genome's origins of replication (*oris*; plot made using Geneious Pro, with inner ring displaying GC% and the *ori* coordinates taken from the GenBank sequence).

past, even if the fascination of some theoretical and algorithmic questions surrounding short reads' *de novo* assembly may remain.

By contrast, some parallel problems of instability or loss of integrity of mtDNA genomes *in vivo* are likely to remain clinically relevant. Human mtDNA deletion-related detriments to health can be serious (Chen et al., 2011; Sequeira et al., 2012; Zaragoza et al., 2011). In fungal pathogens containing repeats that could lead to structural mutants via ectopic pairing, one could imagine virulence being reduced by some deletions or enhanced by others (in relation to mitochondrial activity changes that may be needed by the fungus for pathogenic activities, cf. e.g. Maresca et al., 1981); insights into fitness- or virulence-modulating mtDNA changes, e.g., affecting respiration, might ultimately translate to ways of managing infections.

4.3. Possible interpretations of persistent repeats in mitochondrial genomes

4.3.1. Persisting risks or other disadvantages can indicate presence of a tradeoff

We motivate this section with a parallel example, a hypothesis that also involves mitochondria, namely the CoRR hypothesis for the retention of genes in mtDNA (Allen and de Paula, 2013; Allen, 2003). The rationale for this hypothesis is summarized by Lane (2006, italics in the original): "there must be a very strong *positive* reason" why the existing mitochondrial genes were retained in the mitochondrial genome rather than having been moved to central control in the nuclear genome as one might have expected. The idea is that the genes that have persisted in the mtDNA "have not remained there by chance, but because natural selection has favoured their retention *despite* the manifold disadvantages". The proposed reason is that the speed of respiration is very sensitive to changing circumstances so, to respond effectively to abrupt changes, mitochondria "*need* to maintain a genetic outpost on site".

In a similar way, words in a DNA sequence that consistently occur far more often than expected may suggest that there is a biologically significant benefit or reason for their persisting repetitiveness (Pesole et al., 1992), especially where such repetition is known to be associated with potential disadvantages or risks. In this study, a biological risk that has guided some of our explorations is the risk of ectopic recombination or pairing, and the deletions of DNA segments that such pairing can entail. After addressing the risk itself, we address possible benefits of repetition that could balance those risks, i.e., tradeoff scenarios that might be relevant.

4.3.2. The risk of ectopic recombination-mediated deletion is not easy to quantify

The risk, or efficiency, of an ectopic pairing or recombination event in a genome that could result in the deletion of intervening

DNA has not been easy to quantify, or quantitatively predict, from only information contained in the genome's DNA sequence. Over decades, attempts in this direction have been published for dozens of systems, but the quantitative results on recombination frequencies or efficiencies from those systems have not yet converged to a generally applicable formula or procedure for estimating recombination probabilities directly from a genome sequence. A few of the many reports are Baker et al. (1996), Lichten and Haber (1989), Mezard et al. (1992), Sugawara and Haber (1992), Bollag et al. (1989). The variable findings from studies using different systems illustrates that it is not easy to find factors external to the sequence that could improve sequence-based risk prediction. In mtDNA and if a good predictor were available, risk of a deletion event might be best calculated from the repeats themselves (Samuels et al., 2004; Chen et al., 2011) and/or from joint secondary structures favoring pairing that can preferentially occur in the vicinity of the repeats (Guo et al., 2010).

Although quantification of risk remains difficult, the numerous studies focusing on ectopic recombination do, however, firmly consolidate the *presence* of an ectopic pairing risk, across many contexts and scenarios. The studies strongly suggest that such risks, and their possible consequences, should be anticipated when thinking about non-unique DNA copies located at relatively short distances from each other, as is the case in the relatively small mtDNA chromosomes.

4.3.3. Possible benefits of maintaining structural diversity of mtDNAs

We now consider ways in which risky (potentially genome-destabilizing) ectopic recombination or deletion events, which it would be advantageous to prevent in some circumstances, could in fact be beneficial under other circumstances.

Roles in mechanisms of replication and/or transcription have been invoked as one possible benefit of repeat presence in the mitochondrial genomes of some fungi, and we have considered them in some concrete fungal contexts in this article so far (see Results). Proof of relevance for replication appears to be strong in the well-studied case of the yeast *S. cerevisiae*, more tentative for some other fungi, and absent for the clinically important and also well-studied common deletion in human (see also Section 4.3.6 below).

The discovery of CRISPR sequences in prokaryotes, associated with the organism's self-defense against viruses, raises the question if any eukaryotic genomes (mtDNA or nuclear) might contain sequences derived from a similar or related defense function, and that have possibly now been adapted to a different role. In the context of repeats with stemloop potential, such we have described here in fungal mtDNA, we note that prokaryotic CRISPR sequences can contain two or more direct repeats, often around 30–40 bp or so in length, separated by spacer(s) of a similar length. The direct repeats are, in turn, often palindrome-like or with some dyad symmetry, possibly corresponding to stemloop-like structures. A few fungal mtDNA repeats observed in this study were located in sequences that appear compatible with a CRISPR sequence, but the agreement was not extensive enough to suggest a CRISPR-like or defense-derived function of repeated sequences in fungal mtDNA (see Supplementary Material for details and references).

In contrast to the possible or shown benefits or needs for replication, transcription, defense or other cellular processes, we now mention a different type of benefit that could in principle be associated with the robust presence of repeats in mitochondrial genomes. This hypothesized benefit partly extrapolates insights from angiosperms' mitochondrial genomes (e.g., Kubo and Newton, 2008 [2]), which can be much larger than those of most fungi (e.g., 588 kb in the legume *Vicia faba*; Negruk, 2013).

A strong form of the hypothesis we now consider is that selection could have maintained some repeat pairs because they provide, via ectopic pairing, a mechanism to easily switch between alternative structural forms when these become advantageous, for example when some environmental condition favors such a 'structural switch'. A consequence could be the creation of structural diversity among mtDNAs depending on environmental circumstances. In larger organisms such as human, a related principle might lead to structural heteroplasmy (an example of this class of phenomena, although not necessarily generated in this way, might be the condition-dependent dimeric molecules, branched structures and four-way junctions observed in adult cardiac muscle mtDNAs by Pohjoismäki et al. (2010)). A down side would be the creation of mtDNA deletion mutants causing clinical phenotypes in human. Such consequences are what we indeed observe, in human and (where applicable) in fungi.

In a related context, namely the stretches of tandem repeats often found in nuclear genomes, Trifonov (Trifonov, 1999, 2004; King et al., 2006) has introduced the metaphor of an evolutionary or adaptive "tuning knob" to emphasize how small changes in the repeats can produce fast adaptation. Tandem repeats can also, as their down side, cause clinical disease in human, as in the case of excessively expanding or 'runaway' trinucleotide repeats in the *FMR-1* gene/Fragile X syndrome or in Huntington's disease. Tandem repeats can expand or contract by slippage; non-tandem repeats, on which we focus in our study, cannot use this mechanism, but can typically only use ectopic pairing in order to produce fast adaptation, in this case by generating structural variants.

Interestingly, a recent analysis of 15- and 16-mer repeats (tandem and non-tandem) in the human nuclear genome by Zahradnik et al. (2014) showed that the most repeated words were various mono- and dinucleotide repeats, subsequences of *Alu* repeats, and the sequence $T_{11}GAGA/TCTCA_{11}$ and its truncations. The authors point out that such massively repeated words are likely to have been the *generator* sequences that then led also to a smaller number of mutated sequences, and propose as a conservative suggestion "that all high occurrences in the vocabulary are due to higher use of the words for whatever special intragenomic function they serve", a proposal that is paralleled by our view that frequency is likely to imply function also in mtDNAs.

4.3.4. Updating the traditional demand for a single complete genome sequence

The view considered here has consequences for the understanding of genome sequences. It suggests that the notion of a consensus or *master* structure or *master chromosome* (Kubo and Newton, 2008), constituting 'the' mitochondrial genome, may continue to be useful as an idealized placeholder, but that this concept may not satisfactorily mirror the existing, substantial structural diversity or structural heteroplasmy that can often be observed among mtDNA chromosomes of the same organism or isolate. The view postulates that, whether in fungi or human, we may sometimes need to consider the possibility of a *distribution* of different mtDNA chromosomal sizes, structures or forms (see also Supplementary Material, Section A5).

[2] In 2008 these authors could already summarize the situation for angiosperm mtDNAs as follows: "With one reported exception... each angiosperm master chromosome harbors one to several sets of repeated sequences where active recombination events occur between the repeat copies. Due to this property, isomeric forms of the master chromosome would be expected when the repeat copies of a set are present in inverted orientation to each other; alternatively, two subdivided molecules (subgenomes) would be expected when two copies of the same repeat are present in direct orientation."

A corollary is that for many species we may never succeed in sequencing *the* mitochondrial genome of a species or strain, because there may actually be several such viable genomes, which may be structurally different. Moreover, in some conditions a smaller variant(s) could be 'better' (e.g., replicate more efficiently when needed while continuing to serve some useful function) than the maximal sequence, which one might traditionally have thought of as being the only complete one, and for which one wished 'stability'. The maximal or 'complete' circular construct that has often interested us so much in the past may, however, be just one of many constructs or topologies (e.g., circular, linear or branched) that co-exist in an organism, or in a Petri dish of fungal colonies deriving from a single isolate (e.g., in co-existing petite and wild-type colonies of yeast).

4.3.5. Practical consequences of structural mtDNA diversity

The sooner we realize that a distribution, not a single 'wild type' or variant chromosome, could often be the relevant paradigm when we try to parse or assemble raw NGS mtDNA reads of a eukaryotic individual or sample, the quicker we will be able to appropriately adapt existing assembly tools or develop new ones in order to arrive at assembled sequences of the major structural types, without getting into a tug-of-war of apparently contradictory assemblies, of which some are believed to be more correct than others.

Furthermore, as the title of this article indicates, there may be a duality between the *in vivo* or *in vitro* situations in which repeats can confuse 'faithful' transmission of a chromosomal sequence to the future, e.g., via replication, and *in silico* situations in short-read *de novo* or reference assembly where repeats can confuse the reconstruction of a master chromosome sequence of a fungal mitochondrion. The duality may not stop here, however. Just as we are now imagining a distribution of possible structural chromosome variants, rearranged in accordance with the repeat pairs that could potentially favor the creation of those variants, we can also, in parallel, imagine a distribution of possible assemblies of a set of short NGS reads that is compatible with the information in the read set. To what extent the possible assemblies reflect the possible or observed genomes remains speculative at this point. Figuring out which (more global) assembly topologies or structures are compatible with a set of (more local) segments or fragments of those structures (e.g., short reads) is a classic problem that goes back at least as far as Benzer's early analyses (Benzer, 1959, 1962; Golumbic, 2004); for reconstructions of topologies in yeast mitochondria, see Rayko and Goursot (1999). This classic problem received much attention from combinatoricists in the 1970s and 1980s, e.g., in Golumbic (2004), Berge (1983), Ch. 16, and Berge (1987), Ch. 5. It might soon merit revisiting in light of the various interrelated open problems of structural mitogenomics we have briefly touched upon in this article (cf. also Supplementary Material, Section A6.3).

4.3.6. Factors influencing differential fitness of structural mutants

An initial ectopic pairing event may or may not lead to a viable mutant mtDNA chromosome that can survive, replicate and succeed. In yeast mitochondria, some petite mutants can survive even if they have only a single origin of replication (*ori* sequence) left but no genes at all (Bernardi, 2004, 2005; de Zamaroczy et al., 1981; Rand, 2001), and these can have replicative advantage over the wild-type. Conversely, in human mitochondria, large-scale deletion mutants that do not contain both the O_H and the O_L origins of replication are only very rarely observed (Chen et al., 2011). Because replication and replicative efficiency are so important, they provide hard constraints shaping the evolution of the population.

The common *ori*2–*ori*7 mtDNA deletion (Marotta et al., 1982) of the yeast *S. cerevisiae* might be seen as partly analogous to the common 13-base pairing mtDNA deletion in human, although in yeast the portion of the genome that is deleted is far smaller than in human, only around 2 kb, and contains no gene or other feature of known function (Foury et al., 1998). In *S. cerevisiae*, the origins of replication *ori*2 and *ori*7 have been associated with repetitiveness (see also Section 3.5.6 above) as well as with roles and utility in replication (reviewed with references in Bernardi (2004)). By contrast, to our knowledge the extensively studied 13-bp repeats in human mtDNA leading to the common deletion (Samuels et al., 2004), and the putative duplex pairing structures within 100 bp of the 13-bp repeats, which may be important for efficiency of the common deletion (Guo et al., 2010), have not been associated with a role in mtDNA replication in human; we also note that, unlike the *ori*2–*ori*7 context in yeast, the human origins of replication O_H and O_L are both far from the two 13-bp repeats.

4.3.7. Are repeats the primum movens *of ectopic pairing?*

An analysis of human deletion spectra by Guo et al. (2010) suggests that the 100-bp regions around the two 13-bp repeats of the human common mtDNA deletion might be even more important for efficient pairing, and the resulting deletion of a large region of DNA, than the 13-bp repeats themselves. Based on their results, the authors suggest that in the sample studied, common deletions can occur even without having two intact copies of the 13-bp repeat, because the pairing events could be mediated by the surrounding sequences, which however contain no such perfect repeats. Such a scenario, in which 13-bp repeats are also robustly 'backed up' by strong pairing tendencies of the surrounding DNA, would seem particularly incompatible with the idea that the propensity for pairing of different regions of the mtDNA is just some erroneous danger left uncorrected or unheeded by evolution, without any advantage. The situation would be more compatible with a tradeoff involving a possible direct benefit of maintaining structural mtDNA variety or heteroplasmy in an individual or population. More strongly: we might ask if an occasional need for the pairing event itself might be a major, or conceivably even the main, pressure maintaining the propensity for the common deletion in many human individuals.

More generally, we might ask if a possible benefit of furthering structural mtDNA variants, one that has not received much attention in the literature, could be to facilitate fast adaptation or fine-tuning of respiration in the cell, and respiration rates in the mitochondria, by differential generation/elimination of structural variants.

4.4. The 'neglected' mitochondrial genome: an overdue focus

The small mtDNA genomes or chromosomes of the mitochondria of fungi, which are usually much smaller than 100 kb, belong essentially to a genetically closed system, physically separate from the nucleus and the nuclear genome, with DNA being transferred only rarely or transiently between the two types of genome. This fact, together with the proneness of intra-chromosomal repeats to pair if they are separated by typical short distances within a fungal mtDNA chromosome (kb to tens of kb), suggest that significant repetition of sequences within the mtDNA could often pose a threat of genome instabilities (e.g., large block deletions); for this reason, fungal mtDNA sequences should be an attractive option for further exploring some of the hypotheses we have considered here and extending or consolidating our results.

At least in the case of fungal sequencing, however, the mitochondrial genome has gone through a phase of being 'the neglected genome' (Picardi and Pesole, 2012; Pesole et al., 2012; Rand, 2001). Many complete nuclear genome sequences have only incomplete or missing mitochondrial genomes to accompany them. Obtaining full mitochondrial genome assemblies of fungi from short reads or 454 sequencing can be particularly difficult. Sometimes the reasons are fundamental (repetitiveness of the genome *vs.* short reads), i.e., the assembly limits cannot be avoided by using a different

assembly program (Muñoz et al., 2014). Thus, Lynch et al. (2008) reported specific mtDNA assembly problems for Scer isolates from 454 reads, and Dimitrov et al. (2009) and Steinmetz et al. (2002) addressed variation in mitochondrial properties among Scer isolates, but in the end they analyzed only nuclear DNA sequences to explain the mitochondrial variation.

The mitochondrial genome has, in addition, well-known fundamental differences compared to the nuclear genome: different genetic codes, different gene transcription and regulation paradigms, haploidy, heteroplasmy, absence of nuclear chromatin processes, and different genetics and evolution due to e.g. maternal inheritance, no crossing over, transitions more common than transversions, and phylogeny correlating with land geography. Programs optimized for nuclear genomes do not always have a 'mitochondrial' switch, presumably because program authors do not anticipate frequent use for mtDNA; incomplete mtDNA gene annotations in UCSC human genome releases (e.g., of hg19) may be partly due to such problems.

The period of neglected mitochondrial genomics may soon be over. Simple, bulk screening strategies such as those described here should then aid exploratory analyses of new, previously uncharacterized fungal mtDNA sequences, and may give us deeper insight into the roles of repeat landscapes across fungal mitochondrial genomes.

5. Conclusions

1 The degree of sequence repetition within mitochondrial genomes of fungi spans a wide range, from only trivial or no repeats to rich, complex repeat systems with clear propensity to form secondary structures. A good example of a fungus with such rich, complex repeat systems is baker's yeast (Scer).
2 Ectopic intrachromosomal recombining or pairing entails risks of segmental deletions and genome instability, but these risks may be compensated by a (possibly unknown) benefit of a particular repeat, repeat system, or repeat-pairing product. Persistence can be indicative of functional importance.
3 We have illustrated the potential of the above principle, for mitochondrial sequence mining, by screening 11 complete fungal mitochondrial genomes for repeated words of at least 17 bp occurring at least 4 times, and then using those occurrences to identify elements of repeat landscapes and possible structure/shape vocabularies in the genomes. Many such elements were part of genome features that had previously attracted interest, but some were novel.
4 In the 11 complete mitochondrial genomes we analyzed, we used a simple 'layer-peeling' strategy for exhaustively analyzing and interpreting the set of persistently repetitive 17-mers; this proved to be effective, and a similar strategy should also be useful for other w-mer sizes close to 17. Briefly, the strategy is to successively remove first the w-mer repeats that consist of only A and T (up to perhaps one instance of another base), and then those w-mer repeats that decompose into much smaller subrepeats. For the w-mers that remain, we apply methods such as: screening for intrinsically likely secondary structures by calculating ΔG of the the w-mers themselves and looking for strongly negative values; local assembly to larger repeated sequences; and recognition of highly repetitive core sequences within the w-mers. For the fungi we studied this was a tractable protocol.
5 Some complexity measures, when applied to DNA sequences, will assign low complexities when repeats are frequent. From a semantic or a task viewpoint, however, the same repeat systems and/or their associated structures would increase structural or functional complexity. This reasoning suggests the need for a new notion, *repeat-associated complexity*.

Acknowledgements

We thank Wentian Li for sharing key method strategies prior to publication, for pointing out the parallels with CRISPRs in prokaryotes, and for helpful discussions and references. OKC would like to thank Regina Gourset for recent stimulating discussions and for sharing unpublished results on mutant mtDNA structures, Giorgio Bernardi, Regine Gourset and Edda Rayko for introducing him to this field, and Edward Trifonov for reading the manuscript and for helpful comments. This work was partly supported by Colciencias grants 1222-56934875, "A gene atlas for human pathogenic fungi" and 2213-65842971, "A comprehensive genomic and transcriptomic analysis of dimorphic human pathogen fungi and their relation with virulence," and by Sostenibilidad grants (2013/2014, 2014/2015) from the University of Antioquia. Colciencias National Doctorate Program funding supported JFM and EM and the University of El Rosario partly supported JEG.

Appendix A. Supplementary Data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.compbiolchem.2016.02.016.

References

Allen, J.F., 2003. The function of genomes in bioenergetic organelles. Philos. Trans. R. Soc. Lond. B 358, 19–38.
Allen, J.F., de Paula, W.B.M., 2013. Mitochondrial genome function and maternal inheritance. Biochem. Soc. Trans. 41, 1298–1304.
Baisnee, P.F., Hampson, S., Baldi, P., 2002. Why are complementary DNA strands symmetric? Bioinformatics 18, 1021–1033.
Baker, M.D., Read, L.R., Beatty, B.G., Ng, P., 1996. Requirements for ectopic homologous recombination in mammalian somatic cells. Mol. Cell. Biol. 16, 7122–7132.
Bandelt, H.J., Kloss-Brandstatter, A., Richards, M.B., Yao, Y.G., Logan, I., 2014. The case for the continuing use of the revised Cambridge Reference Sequence (rCRS) and the standardization of notation in human mitochondrial DNA studies. J. Hum. Genet. 59, 66–77.
Benzer, S., 1959. On the topology of the genetic fine structure. Proc. Natl. Acad. Sci. U. S. A. 45, 1607–1620.
Benzer, S., 1962. The fine structure of the gene. Sci. Am. 206, 70–84.
Berge, C., 1983. Graphes, 3rd ed. Gauthiers-Villars, Bordas, Paris.
Berge, C., 1987. Hypergraphes: Combinatoires des Ensembles Finis. Gauthiers-Villars, Bordas, Paris.
Bernardi, G., 2004. Structural and Evolutionary Genomics: Natural Selection in Genome Evolution. Elsevier, Amsterdam.
Bernardi, G., 2005. Lessons from a small, dispensable genome: the mitochondrial genome of yeast. Gene 354, 189–200.
Bollag, R.J., Waldman, A.S., Liskay, R.M., 1989. Homologous recombination in mammalian cells. Annu. Rev. Genet. 23, 199–225.
Britten, R.J., Kohne, D.E., 1968. Repeated sequences in DNA. Science 161, 529–540.
Bultrini, E., Pizzi, E., Del Giudice, P., Frontali, C., 2003. Pentamer vocabularies characterizing introns and intron-like intergenic tracts from *Caenorhabditis elegans* and *Drosophila melanogaster*. Gene 304, 183–192.
Chen, T., He, H., Huang, Y., Zhao, W., 2011. The generation of mitochondrial DNA large-scale deletions in human cells. J. Hum. Genet. 56, 689–694.
Compeau, P.E.C., Pevzner, P.A., Tesler, G., 2011. How to apply de Bruijn graphs to genome assembly. Nat. Biotechnol. 29, 987–991.
Davidson, E.H., 1976. Gene Activity in Early Development, 2nd ed. Academic Press, New York.
de Zamaroczy, M., Marotta, R., Faugeron-Fonty, G., Gourset, R., Mangin, M., Baldacci, G., Bernardi, G., 1981. The origins of replication of the yeast mitochondrial genome and the phenomenon of suppressivity. Nature 292, 75–78.
Dimitrov, L., Brem, R., Kruglyak, L., Gottschling, D., 2009. Polymorphisms in multiple genes contribute to the spontaneous mitochondrial genome instability of *Saccharomyces cerevisiae* S288C strains. Genetics 183, 365–383.
Forsdyke, D.R., 1995. Conservation of stem-loop potential in introns of snake venom phospholipase A_2 genes: an application of FORS-D analysis. Mol. Biol. Evol 12, 1157–1165.
Foury, F., Roganti, T., Lecrenier, N., Purnelle, B., 1998. The complete sequence of the mitochondrial genome of *Saccharomyces cerevisiae*. FEBS Lett. 440, 325–331.
Gallo, J.E., Muñoz, J.F., Misas, E., McEwen, J.G., Clay, O.K., 2014. The complex task of choosing a *de novo* assembly. Lessons from fungal genomes. Comp. Biol. Chem. 53, 97–107.
Gerhold, J.M., Aun, A., Sedman, T., Joers, P., Sedman, J., 2010. Strand invasion structures in the inverted repeat of *Candida albicans* mitochondrial DNA reveal a role for homologous recombination in replication. Mol. Cell 39, 851–861.

Golumbic, M.C., 2004. Algorithmic Graph Theory and Perfect Graphs, 2nd ed. Elsevier, Amsterdam.

Guo, X., Popadin, K.Y., Markuzon, N., Orlov, Y.L., Kraytsberg, Y., et al., 2010. Repeats, longevity and the sources of mtDNA deletions: evidence from 'deletional spectra'. Trends Genet. 26, 340–343.

Joardar, V., Abrams, N.F., Hostetler, J., Paukstelis, P.J., Pakala, S., et al., 2012. Sequencing of mitochondrial genomes of nine *Aspergillus* and *Penicillium* species identifies mobile introns and accessory genes as main sources of genome size variability. BMC Genom. 13, 698.

Jorgensen, J.H., Pfaller, M.A., Carroll, K.C., Funke, G., Landry, M.L., Richter, S.S., Warnock, D.W., 2015. Manual of Clinical Microbiology, 11th ed. ASM Press, Washington, DC.

Kann, L.M., Rosenblum, E.B., Rand, D.M., 1998. Aging, mating, and the evolution of mtDNA heteroplasmy in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. U. S. A. 95, 2372–2377.

Khinchin, A.I., 1957. Mathematical Foundations of Information Theory. Dover, Mineola, NY.

King, D., Trifonov, E., Kashi, Y., 2006. Tuning knobs in the genome: evolution of simple sequence repeats by indirect selection. In: Caporale, L. (Ed.), The Implicit Genome. Oxford University Press, New York, pp. 77–90.

Kubo, T., Newton, K.J., 2008. Angiosperm mitochondrial genomes and mutations. Mitochondrion 8, 5–14.

Lakshmanan, L.N., Gruber, J., Halliwell, B., Gunawan, R., 2012. Role of direct repeat and stem-loop motifs in mtDNA deletions: cause or coincidence? PLoS ONE 7 (4), e35271.

Lane, N., 2006. Power, Sex, Suicide: Mitochondria and the Meaning of Life. Oxford University Press.

Lewis, D.L., Farr, C.L., Farquhar, A.L., Kaguni, L.S., 1994. Sequence, organization, and evolution of the A+T region of *Drosophila melanogaster* mitochondrial DNA. Mol. Biol. Evol. 11, 523–538.

Li, W., Freudenberg, J., 2014a. Characterizing regions in the human genome unmappable by next-generation-sequencing at reads length of 1000 bases. Comput. Biol. Chem. 53 (Pt A), 108–117.

Li, W., Freudenberg, J., 2014b. Mappability and read length. Front. Genet 5, 381.

Li, W., Freudenberg, J., Miramontes, P., 2014. Diminishing return for increased mappability with longer sequencing reads: implications of the *k*-mer distributions in the human genome. BMC Bioinform. 15, 2.

Lichten, M., Haber, J.E., 1989. Position effects in ectopic and allelic mitotic recombination in *Saccharomyces cerevisiae*. Genetics 123, 261–268.

Lynch, M., Sung, W., Morris, K., Coffey, N., Landry, C., et al., 2008. A genome-wide view of the spectrum of spontaneous mutations in yeast. Proc. Natl. Acad. Sci. U. S. A. 105, 9272–9277.

Maresca, B., Lambowitz, A.M., Kumar, V.B., Grant, G.A., Kobayashi, G.S., Medoff, G., 1981. Role of cysteine in regulating morphogenesis and mitochondrial activity in the dimorphic fungus *Histoplasma capsulatum*. Proc. Natl. Acad. Sci. U. S. A. 78, 4596–4600.

Marotta, R., Colin, Y., Goursot, R., Bernardi, G., 1982. A region of extreme instability in the mitochondrial genome of yeast. EMBO J. 1, 529–534.

Mezard, C., Pompon, D., Nicolas, A., 1992. Recombination between similar but not identical DNA sequences during yeast transformation occurs within short stretches of identity. Cell 70, 659–670.

Muñoz, J.F., Misas, E., Gallo, J.E., McEwen, J.G., Clay, O.K., 2014. Limits to sequencing and *de novo* assembly: classic benchmark sequences for optimizing fungal NGS design. Adv. Intell. Syst. Comput. 232, 221–230.

Nargang, F.E., Bell, J.B., Stohl, L.L., Lambowitz, A.M., 1983. A family of repetitive palindromic sequences found in *Neurospora* mitochondrial DNA is also found in a mitochondrial plasmid DNA. J. Biol. Chem. 258, 4257–4260.

Negruk, V., 2013. Mitochondrial genome sequence of the legume *Vicia faba*. Front. Plant Sci. 4, 128.

Pesole, G., Allen, J.F., Lane, N., Martin, W., Rand, D.M., Schatz, G., Saccone, C., 2012. The neglected genome. EMBO Rep. 13, 473–474.

Pesole, G., Prunella, N., Liuni, S., Attimonelli, M., Saccone, C., 1992. WORDUP: an efficient algorithm for discovering statistically significant patterns in DNA sequences. Nucleic Acids Res. 20, 2871–2875.

Picardi, E., Pesole, G., 2012. Mitochondrial genomes gleaned from human whole-exome sequencing. Nat. Methods 9, 523–524.

Pohjoismäki, J.L.O., Goffart, S., Taylor, R.W., Turnbull, D.M., Suomalainen, A., Jacobs, H.T., Karhunen, P.J., 2010. Developmental and pathological changes in the human cardiac muscle mitochondrial DNA organization, replication and copy number. PLoS ONE 5 (5), e10426.

Price, A.L., Jones, N.C., Pevzner, P.A., 2005. *De novo* identification of repeat families in large genomes. Bioinformatics 21 (Suppl. 1), i351L 358.

Rand, D.M., 2001. The units of selection on mitochondrial DNA. Ann. Rev. Ecol. Syst. 32, 415–448.

Rayko, E., Goursot, R., 1999. Amphimeric mitochondrial genomes of petite mutants of yeast, III. Generation by linking two secondary–structure-dependent illegitimate recombination events. Curr. Genet. 35, 14–22.

Rizk, G., Lavenier, D., Chikhi, R., 2013. DSK: k-mer counting with very low memory usage. Bioinformatics 29, 652–653.

Samuels, D.C., Schon, E.A., Chinnery, P.F., 2004. Two direct repeats cause most human mtDNA deletions. Trends Genet. 20, 393–398.

Sequeira, A., Martin, M.V., Rollins, B., Moon, E.A., Bunney, W.E., Macciardi, F., Lupoli, S., Smith, E.N., Kelsoe, J., Magnan, C.N., van Oven, M., Baldi, P., Wallace, D.C., Vawter, M.P., 2012. Mitochondrial mutations and polymorphisms in psychiatric disorders. Front. Genet. 3, 103.

Steinmetz, L.M., Scharfe, C., Deutschbauer, A.M., Mokranjac, D., Herman, Z.S., et al., 2002. Systematic screen for human disease genes in yeast. Nat. Genet. 31, 400–404.

Sugawara, N., Haber, J.E., 1992. Characterization of double-strand break-induced recombination: homology requirements and single-stranded DNA formation. Mol. Cell. Biol. 12, 563–575.

Trifonov, E., 1999. Tandem repeats as tuning knobs for fast adaptation and differentiation (Abstract). In: Evolutionary Genomics, Punta Leona, Costa Rica, January 10-16.

Trifonov, E.N., 2004. Tuning function of tandemly repeating sequences: a molecular device for fast adaptation. In: Wasser, S. (Ed.), Evolutionary Theory and Processes: Modern Horizons. Papers in Honour of Eviatar Nevo. Springer Science + Business Media/Kluwer, Dordrecht, pp. 115–138.

van de Sande, W.W.J., 2012. Phylogenetic analysis of the complete mitochondrial genome of *Madurella mycetomatis* confirms its taxonomic position within the order Sordariales. PLoS ONE 7, e38654.

Wallace, D.C., 2013. A mitochondrial bioenergetic etiology of disease. J. Clin. Invest. 123, 1405–1412.

Yin, S., Heckman, J., RajBhandary, U.L., 1981. Highly conserved GC-rich palindromic DNA sequences flank tRNA genes in *Neurospora crassa* mitochondria. Cell 26 (3 Pt 1), 325–332.

Zahradnik, D., Trifonov, E.N., Zemková, M., 2014. The evolutionary landscape of human genome vocabulary. Czech Slovak Linguistic Rev. 2014 (1), 106–119.

Zaragoza, M.V., Brandon, M.C., Diegoli, M., Arbustini, E., Wallace, D.C., 2011. Mitochondrial cardiomyopathies: how to identify candidate pathogenic mutations by mitochondrial DNA sequencing, MITOMASTER and phylogeny. Eur. J. Hum. Genet. 19, 200–207.

Zuker, M., 2003. Mfold web server for nucleic acid folding and hybridization prediction. Nucl. Acids Res. 31, 3406–3415, http://mfold.rna.albany.edu/?q=mfold; http://mfold.rna.albany.edu/?q=DINAMelt/Quickfold.

Chapter 8

Genome diversity, recombination, and virulence across the major lineages of *Paracoccidioides*

AMERICAN SOCIETY FOR MICROBIOLOGY | mSphere™

RESEARCH ARTICLE
Ecological and Evolutionary Science

Genome Diversity, Recombination, and Virulence across the Major Lineages of *Paracoccidioides*

José F. Muñoz,[a,b,c] Rhys A. Farrer,[c] Christopher A. Desjardins,[c] Juan E. Gallo,[a,d] Sean Sykes,[c] Sharadha Sakthikumar,[c] Elizabeth Misas,[a,b] Emily A. Whiston,[e] Eduardo Bagagli,[f] Celia M. A. Soares,[g] Marcus de M. Teixeira,[h,i] John W. Taylor,[e] Oliver K. Clay,[a,j] Juan G. McEwen,[a,k] Christina A. Cuomo[c]

Cellular and Molecular Biology Unit, Corporación para Investigaciones Biológicas, Medellín, Colombia[a]; Institute of Biology, Universidad de Antioquia, Medellín, Colombia[b]; Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA[c]; Doctoral Program in Biomedical Sciences, Universidad del Rosario, Bogotá, Colombia[d]; Department of Plant and Microbial Biology, University of California, Berkeley, Berkeley, California, USA[e]; Instituto de Biociências, Universidade Estadual Paulista, Botucatu, São Paulo, Brazil[f]; Laboratório de Biología Molecular, Instituto de Ciências Biológicas, ICBII, Goiânia, Brazil[g]; Instituto de Ciências Biológicas, Universidade de Brasília, Brasília, Distrito Federal, Brazil[h]; Division of Pathogen Genomics, Translational Genomics Research Institute North, Flagstaff, Arizona, USA[i]; School of Medicine and Health Sciences, Universidad del Rosario, Bogotá, Colombia[j]; School of Medicine, Universidad de Antioquia, Medellín, Colombia[k]

ABSTRACT The *Paracoccidioides* genus includes two species of thermally dimorphic fungi that cause paracoccidioidomycosis, a neglected health-threatening human systemic mycosis endemic to Latin America. To examine the genome evolution and the diversity of *Paracoccidioides* spp., we conducted whole-genome sequencing of 31 isolates representing the phylogenetic, geographic, and ecological breadth of the genus. These samples included clinical, environmental and laboratory reference strains of the S1, PS2, PS3, and PS4 lineages of *P. brasiliensis* and also isolates of *Paracoccidioides lutzii* species. We completed the first annotated genome assemblies for the PS3 and PS4 lineages and found that gene order was highly conserved across the major lineages, with only a few chromosomal rearrangements. Comparing whole-genome assemblies of the major lineages with single-nucleotide polymorphisms (SNPs) predicted from the remaining 26 isolates, we identified a deep split of the S1 lineage into two clades we named S1a and S1b. We found evidence for greater genetic exchange between the S1b lineage and all other lineages; this may reflect the broad geographic range of S1b, which is often sympatric with the remaining, largely geographically isolated lineages. In addition, we found evidence of positive selection for the *GP43* and *PGA1* antigen genes and genes coding for other secreted proteins and proteases and lineage-specific loss-of-function mutations in cell wall and protease genes; these together may contribute to virulence and host immune response variation among natural isolates of *Paracoccidioides* spp. These insights into the recent evolutionary events highlight important differences between the lineages that could impact the distribution, pathogenicity, and ecology of *Paracoccidioides*.

IMPORTANCE Characterization of genetic differences between lineages of the dimorphic human-pathogenic fungus *Paracoccidioides* can identify changes linked to important phenotypes and guide the development of new diagnostics and treatments. In this article, we compared genomes of 31 diverse isolates representing the major lineages of *Paracoccidioides* spp. and completed the first annotated genome sequences for the PS3 and PS4 lineages. We analyzed the population structure and characterized the genetic diversity among the lineages of *Paracoccidioides*, including a deep split of S1 into two lineages (S1a and S1b), and differentiated S1b, associated

Received 28 July 2016 **Accepted** 6 September 2016 **Published** 28 September 2016

Citation Muñoz JF, Farrer RA, Desjardins CA, Gallo JE, Sykes S, Sakthikumar S, Misas E, Whiston EA, Bagagli E, Soares CMA, Teixeira MDM, Taylor JW, Clay OK, McEwen JG, Cuomo CA. 2016. Genome diversity, recombination, and virulence across the major lineages of *Paracoccidioides*. mSphere 1(5):e00213-16. doi:10.1128/mSphere.00213-16.

Editor Aaron P. Mitchell, Carnegie Mellon University

Address correspondence to Christina A. Cuomo, cuomo@broadinstitute.org.

with most clinical cases, as the more highly recombining and diverse lineage. In addition, we found patterns of positive selection in surface proteins and secreted enzymes among the lineages, suggesting diversifying mechanisms of pathogenicity and adaptation across this species complex. These genetic differences suggest associations with the geographic range, pathogenicity, and ecological niches of *Paracoccidioides* lineages.

KEYWORDS: *Paracoccidioides*, evolution, genetic recombination, genome analysis, mycology, population genetics

Paracoccidioides spp. are the cause of paracoccidioidomycosis (PCM), a systemic mycosis that mainly affects people in Latin America. In this region where PCM is endemic, PCM has an estimated incidence of 1 to 3 cases per 100,000 inhabitants (1, 2). The vast majority of PCM cases (roughly 80%) occur in Brazil, while Colombia and Venezuela have the next highest numbers of infections (3). *Paracoccidioides* is a thermally dimorphic fungus closely related to *Histoplasma* and *Blastomyces*, which cause similar infections worldwide or predominantly in regions of North America, respectively.

Multilocus sequencing studies elucidated species boundaries within the *Paracoccidioides* genus and supported the existence of two distinct species, *P. brasiliensis* and *P. lutzii* (4). *P. lutzii* is a single monophyletic and recombining population found to date in central, southwest, and north Brazil and Ecuador (4). *P. brasiliensis* is monophyletic and is comprised of distinct lineages classified as S1, PS2, PS3, and PS4 (4–6). The S1 lineage is associated with the majority of PCM cases and is widely distributed in South America (4–6). PS2 has been identified to date only in Brazil and Venezuela, whereas PS3 is mainly found in regions of endemicity in Colombia (4, 5). Recently, a novel lineage, PS4, was described from a region of Venezuela (6). Evidence of recombination was noted for *P. brasiliensis* S1 and *P. lutzii*, but not other lineages, based on a small number of genomic loci (4, 5).

Isolates from each of these phylogenetic lineages of *Paracoccidioides* can infect humans; however, different lineages can vary in virulence and culture adaptation and can induce different immune responses by the host (7, 8). One feature that is correlated with the differential rates of infection is variation in the number of infective conidia. For example, isolates from S1 produce many more conidia than PS2 isolates, which could be related to the disproportional 9:1 rate of S1 to PS2 infection in both human and armadillo isolates (8). In addition to interspecific variation between lineages and between species, *Paracoccidioides* isolates have been shown to contain extensive intraspecific genetic variability between strains of the same lineage (9–11).

To enable genome-based studies of this medically important fungus, isolates of *P. brasiliensis* S1 and PS2 and *P. lutzii* were previously sequenced and compared to related dimorphic and nondimorphic fungi (12). Notably, *Paracoccidioides* and related dimorphic pathogens have a reduced number of genes involved in carbohydrate metabolism, protein metabolism, and synthesis of secondary metabolites (12), an observation that allows new insights into the differences between these related fungi and their physiological potential for pathogenicity. Recently, the genome assemblies and gene annotations of those reference strains were improved using Illumina resequencing, increasing the overall accuracy of assembly bases and gene structures (13). These improved reference genomes of *Paracoccidioides* spp. provide an opportunity to map the population structure and examine variation with finer resolution.

In this study, we used genome sequences of 31 isolates for a comprehensive comparison of gene conservation, genetic diversity, and genome evolution across the major lineages of *Paracoccidioides*. The panel of isolates sequenced in this study included clinical isolates from acute and chronic PCM cases and environmental isolates from soil or two species of armadillos (*Dasypus novemcinctus* and *Cabassous centralis*). We assembled the first reference genomes for the PS3 and PS4 lineages: compared to the previously assembled references from other lineages, gene content and order are

highly conserved, with few rearrangements. We characterized genetic diversity among the lineages, as well as lineage-specific evolutionary patterns within the *Paracoccidioides* genus and found evidence of recombination and ancestral hybridization patterns between some of the lineages. Additionally, we identified genomic regions or genes that are highly diverse within or between lineages; these include genes with potential roles in virulence. We found that genes with the strongest evidence of positive selection include the *GP43* antigen gene and genes coding for other secreted proteins and proteases (e.g., *PGA1*, *CBP1*, *SOD3*, and *ENG1*), as well as loss-of-function mutation in genes that are specific to some lineages. Our analyses provide insight into the recent evolutionary events highlighting genetic differences between the lineages that could impact the distribution, pathogenicity, and ecology of *Paracoccidioides*. These potential virulence factors and genetic differences at the population level will be important for future studies to compare with the infectious potency of *Paracoccidioides* clinical isolates and clinical symptoms attributable to paracoccidioidomycosis.

RESULTS

Conserved genome organization across *Paracoccidioides* spp. To compare genome structures and gene contents between the major *Paracoccidioides* lineages, we sequenced, assembled, and annotated a representative isolate from the PS3 lineage (strain PbCnh) and from the PS4 lineages (strain Pb300) (see Fig. S1A and Text S1 in the supplemental material). The 29.4-Mb assemblies of both strains are intermediate in size between those of the 29.95-Mb assembly of Pb18 (S1b) and the 29.06-Mb assembly of Pb03 (PS2) (see Fig. S1B) (13). The gene annotation of strains PbCnh and Pb300 resulted in 8,324 and 8,070 predicted protein-coding genes, respectively. High representation of core eukaryotic genes provides evidence that those genomes are nearly complete; 96 to 98% of these conserved genes are found in all assemblies (see Fig. S1C). Predicted gene contents were highly similar across all four *P. brasiliensis* genomes, comparing the new assemblies to the previously sequenced genomes (see Fig. S1B), which suggests that the gene content is very consistent across the lineages.

The *Paracoccidioides* genomes of both species and all lineages are highly conserved in terms of whole-genome sequence similarity and gene synteny (see Text S1 and Fig. S1D). The genomes of *P. brasiliensis* share an average of 98.5% identity, whereas the genome of the more distant species *P. lutzii* shares an average of 94.8% with *P. brasiliensis*. The genomes of PbCnh (PS3) and Pb300 (PS4) share the highest percentage aligned (98.9%), which correlates with the phylogenetic and population structure relationships (see below). We identified syntenic regions of conserved gene order and found an average of 6,907 genes within syntenic blocks among the *Paracoccidioides* lineages (see Fig. S1D to F). The percentage of genes in syntenic blocks ranged from 75.3% (interspecies, *P. brasiliensis* versus *P. lutzii*) to 89.9% (intraspecies, *P. brasiliensis* versus *P. brasiliensis*). In contrast, the dimorphic fungus *Blastomyces* has only ~69% genes in syntenic blocks due to the presence of isochore-like structures of repeat-rich GC-poor and GC-rich blocks rarely observed in *Paracoccidioides* (14).

While the *Paracoccidioides* genomes were largely colinear, a few chromosomal rearrangements were detected. A large rearrangement was detected between *P. brasiliensis* S1/PS3 and *P. lutzii*/*P. brasiliensis* PS2 in chromosome 4, where the regions at the beginning and the end of the supercontig 4 are inverted, and the gene order across the middle of the supercontig is conserved (see Fig. S1F). A large chromosomal rearrangement was also detected between *P. brasiliensis* and *P. lutzii*, where syntenic blocks in *P. brasiliensis* chromosomes 3 and 5 (Pb18, supercontigs 2, 13, and 10) are combined into a single supercontig in *P. lutzii* (see Fig. S1F). We found no evidence of assembly errors across the junctions of these rearrangements based on even coverage of aligned reads across these regions. In addition, we called structural variants based on the read alignments to the Pb18 assembly, and recovered each of the rearrangements present in the assemblies (see Text S1 and Data Set S1 in the supplemental material). Chromosomal rearrangements may impact the capacity for genetic exchange, as some crossover events will generate missing chromosomal regions or other aneu-

TABLE 1 *Paracoccidioides* species isolates selected for this study

Isolate ID	Other name(s)	Origin	Source	Lineage	Provider	Reference
P. brasiliensis						
Pb18[a]	B17	Sao Paulo, Brazil	Chronic PCM[b]	S1b	R. Puccia	44
PbCaz	Cazon; A1	Chaco, Argentina	Acute PCM	S1b	R. Negroni	45
Pb113		Manaus-AM, Brazil	PCM	S1b	C. de Almeida Soares	46
PbBlo		Brazil	PCM	S1b	C. de Almeida Soares	This study
MS1		Mato Grosso do Sul, Brazil	PCM	S1a	M. Sueli Felipe	47
D03		Piracicaba, SP, Brazil	PCM	S1a	E. Bagagli	This study
MS2		Mato Grosso do Sul, Brazil	PCM	S1a	M. Sueli Felipe	47
Pb1445	A5	Argentina	Chronic PCM	S1a	R. Negroni	5
Pb337	T15LN1; B10	Brazil	*D. novemcinctus*	S1a	E. Bagagli	48
Pb66		Brazil	PCM	S1a	C. de Almeida Soares	49
PbBer	Bercelli; A3	Argentina	PCM	S1a	R. Negroni	5
D02		Laranjal Paulista, SP, Brazil	PCM	S1a	E. Bagagli	This study
T1F1	B1	Pratanea, SP, Brazil	*D. novemcinctus*	S1a	E. Bagagli	48
T15N1		Botucatu, Brazil	*D. novemcinctus*	S1a	E. Bagagli	This study
T16B1		Brazil	*D. novemcinctus*	S1a	E. Bagagli	This study
Pb300[a]	V1	Miranda, Venezuela	Soil	PS4	M. B. Albornoz	50
EPM83		Bogotá, Colombia	Chronic PCM	PS3	A. Restrepo	49
Pb339	B18	Sao Paulo, Brazil	PCM	PS3	A. Restrepo	51
Pb60855	C4	Antioquia, Colombia	Chronic PCM	PS3	A. Restrepo	52
PbBac		Colombia	PCM	PS3	A. Restrepo	This study
PbCab	P196; C6	Caldas, Colombia	*C. centralis*	PS3	A. Restrepo	53
PbCnh[a]		Colombia	Chronic PCM	PS3	A. Restrepo	This study
PbJam		Colombia	Chronic PCM	PS3	A. Restrepo	This study
Pb02	V2	Caracas, Venezuela	Chronic PCM	PS2	R. Puccia	54
Pb03[a]	B26	Sao Paulo, Brazil	Chronic PCM	PS2	R. Puccia	54
Pb262		Uberlândia, MG, Brazil	Dog food	PS2	Z. Pires de Camargo	55
T10B1	B7	Botucatu, Brazil	*D. novemcinctus*	PS2	E. Bagagli	5
P. lutzii						
Pb01[a]		Goiás, Brazil	PCM	*P. lutzii*	R. Puccia	56
Pl1578		Goiás, Brazil	PCM	*P. lutzii*	C. de Almeida Soares	47
ED01		Goiás, Brazil	PCM	*P. lutzii*	C. de Almeida Soares	47
PlEE	EE	Mato Grosso, Brazil	PCM	*P. lutzii*	M. Sueli Felipe	4

[a]Reference strain assembled and annotated genome.
[b]PCM, paracoccidioidomycosis.

ploidies and nonviable progeny. The rearrangement between PS2 and the other lineages of *P. brasiliensis* could potentially prevent genetic exchange between these groups.

In addition to chromosomal rearrangements, we looked for evidence of copy number and ploidy variation. We calculated the normalized read alignment density in 10-kb nonoverlapping windows (see Materials and Methods) and examined the variation across the Pb18 chromosomes. The alignment density showed no large regions of higher sequencing depth, supporting that *Paracoccidioides* species do not maintain aneuploid chromosomes or segments (see Fig. S2 in the supplemental material) such as those found in other fungi.

Phylogenetic for S1 group split and allele sharing between lineages. To examine the *Paracoccidioides* phylogenetic relationships, we identified polymorphisms across an extended panel of 31 isolates (Table 1). Using 614,570 positions (single-nucleotide polymorphisms [SNPs]) in the sequenced isolates, maximum likelihood and Bayesian phylogenies were constructed to examine intralineage relationships (Materials and Methods). Both maximum likelihood and Bayesian analyses highly supported that *P. brasiliensis* isolates are clustered into five distinct lineages (Fig. 1; see Fig. S3 in the supplemental material). Of the four previously identified lineages, S1 is the most highly variable, with two distinct clades we denoted as the S1a and S1b lineages. The S1b lineage includes the reference strain Pb18, along with three clinical isolates from Brazil and Argentina. The S1a lineage includes clinical and environmental isolates and was split into two subclades: one includes clinical isolates from Argentina and central-west regions of Brazil, while the second includes three clinical and four armadillo isolates

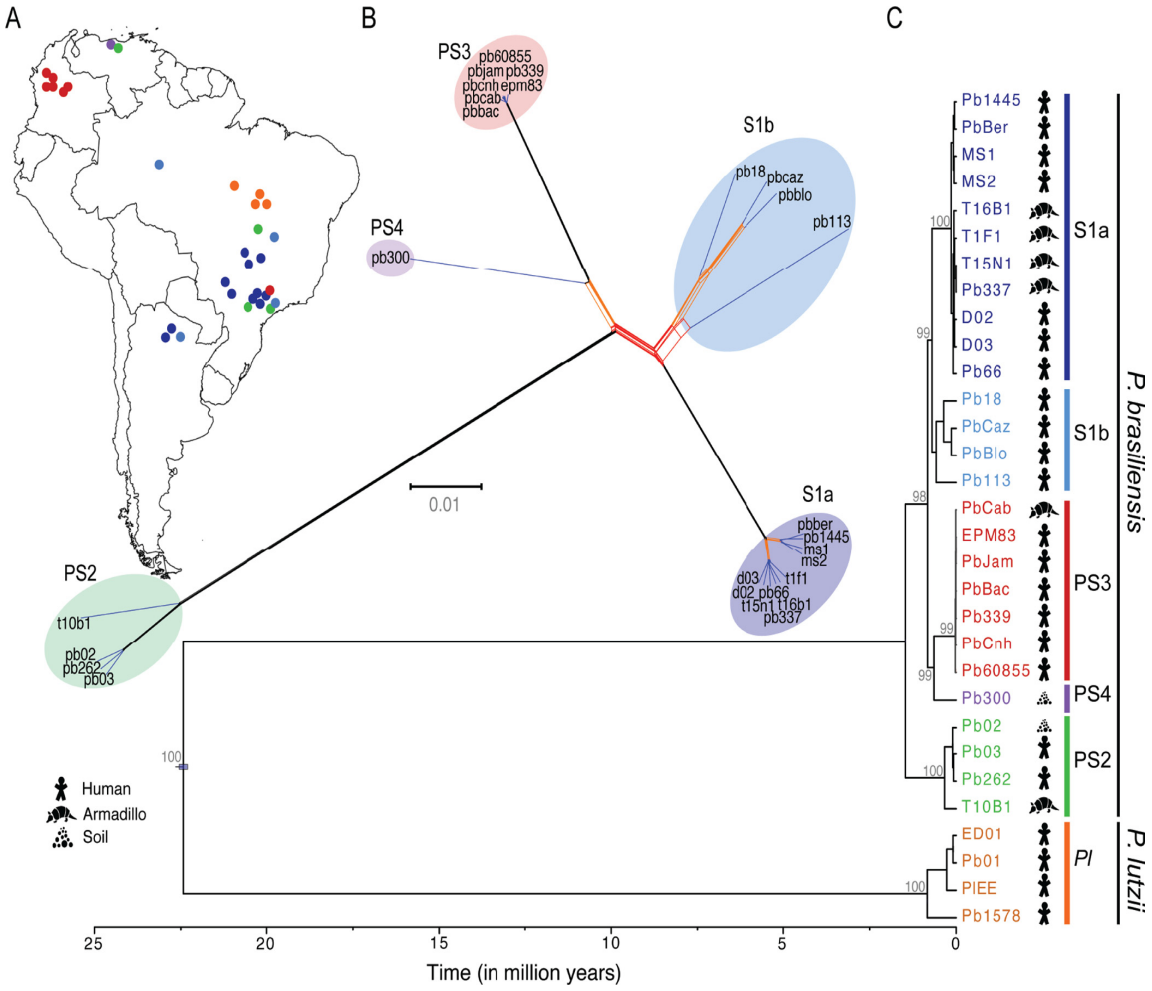


FIG 1 Phylogeny and recombination in *Paracoccidioides*. Two methods were used to examine strain relationships originating from across South America (A): using 614,570 SNPs, including a phylogenetic network constructed with SplitsTree4 (B), and a Bayesian calibrated phylogeny constructed with BEAST (C); bootstrap values from maximum likelihood phylogeny constructed with RAxML were included for major subdivisions. Both methods show evidence of five distinct lineages in *P. brasiliensis*: S1 (blue), which is divided into two groups S1a (dark blue) and S1b (light blue), PS2 (green), PS3 (red), and the recently described PS4 (purple). Also, this phylogeny supports the divergence between *P. brasiliensis* and *P. lutzii* (*PI* [orange]) as a different species. In addition, the phylogenetic network of *P. brasiliensis* suggests patterns of recombination (red branches).

from southeast Brazil. This phylogenetic analysis also supports that PS3 is a monophyletic group with very limited diversity of mostly isolates from Colombia, including both chronic PCM isolates and one isolate from armadillo. While previous phylogenies using short loci had suggested that PS3 only includes isolates from Colombia (4, 5, 8), we found that the Pb339 isolate from southeast Brazil was placed in this lineage (Fig. 1), suggesting that this lineage may be more widespread than previously described. In addition, our phylogenetic analysis provides strong evidence for the separation of the PS4 lineage, as recently proposed (6), sharing a common ancestor with PS3. This was also supported by population structure analyses (see below). To confirm that the phylogeny was not influenced by the use of a reference genome from one lineage, we identified SNPs from reads aligned to reference genomes representing each lineage (Pb03, PbCnh, Pb300, and Pb01) and found the same topology in phylogenies of each set (see Fig. S3).

To better understand the evolutionary history of *Paracoccidioides*, we estimated the divergence and most common ancestor dates (time to most recent common ancestor [TMRCA]). This analysis suggests that *P. lutzii* and *P. brasiliensis* diverged about 22.5 million years ago (MYA). The *P. brasiliensis* lineages separated more recently at

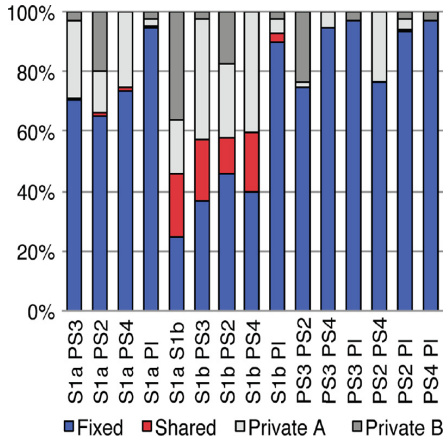


FIG 2 Classification of population-specific SNP alleles. The percentage of the distribution of SNP alleles is shown for each pairwise comparison among the six lineages of *Paracoccidioides* spp. SNP alleles were classified as fixed (blue), shared (red), private A (light gray [for each pairwise comparison specific to the first lineage in the *x* axis]), and private B (dark gray [specific to the second lineage]).

1.47 MYA (Fig. 1). Within *P. brasiliensis* lineages, the shortest time of divergence was found for PS3 (TMRCA, 23,000 years ago [23 KYA]) and the longest was found in S1b (TMRCA, 575 KYA), with PS2 and S1a showing intermediate values (Fig. 1; see Fig. S4 in the supplemental material). We estimated that *P. lutzii* diverged slightly earlier than the *P. brasiliensis* groups, around 833 KYA. Using whole-genome SNPs, the separation of *P. brasiliensis* and *P. lutzii* is very similar to a previously reported estimate (8); however, we estimate more recent separation within each lineage.

While the SNP data strongly supported a single tree, we also examined the relationship of the sequenced isolates using a network approach to look for evidence of alternative topologies. The NeighborNet algorithm identified five major groups within the *P. brasiliensis* species, including the clear separation of S1 into the S1a and S1b lineages. In addition, the network suggests some level of ancestral recombination between the groups as well as more recent recombination involving the S1b lineage (Fig. 1B). We tested for evidence of phylogenetic heterogeneity using the pairwise homoplasy index test (15) and found statistically significant support for recombination ($P = 4.3e{-}13$).

Next, we compared the distribution of SNPs across the genome and examined how sites were shared between lineages. We found a similar pattern of polymorphism frequency across the genome for isolates of the same lineage; differences between lineages include two distinct patterns for the related phylogenetic lineages S1a and S1b, supporting this subdivision (see Fig. S2 in the supplemental material). To more finely compare variant sites between lineages, we classified SNP alleles based on populations as fixed, shared, or private based on pairwise comparisons between lineages (S1a, S1b, PS2, PS3, PS4, and *P. lutzii*). S1b has the highest shared SNP allele component, ranging from 12.4% to 21.1% compared with the S1a, PS2, PS3 and PS4 lineages and 2.5% compared with *P. lutzii* (Fig. 2). In contrast, the next highest shared frequency is for S1a, which ranged from 0.17% to 1.1% compared with PS2, PS3, PS4, and *P. lutzii*. To determine whether the large fraction of shared alleles was due to recombination rather than just a higher genetic diversity in S1b, we combined SNPs for PS3 and PS4 to make an artificial lineage with diversity relatively equivalent to that of S1b. In this combined set, we did not observe the fraction of shared alleles increase from the values of PS3 and PS4 alone, suggesting the effect in S1b is not due to high diversity alone. While recombination with other lineages is also supported by the network tree analysis, it is unclear whether this recombination was relatively recent or ancestral.

Recombination and hybridization between the major lineages of *Paracoccidioides*. To further examine the *Paracoccidioides* lineages for evidence of recombi-

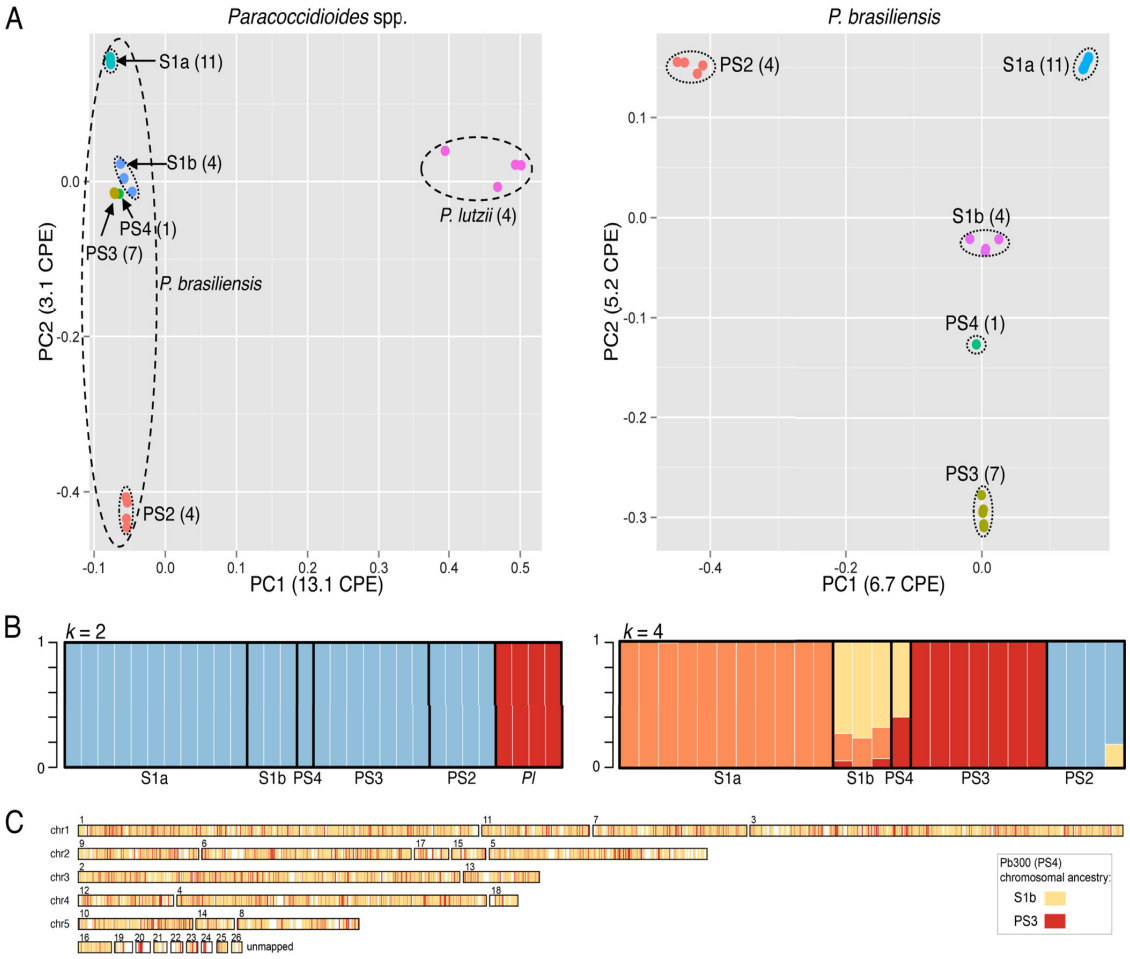


FIG 3 Genetic population structure of *Paracoccidioides*. (A) PCA of genetic variants in *Paracoccidioides* is shown as a two-dimensional plot for all isolates (n = 31) (left) or for all *P. brasiliensis* isolates (n = 26) (right). In each plot, circles indicate isolates and colors indicate the lineage. (B) Population structure of *Paracoccidioides* spp. (left) and *P. brasiliensis* (right) inferred from 476,589 and 339,966 SNPs, respectively, using the Structure software program with different k values of 2 and 4, respectively. An admixture model with correlated allele frequencies and site-by-site analysis was used. Each isolate is represented by a single vertical line broken into k-colored segments, with lengths proportional to each of the k inferred clusters. (C) Whole-genome plot for Pb300 (PS4) deducted from Structure site-by-site analysis showing the chromosomal ancestry.

nation, we performed a principal-component analysis (PCA) of the SNP data. For all *Paracoccidioides* isolates (n = 31), we found clear separation of the two species; PC1 cleanly separates *P. brasiliensis* and *P. lutzii* isolates as two distinct species. Within *P. brasiliensis*, PC2 separates PS2 from all other groups and divides S1 into two distinct groups (S1a and S1b lineages) (Fig. 3A). In this comparison, the S1b, PS3, and PS4 lineages appeared more closely related, suggesting less genetic divergence between these lineages. Comparing only the *P. brasiliensis* isolates (n = 26), PC2 separates the S1b, PS4, and PS3 lineages, where PS4 is located between S1b and PS3 (Fig. 3A). This analysis supports the major subdivisions in the *P. brasiliensis* population and suggests that some lineages have similar relationships to multiple other lineages, appearing more centrally in PCA plots, suggesting the impact of recombination.

To examine the relationships of these groups, we predicted the population ancestry for each strain using a model-based clustering algorithm implemented by the Structure software (16). We identified populations within *Paracoccidioides* spp. using (i) all isolates and (ii) only *P. brasiliensis* isolates. There is clear separation between *P. brasiliensis* and *P. lutzii*, with only two major clusters of ancestry (Fig. 3B). When only *P. brasiliensis* isolates are examined, four distinct ancestry clusters are most highly supported (see Materials and Methods), corresponding to S1a, PS2, PS3, and partially S1b. Evaluation

of different values of *k* showed overall support for these major subdivisions; however, *P. brasiliensis* could also be inferred to have 3 primary clusters of S1a, PS2, and PS3 (see Materials and Methods and Fig. S5 in the supplemental material). In both the 3- and 4-cluster analyses, isolates from S1b and PS4 have a subset of sites found in different clusters; there is a unique set of S1b/PS4 sites in the 4-cluster analysis, similar to the private alleles found for S1b (Fig. 2). These patterns support a hybrid ancestry for the S1b and PS4 *P. brasiliensis* isolates, which share SNP markers in different proportions with the S1a and PS3 groups (Fig. 3B). Plotting SNPs colored by ancestry across the genome for the PS4 isolate Pb300 revealed a highly intermixed pattern of small blocks of S1b and PS3 ancestry, suggesting a relatively ancient hybridization event (Fig. 3C).

Equal frequencies of mating types in each lineage. Although the sexual phase of *Paracoccidioides* had not been completely characterized in nature or in the laboratory, the patterns of recombination and ancestral hybridization we found may be most parsimoniously explained by sexual reproduction. We identified the mating type of each isolate based on the genomic sequenced data. While the previously sequenced reference genomes Pb18 (S1b) and Pb03 (PS2) contain the mating type HMG (*MAT1-2*), here we generated the first assembled genome of mating type α (*MAT1-1*) for *P. brasiliensis* (strain PbCnh [PS3]). By aligning the assemblies, we observed that there are not any chromosomal rearrangements near the mating locus that could impact the capacity for interlineage genetic exchange (see Fig. S1F in the supplemental material). The assembly of Pb300 (PS4) also contains the mating type HMG (*MAT1-2*). As previously noted, conservation of mating- and meiosis-specific genes suggests that *P. brasiliensis* has the necessary machinery for sexual reproduction (12), and we see no additional gene loss in these lineages. All of the 31 sequenced *Paracoccidioides* isolates were heterothallic with a population ratio of 1:1 of each mating type, both in the population as a whole as well as among each lineage, with a total of 15 isolates α (*MAT1-1*) and 16 isolates HMG (*MAT1-2*) (see Fig. S6 in the supplemental material). The evidence for genetic exchange and recombination and these roughly equal numbers of both mating types in each lineage support the potential for sexual reproduction in *Paracoccidioides*.

Genome-wide population genetic variation between the major lineages of *Paracoccidioides*. Both polymorphism and phylogenetic analyses suggest that PS3 is a monophyletic group derived from a shared common ancestor with limited diversity, and S1b is the most variable lineage, reflecting a more widespread geographic distribution. We found the highest level of nucleotide diversity in S1b ($\pi = 0.00256$), which was 30-fold greater than the very low nucleotide diversity found in PS3 ($\pi = 0.00008$). Nucleotide diversities in S1a, PS2, and *P. lutzii* were intermediate: 0.00053, 0.00066, and 0.00015, respectively (Fig. 4A). In addition, we tested for genome-wide allele frequency distribution using Tajima's *D* (TD) to scan for signatures of demography and selection. We found that all populations showed genome-wide Tajima's *D* values not significantly different from the null expectation. This suggests that each lineage is evolving neutrally at the whole-genome level (Fig. 4A).

Additionally, we analyzed interspecific divergence between *P. brasiliensis* and *P. lutzii* and between *P. brasiliensis* lineages (S1a, S1b, PS2, and PS3) across the genome using Wright's F_{ST} statistic and a 10-kb sliding-window approach. Across the *P. brasiliensis* pairwise comparisons, values ranged between 0.70 and 0.94, suggesting limited genetic exchange between the lineages (Fig. 4C; see Fig. S7 in the supplemental material). Furthermore, there were no regions of very low F_{ST} that would indicate recent genetic exchange (see Fig. S7). Comparisons of *P. lutzii* and *P. brasiliensis* lineages had values ranging from 0.97 to 0.99, suggesting strong local divergence between these species, similar to the 0.95 F_{ST} value observed for most regions of *Coccidioides immitis* compared to *Coccidioides posadasii* (17). S1b had the lowest pairwise F_{ST} values (0.70 to 0.76) compared to the remaining lineages (0.91 to 0.94) (Fig. 4B; see Fig. S7). This supports the hypothesis of recombination between S1b and the other lineages, which correlates with the intermediate position of S1b in the population structure analysis and higher degree of shared alleles (Fig. 3 and 2, respectively). This may have clinical implications

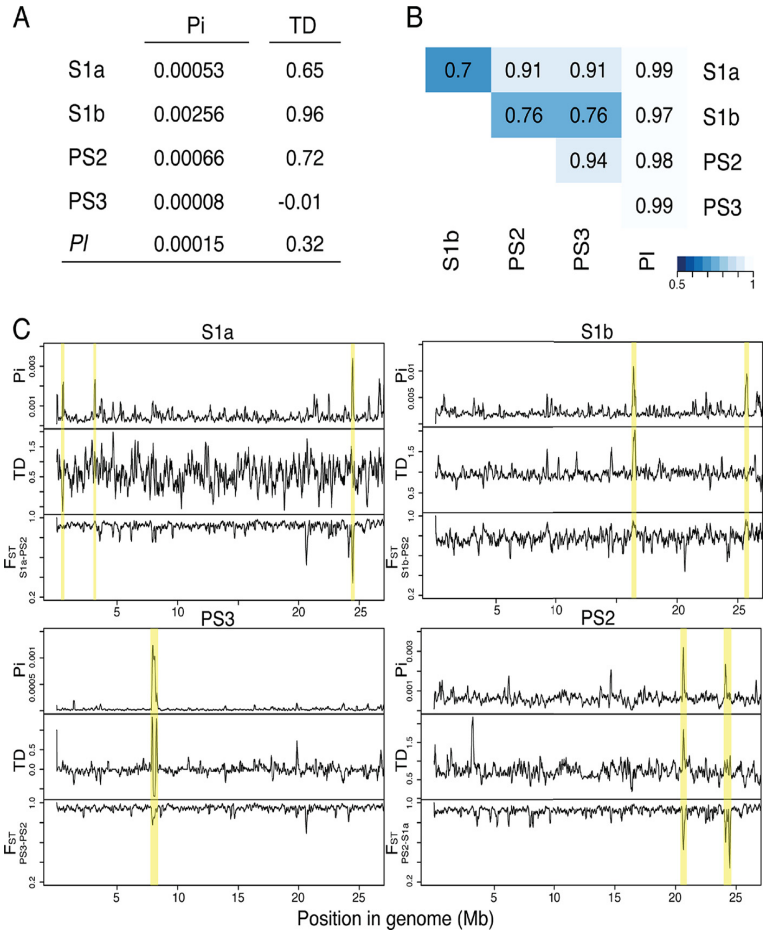


FIG 4 Genome-wide nucleotide diversity (π [Pi]), Tajima's *D* (TD), and population divergence analysis (F_{ST}) in *Paracoccidioides*. (A) Genome-wide average of nucleotide diversity (π) and TD for sliding windows of 10-kb regions within the main *Paracoccidioides* lineages. (B) Average of genome-wide (10-kb windows) variation in (θ), Weir's formulation of Wright's fixation index (F_{ST}), for pairwise comparisons in each lineage. (C) Distribution of the average nucleotide diversity (π), TD, and F_{ST} for sliding windows of 10-kb regions for the S1a, S1b, PS3, and PS3 *Paracoccidioides* lineages.

since the S1b lineage is widely distributed in South America, includes highly virulent strains, and has been associated with the vast majority of cases of PCM (Fig. 1).

Lineage-specific gain and loss and rapid evolution of virulence-associated genes. Comparing the *P. brasiliensis* lineages, a total of 6,670 core ortholog clusters had representative genes from all five reference genomes. We found 720 ortholog groups in at least two strains and 459 ortholog groups that were present in all *P. brasiliensis* strains but absent in *P. lutzii*. We did not find any significant enrichment of functional categories among *P. lutzii* and *P. brasiliensis* strains or among their lineages, which suggests that the phenotypic differences between *P. lutzii* and *P. brasiliensis* and between the lineages are not due to large protein family expansion or contraction. However, unique genes (those without orthologs in other lineages) could contribute to different phenotypic differences observed in each lineage, as well as genes duplicated in only one species. Among the unique genes we identified, there were genes coding for several protein kinases, proteases, transcription factors, and transporters (see Data Set S1 in the supplemental material). More specifically, in *P. lutzii* there were several unique genes coding for anhydrolases, glycosyl hydrolases, peptidases (M24 and C12), and methyltransferases, while in *P. brasiliensis* there were several unique genes coding for actin, transporters, aspartyl proteases, peptidases (M16 and M28), and transcription factors. These unique proteins may provide a more diverse

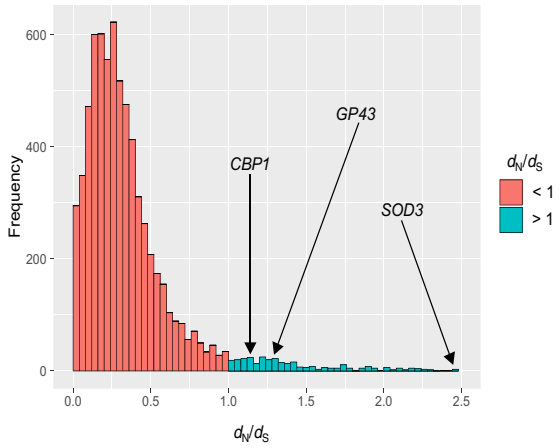


FIG 5 Genes under positive selection in *Paracoccidioides*. Shown is a histogram of the *dN*/*dS* values comparing *P. brasiliensis* (Pb18) and *P. lutzii* (Pb01). Genes undergoing positive selection (*dN*/*dS* of >1) are in blue.

functional repertoire to each *P. lutzii* or *P. brasiliensis* lineage, enabling the fungus different mechanisms and strategies to produce infection and disease.

In addition to identifying strain-specific genes, we identified nonsense mutations that were specific to each lineage. For example, the serine carboxypeptidase gene (*CPDS;* PADG_07980), the predicted mannan endo-1,6-α-mannosidase gene (PADG_00193), and the predicted transmembrane gene (PADG_08161) have nonsense mutations in all lineages of *P. brasiliensis* (S1a, S1b, PS2, PS3, and PS4) but not in *P. lutzii*, suggesting these genes are not functional in *P. brasiliensis*. Other genes that do not have nonsense mutations in *P. brasiliensis* (S1a, S1b, PS2, PS3, and PS4) but have nonsense mutations in all *P. lutzii* strains included genes coding for the separase protease (peptidase_C50; PADG_07698), a DNA-binding protein protease (peptidase_M24; PADG_01032), other secreted proteins (e.g., PADG_00436), as well as other genes coding for proteins involved in different cellular processes (e.g., sugar transporter, PADG_04202; sterigmatocystin biosynthesis monooxygenase StcW, PADG_04703; and a glucose-6-phosphate isomerase, PADG_00451) (see Data Set S1). Differences in gene content between the species, including genes coding for predicted proteases and cell wall genes, could impact how each species/lineage interacts with the host or environment, although other genes with overlapping functions could compensate for loss of these genes.

To more finely examine gene sequence differences and the impact of selection, we calculated the ratio of nonsynonymous to synonymous evolutionary changes (*dN*/*dS*) for each gene across the lineages of *Paracoccidioides*. An average of 485 genes were found to be under positive selection (*dN*/*dS* of >1) in each of the lineages (Fig. 5; see Data Set S1). The set of genes evolving under positive selection includes the surface antigen gene *GP43* (PADG_07615), the superoxide dismutase gene *SOD3* (PADG_02842), the alternative oxidase gene *AOX* (PADG_03747), and the thioredoxin gene (PADG_05504), virulence-associated genes of central importance in *Paracoccidioides* and other dimorphic fungi (Table 2). Other notable genes under positive selection included several protease genes, including a glutamate carboxypeptidase gene (PADG_00686), a multifunctional tryptophan biosynthesis protein gene (PADG_07274), the a-pheromone processing metallopeptidase gene *STE23* (PADG_07053), and the subtilase-type proteinase gene *PSP3* (PADG_07422). In addition, many genes coding for secreted proteins appear under positive selection, including the calcium binding protein gene *CBP1* (PADG_02399) and the mannan endo-1,6-α-mannosidase gene *DCW1* (PADG_01494). Among the 88 total secreted proteins under positive selection, seven were found to be significantly differentially upregulated during a mouse model of infection in *Blastomyces* (14), including a

TABLE 2 Selection of candidate virulence factors in *Paracoccidioides* isolates found in highly diverse regions and/or under selection

Locus ID	Description of protein	π	TD	SNPs	*dN/dS* >1	F_{ST}
PADG_02498	3-Hydroxyanthranilate 3,4-dioxygenase	+	−	−	−	−
PADG_00940	Acetate kinase; *B. dermatitidis* ortholog induced during macrophage interaction	−	+	−	+	−
PADG_01835	Aldehyde reductase; *B. dermatitidis* ortholog induced during macrophage interaction	−	−	−	+	+
PADG_07461	α-1,3-Glucanase	+	+	+	+	+
PADG_03747	Alternative oxidase; *AOX* gene	−	−	−	+	−
PADG_02460	Antigenic GPI-protein; secreted; antigen; *PGA1* gene	+	−	+	+	+
PADG_04167	Aspartyl aminopeptidase; peptidase family M18	+	−	−	−	−
PADG_06131	BUD32 protein kinase; vesicles	−	−	−	+	−
PADG_02399	Calcium binding protein; secreted; *CBP1 gene*	−	−	−	+	+
PADG_00743	Class II aldolase; *B. dermatitidis* ortholog induced during macrophage interaction	−	−	−	+	−
PADG_12370	Endo-1,3(4)-β-glucanase; secreted; *ENG1* gene; *B. dermatitidis* ortholog induced during macrophage interaction	−	+	−	+	−
PADG_05497	GATA-binding protein; *B. dermatitidis* ortholog induced during macrophage interaction	−	−	−	+	−
PADG_07615	Glucan 1,3-β-glucosidase; secreted; antigen; *GP43* gene; induced during macrophage interaction	−	+	+	+	−
PADG_05345	High-affinity nickel transporter; *B. dermatitidis* ortholog induced during *in vivo* infection	−	−	−	+	−
PADG_07274	Hypothetical protein	−	−	−	+	−
PADG_06699	Hypothetical protein; *B. dermatitidis* ortholog induced during macrophage interaction	−	+	−	+	−
PADG_01238	Hypothetical protein; *B. dermatitidis* ortholog induced during macrophage interaction	−	−	−	+	−
PADG_01283	Hypothetical protein; *B. dermatitidis* ortholog induced during macrophage interaction	−	−	−	+	−
PADG_03908	Hypothetical protein; *B. dermatitidis* ortholog induced during macrophage interaction	−	−	−	+	−
PADG_07534	Hypothetical protein; *B. dermatitidis* ortholog induced during macrophage interaction	−	−	−	+	−
PADG_11963	Hypothetical protein; *B. dermatitidis* ortholog induced during macrophage interaction	−	−	−	+	+
PADG_02535	Hypothetical protein; secreted; vesicles; *B. dermatitidis* ortholog induced during *in vivo* infection	+	+	+	+	+
PADG_02542	Hypothetical protein; vesicles	−	−	−	+	+
PADG_12450	Hypothetical protein; vesicles	+	−	+	+	−
PADG_04741	Hypothetical protein; vesicles	+	+	−	+	−
PADG_02521	Hypothetical protein; vesicles	+	−	−	+	−
PADG_12101	Hypothetical protein; vesicles	−	−	−	+	−
PADG_01494	Mannan endo-1,6-α-mannosidase; secreted; *DCW1* gene	−	+	−	+	−
PADG_00948	Oxidoreductase; *B. dermatitidis* ortholog induced during *in vivo* infection	−	+	−	+	−
PADG_07460	Predicted aminopeptidase; peptidase family M18; induced during macrophage interaction	+	+	+	−	+
PADG_05820	Predicted aminopeptidase; peptidase family M24; induced during macrophage interaction	+	−	+	−	−
PADG_07369	Predicted dehydrogenase	+	+	+	−	+
PADG_02527	Predicted dehydrogenase	+	−	−	−	−
PADG_02562	Predicted dehydrogenase	+	+	−	−	−
PADG_02492	Predicted dehydrogenase	+	−	−	−	−
PADG_07365	Predicted dehydrogenase	+	−	−	−	+
PADG_07411	Predicted dehydrogenase	+	−	−	−	−
PADG_02575	Predicted nonribosomal peptide synthetase	+	+	−	−	−
PADG_06309	Predicted oxidoreductase	+	+	+	−	+
PADG_02592	Predicted oxidoreductase	+	+	+	−	−
PADG_06322	Predicted peroxidase	+	+	−	−	+
PADG_02507	Predicted peroxidase; secreted	+	−	−	+	−
PADG_07053	Predicted protease; peptidase family 16	−	−	−	+	−
PADG_00686	Predicted protease; peptidase family M28	−	−	−	+	+
PADG_06314	Predicted protease; peptidase family S10; secreted; induced during macrophage interaction	+	+	−	−	−
PADG_06167	Predicted protease; peptidase family S24	+	−	−	+	−
PADG_07422	Predicted protease; peptidase family S8; secreted	−	+	−	+	+

(Continued on following page)

TABLE 2 (Continued)

Locus ID	Description of protein	π	TD	SNPs	*dN/dS* >1	F_{ST}
PADG_07454	Predicted scramblase; secreted	+	−	+	+	+
PADG_06308	Predicted transporter	+	+	+	+	+
PADG_02081	RING finger domain-containing protein; vesicles	−	+	−	+	−
PADG_08583	Secreted protein	−	−	−	+	−
PADG_02569	Secreted protein	+	+	−	−	−
PADG_00954	Secreted protein immunoreactive protein; secreted	+	+	+	+	−
PADG_07830	Secreted protein; *B. dermatitidis* ortholog induced during *in vivo* infection	−	−	−	+	−
PADG_05055	Secreted protein; *B. dermatitidis* ortholog induced during macrophage interaction	+	+	+	+	−
PADG_03277	Secreted protein; vesicles; *B. dermatitidis* ortholog induced during *in vivo* infection	−	−	−	+	−
PADG_02842	Superoxide dismutase; secreted; *SOD3* gene; *B. dermatitidis* ortholog induced during *in vivo* infection	−	−	−	+	−
PADG_05504	Thioredoxin	−	−	−	+	−
PADG_02515	Transporter	+	−	−	+	−
PADG_05881	Vacuolar amino acid transporter; vesicles	−	−	−	+	−
PADG_00941	Xylulose-5-phosphate phosphoketolase; *B. dermatitidis* ortholog induced during macrophage interaction	−	+	−	+	−

high-affinity nickel transporter (PADG_05345), an oxidoreductase (PADG_00948) and the *SOD3*-encoded protein. Other genes that were found significantly induced during the interaction of *Blastomyces* with macrophages (14) were also found to evolve under positive selection in *Paracoccidioides*, including the two most highly upregulated in infected macrophages, a gene coding for a secreted protein with unknown function (PADG_01283) and a secreted endo-1,3(4)-β-glucanase gene, *ENG1* (PADG_12370); other genes found in both analyses include genes coding for GATA-binding protein (PADG_05497), an aldehyde reductase (PADG_01835), and a glucan 1,3-β-glucosidase (PADG_06699) (Table 2; see Data Set S1). These genes therefore may be generally important for host interactions of dimorphic pathogens and make good candidates for future studies on their contribution to virulence.

Combining diversity measures reveals lineage-specific targets of positive selection. Examination of the local variation in nucleotide diversity (π) and Tajima's *D* (TD) across the genome revealed small regions with high lineage-specific diversity, which also typically showed high TD (Fig. 4C; see Data Set S1 in the supplemental material). PS2 had the most lineage-specific high-diversity windows, with 14 in total, coinciding with the more ancient divergence of this lineage and in agreement with the phylogenetic analysis. The windows with significantly high π and TD highlighted lineage-specific regions of high diversity and selection, which in some cases were regions with low interspecific divergence (F_{ST}) (Fig. 4B).

These high-diversity regions highlighted different sets of variable and rapidly evolving genes within lineage. These genes encompass diverse cellular functions, including coding for secreted and cell wall proteins, transport, transcription regulation, oxidative stress, and proteolysis (Table 2; see Data Set S1), some of which have experimental evidence of a role in virulence and pathogenicity (see Text S1 in the supplemental material). Two high-diversity regions found for S1b include two aminopeptidases (PADG_05820 and PADG_07460) previously noted to be upregulated in *Paracoccidioides* during macrophage infection (18). In the low-diversity PS3 isolates, one large highly variable region was identified that includes 125 genes (23 with a *dN/dS* of >1, 12 secreted protein genes, and 9 genes involved with the oxidation reduction process). Among these genes are genes encoding two glycosylphosphatidylinositol (GPI)-anchored proteins (including *PGA1*) and the secreted protein PADG_02535, which was identified in extracellular vesicles of *Paracoccidioides* (19), and it was recently shown in *B. dermatitidis* that an ortholog was more highly expressed during mouse pulmonary infection (14). In two regions found for PS2, 11 of 48 genes show evidence of positive selection (*dN/dS* of >1) and include genes coding for a serine carboxypeptidase

(PADG_06314) induced in *Paracoccidioides* during macrophage infection (18) and a peroxidase (PADG_06322), as well as the amino acid permease PADG_07440, which in *B. dermatitidis* was highly induced during mouse pulmonary infection (14). Differences in these genes might impact the virulence phenotypes observed across the lineages and make good candidates for future studies of their contribution to virulence variation between isolates of *Paracoccidioides* (Table 2).

DISCUSSION

Building on previous genomic analysis of *Paracoccidioides* (12, 13), here we reevaluate the major lineages and provide new reference genomes for two lineages, PS3 and PS4. Together these data enable a more comprehensive view of the genome content of *Paracoccidioides* and help trace the genome-level variation following the recent divergence into well-defined lineages. We find clear support for the separation of *P. brasiliensis* into distinct lineages; however, we also find evidence of recombination and highest diversity within a single lineage, the newly described S1b. Dating the timing of the separation of each lineage supported S1b as the earliest diverging *P. brasiliensis* lineage, with the later emergence of PS2, S1a, and PS3.

Our study provides additional support for *P. brasiliensis* and *P. lutzii* as separate species, with some amount of incomplete lineage sorting suggested by intermediate F_{ST} values between the *P. brasiliensis* lineages. Species definitions have been reevaluated in many fungal groups, including the *Onygenales* (20), based on improved phylogenies from wider sets of isolates compared or incorporation of additional loci, or even whole genomes, in addition to the analysis of morphological data of sexual and asexual structures and the evidence of sexual reproduction. Using multiple approaches, we see clear separation of *P. brasiliensis* and *P. lutzii* and evidence of recombination within *P. brasiliensis*; however, this recombination did not appear to be recent. While the phylogenetic separation suggests the *P. brasiliensis* lineages are largely genetically isolated, increased sample size is required to compare variation within and between the lineages to support their separation into different species.

We found support for a split in the S1 group into S1a and S1b lineages; in addition to genetic separation, signatures of recombination clearly differentiate S1b from S1a. While the phylogenetic analysis suggests that S1a and S1b are both monophyletic, we also find evidence from multiple analyses that S1b, the most geographically widespread lineage, has undergone recombination with each of the other lineages. This appears to be mostly ancestral, as F_{ST} while variable did not reveal any large recently introgressed regions between lineages. The mixed ancestry of PS4 also suggests that more ancient recombination created this lineage, as the pattern of SNP ancestry across the genome revealed an intermixed distribution of both S1b and PS3 alleles that were likely shuffled by extensive recombination over time. The sequence of additional isolates of PS4 as well as S1b could help further explore how alleles within these groups are distributed within each lineage and across geographic regions.

All lineages, including S1a and S1b, contain roughly equal numbers of both mating types based on the sequenced isolates; an equal ratio was also previously noted in a wider set of *P. brasiliensis* isolates (21). The equal representation suggests that other differences between the strains explain how S1b has undergone higher levels of recombination compared to S1a. Direct testing of mating potential for different strains requires the further development of laboratory mating experiments, which appear to only initiate but not complete mating to date (22). In addition, further studies comparing larger numbers of isolates from the same geographic region but different lineage groups would be more sensitive to detecting recent exchange between groups in nature.

This genome-wide comparison of the major lineages of *Paracoccidioides* revealed that while the genome organization and gene content are highly conserved, specific genes under positive selection include several previously known to be important for host interaction as well as new candidates. Two of the genes we identified (*GP43* and *CTS20*) were noted in a previous study, which utilized expressed sequence tags (ESTs) and found evidence of positive selection in 11 of 32 examined genes implicated in

virulence (23). Here, we find evidence for positive selection in an average of 485 genes per strain using the updated genome sequences. In addition to *GP43*, we see evidence of selection in the gene coding for the secreted antigenic GPI-anchored protein, *PGA1* (24). Other genes under positive selection include those coding for aminopeptidases highly upregulated during host-pathogen interactions (14), secreted proteins previously found in *Paracoccidioides* extracellular vesicles (19), and other proteins involved in mitigating oxidative stress that were also found induced during the interaction with macrophages, such as the *AOX*, *SOD3*, and *CBP1* genes (14, 18, 25–27). Secreted proteins in *Paracoccidioides* have been associated with nutrient acquisition, cell defense, and modulation of the host defense machinery.

Better knowledge of the differences between the *Paracoccidioides* lineages can help guide the development of new diagnostics, treatments, and models of pathogen evolution. Delineation of the prevalent S1 group into the well-separated S1a and S1b lineages allowed us to differentiate S1b as the more highly recombining and diverse lineage. The higher diversity present in S1b may contribute to its dispersion and survival in a wider geographical range than the other lineages, perhaps enabling local adaptation such as to temperature variation. In our data, all of the S1b strains are clinical isolates and include two highly virulent strains, Pb18 and Pb113, and two strains from acute PCM cases (PbBlo and PbCaz). In contrast, the S1a lineage includes clinical and armadillo isolates, with very few sequence differences between strains from these sources. Sequence of isolates from a wider geographical range, including Brazil, Argentina, Paraguay, and Mexico, would likely increase the sampled genetic diversity of *Paracoccidioides* and allow further study of these trends; additional isolates could also enable genome-wide association studies of clinically relevant phenotypes. Development of sequence-based diagnostics will need to take into consideration that S1b strains share a higher percentage of alleles with all other strains; as these sites could contribute to misidentification of lineages, sufficient power or selection of sites unique to each group that also take into account the higher diversity in S1b would improve diagnostic power.

MATERIALS AND METHODS

Selection and sequencing of *Paracoccidioides* isolates. A total of 31 isolates of *P. brasiliensis* and *P. lutzii* were sequenced and included in the analyses (Table 1): 11 isolates from lineage S1a, 4 isolates from S1b, 7 isolates from PS3, 4 isolates from PS2, 1 isolate from PS4, and 4 isolates from *P. lutzii*. Selected isolates included clinical isolates from acute and chronic PCM cases, environmental isolates from armadillos, and the previously published and updated reference genomes of *P. brasiliensis* (Pb03 from PS2 and Pb18 from S1) and Pb01 from *P. lutzii* (12, 13). Genomic DNA for sequencing was prepared from yeast culture, using phenol-chloroform extraction method. The genomes were sequenced using Illumina sequencing platforms and using different insert sizes (~620, ~180, or ~650 bp), paired-end-read lengths (150, 101, or 100 bp) and sequence coverage ranging from 68 to 419 (see Table S1 in the supplemental material).

Identification and analysis of gene orthologs and selection analysis. For a detailed description of the genome assembly and gene annotation of *P. brasiliensis* Pb300 and PbCnh strains, see the supplemental material. The five assembled and annotated genomes representing each lineage of the *Paracoccidioides* species were used for comparative analysis. Genes were functionally annotated by assigning Pfam domains, GO terms, KEGG classification, SignalP, and TMHMM. OrthoMCL (28) was used to cluster the protein-coding genes of the four chosen genomes by similarity and create sets of genes that have a high probability of being orthologous to each other.

For selection analysis, variant call format (VCF) files were used to align the reference transcript set in coding triplets in the PHYLIP multiple sequence alignment format. Then we calculated the ratio of nonsynonymous to synonymous substitutions (*dN*/*dS*) within the *Paracoccidioides* genus on properly aligned genes. We employed the yn00 program in PAML (29), implementing the Yang and Nielsen method (30). To eliminate the possibility that fast-evolving genes detected in this analysis were biased for false-positive SNP calls, we calculated the mean base quality, mean mapping quality, and quality normalized to depth and compared these parameters for all fast-evolving genes and all other genes.

SNP variant detection and analysis. To detect polymorphisms, we used reference assemblies representing each lineage (Pb18, Pb03, PbCnh, Pb300, and Pb01). Each of the 31 Illumina data sets was independently aligned to the genome assemblies using the short read component *aln* of BWA version 0.5.9 (31) with default settings. SNPs and indels were called with Pilon version 1.4 using the haploid ploidy default setting (32). An average of 38 million reads was aligned per strain, with an average quality and error rate of 33.2 and 2.0E−02, respectively (see Table S1 in the supplemental material). Variant call format (VCF) files were filtered using VCFtools version 0.1.12 (33) or according to further analyses—e.g.,

considering SNP calls in all strains with genotype 1/1 and minimum depth 4. For SNP positions, the total mapping depth was 125 and the mean base quality was 34, averaged across all variants (see Table S1).

To address if any lineage could be uniformly diploid, we examined candidate heterozygous positions predicted by Pilon. The low frequency of such positions (~0.04%), which often overlap repetitive sequence (68% of these positions), suggests there is little evidence for diploidy. This suggests that all sequenced genomes are homozygous haploids, as expected from prior work establishing that *Paracoccidioides* species are haploid (34, 35).

The false discovery rate (FDR) was estimated as an additional parameter of the mapping and SNP calling accuracy (36). As a truth set, we simulated 670,000 random mutations in the reference genome (Pb18). The number of random mutations was the maximum number of mutations detected with the sequenced strains. Next, we aligned the raw reads to the random-mutated reference, called SNPs, and compared them to the known truth set of simulated mutations to calculate the accuracy of our data and process. Both the number of true positives (658,566; precision, 99.8%, and sensitivity, 98.3%) and the number of false positives (386; 0.06%) show the high accuracy of the overall process, including read quality, alignment accuracy, and the SNP calls.

To determine the chromosomal copy number variation (CCNV) and the distribution of SNPs across the genome, we calculated the alignment density (depth of read coverage) and SNP density (frequency of SNPs per site), respectively. All VCF files were summarized in 10-kb nonoverlapping windows. The alignment density was normalized by the average genomic coverage and the window length. The SNP density was normalized by the window length. The genome assembly of Pb18 is anchored to chromosomes (12); we used this reference to map the alignment density and SNP density into chromosomes. Pb03, PbCnh, Pb300, and Pb01 were mapped into scaffolds.

To more finely compare variant sites between lineages, we classified the distribution of biallelic loci between two populations categorized as "shared," "fixed," or "private." If one allele is present in all members of one population and the other allele is present in all members of the other population, that locus was considered "fixed." If both alleles are present in both populations, that locus was considered "shared." If one allele is present in all members of one population and the other population has both alleles, that locus was considered "private."

To analyze each SNP in the context of gene annotations of the reference genomes we used VCFannotator (http://sourceforge.net/projects/vcfannotator/). Variants were placed in the context of genome feature annotations and indicated as "intergenic," "intronic," or "coding." The type of coding mutation was further characterized by its impact on the protein-coding sequence as "synonymous," "nonsynonymous," "nonsense," or "read-through." To evaluate the support for each mutation, the mean base quality, the mean mapping quality, and the quality normalized to depth were retrieved and compared.

Phylogenetic analysis using whole-genome SNPs. Alignments were constructed from SNP matrices extracted from the VCF format. To consider a position in the alignment matrix, we used a minimum depth of coverage of 4, and we kept those positions with at least one variant site in all the sequenced isolates. We built an SNP alignment matrix for each *Paracoccidioides* reference strain used here representing each lineage to obtain four trees and compare their topologies. Maximum likelihood phylogenies were constructed using RAxML version 8.0.20 (37) using the GTRCAT nucleotide substitution model and bootstrap analysis based on 1,000 replicates. To infer Bayesian phylogenetic trees and to estimate the divergence and most common ancestor dates, we used BEAST v1.8.2 (38). Using the SNP matrices, we normalized the mutation rate for genome-wide variable sites (normalized, 4.43E−8; genome-wide, 1E−9 [39]) that was used as the clock rate along with the coalescent model, relaxed clock/uncorrelated lognormal clock model. For nonpartitioned variant sites, we used the strict clock model, with the constant pattern model, general time reversible (GTR) (G+I) substitution model, and 20,000,000 Markov chain Monte Carlo (MCMC) chains. Tracer v1.6.0 (http://beast.bio.ed.ac.uk/Tracer) was used to visualize traces and inspect the effective sample size and MCMC. We determined the relationship of the sequenced isolates using the NeighborNet algorithm with SplitsTree4 (40).

Population genetic structure analyses. We performed a principal-component analysis (PCA) on a matrix of SNP calls for all of the *Paracoccidioides* isolates ($n = 31$) and for only the *P. brasiliensis* isolates ($n = 26$), using SMARTPCA (41). Population structure was performed using the Bayesian model-based clustering program STRUCTURE v2.3 (16) in the site-by-site mode, with successive k values from 2 to 6. We identified populations within *Paracoccidioides* using 476,589 SNPs in all isolates and within *P. brasiliensis* isolates using 339,966 SNPs. We estimated the model evidence for k using thermodynamic integration method as implemented in MavericK v1.0 (42), using replicates for 10 randomly generated 1% subsamples of the *P. brasiliensis* SNP matrix; this supported $k = 4$ as the best choice for *P. brasiliensis* (see Fig. S5B in the supplemental material). Genome-wide nucleotide diversity (π) and Tajima's D were computed for each identified *Paracoccidioides* population (S1a, S1b, PS2, PS3, and *P. lutzii*) using VCFtools v0.1.12 (33). The average nucleotide diversity (π) and Tajima's D were computed for nonoverlapping sliding windows of 10 kb. We calculated Wright's fixation index (F_{ST} [43]) according to the equations given in VCFtools v0.1.12 (33) adjusted for low sample sizes. Sliding-window F_{ST} analyses were conducted using all SNPs found within 10-kb nonoverlapping windows. Variant call format (VCF) files were filtered using an in-house perl script to calculate F_{ST}, and clusters for comparison were chosen based on the whole-genome phylogenetic tree.

Accession number(s). The assemblies and annotations of the *P. brasiliensis* genomes have been deposited in DDBJ/ENA/GenBank under the following accession numbers: *Paracoccidioides brasiliensis* PbCnh, LYUC00000000; *Paracoccidioides brasiliensis* Pb300, LZYO00000000. All of the whole-genome sequence (WGS) raw data for the 31 *Paracoccidioides* strains have been deposited in the NCBI Sequence

Read Archive (BioProject, PRJNA322632; SRA, SRP077566). BioSample and SRA accession numbers for individual strains are included in Table S1 in the supplemental material.

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at http://dx.doi.org/10.1128/mSphere.00213-16.

Text S1, DOCX file, 0.1 MB.
Data Set S1, XLSX file, 0.6 MB.
Figure S1, PDF file, 1.9 MB.
Figure S2, PDF file, 2.1 MB.
Figure S3, PDF file, 0.5 MB.
Figure S4, PDF file, 1.1 MB.
Figure S5, PDF file, 0.2 MB.
Figure S6, PDF file, 0.3 MB.
Figure S7, PDF file, 0.5 MB.
Table S1, XLSX file, 0.1 MB.

ACKNOWLEDGMENTS

We thank Angela Restrepo, Rosana Puccia, Zoilo Pires de Camargo, and Maria Sueli Felipe for kindly providing the isolates for this study.

This project has been funded in whole or in part with federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under contract no. HHSN272200900018C. This work was partly supported by Colciencias via the grants "A Gene Atlas for Human Pathogenic Fungi" (122256934875) and "A Comprehensive Genomic and Transcriptomic Analysis of Dimorphic Human Pathogen Fungi and Its Relation with Virulence" (221365842971) and by the Universidad de Antioquia via a "Sostenibilidad 2015/2016" grant. Colciencias National Doctorate Program funding supported J.F.M.; Enlaza Mundos partly supported his fellowship. The Wellcome Trust supported R.A.F.

J.F.M., J.G.M., O.K.C., and C.A.C. conceived and designed the experiments. J.F.M., R.A.F., C.A.D., and C.A.C. analyzed the data. J.F.M. and C.A.C. assembled and annotated *Paracoccidioides* genomes. J.E.G., S.M.S., S.S., E.M., E.A.W., E.B., C.M.A.S., M.D.M.T., and J.W.T. contributed reagents, materials, and/or analysis tools. J.F.M., C.A.C., C.A.D., J.G.M., and O.K.C. wrote the manuscript. J.F.M., R.A.F., C.A.D., C.A.C., J.E.G., S.M.S., S.S., E.M., E.A.W., E.B., C.M.A.S., M.D.M.T., and J.W.T. performed the experiments.

FUNDING INFORMATION

This work, including the efforts of José F. Muñoz, Juan E. Gallo, Elizabeth Misas, Oliver K. Clay, and Juan G. McEwen, was funded by Colciencias (122256934875 and 221365842971). This work, including the efforts of José F. Muñoz, Rhys A. Farrer, Christopher A. Desjardins, and Christina A. Cuomo, was funded by HHS | NIH | National Institute of Allergy and Infectious Diseases (NIAID) (HHSN272200900018C). This work, including the efforts of Rhys A. Farrer, was funded by Wellcome Trust. This work, including the efforts of Marcus de M. Teixeira, was funded by Conselho Nacional de Desenvolvimento Científico e Tecnológico (460999/2014-1).

REFERENCES

1. **Bocca AL, Amaral AC, Teixeira MM, Sato PK, Shikanai-Yasuda MA, Soares Felipe MS**. 2013. Paracoccidioidomycosis: eco-epidemiology, taxonomy and clinical and therapeutic issues. Future Microbiol **8:**1177–1191. http://dx.doi.org/10.2217/fmb.13.68.
2. **Bellissimo-Rodrigues F, Vitali LH, Martinez R**. 2010. Serological diagnosis of paracoccidioidomycosis in HIV-coinfected patients. Mem Inst Oswaldo Cruz **105:**904–907. http://dx.doi.org/10.1590/S0074-02762010000700011.
3. **Brummer E, Castaneda E, Restrepo A**. 1993. Paracoccidioidomycosis: an update. Clin Microbiol Rev **6:**89–117. http://dx.doi.org/10.1128/CMR.6.2.89.
4. **Teixeira MM, Theodoro RC, de Carvalho MJ, Fernandes L, Paes HC, Hahn RC, Mendoza L, Bagagli E, San-Blas G, Felipe MS**. 2009. Phylogenetic analysis reveals a high level of speciation in the *Paracoccidioides* genus. Mol Phylogenet Evol **52:**273–283. http://dx.doi.org/10.1016/j.ympev.2009.04.005.
5. **Matute DR, McEwen JG, Puccia R, Montes BA, San-Blas G, Bagagli E, Rauscher JT, Restrepo A, Morais F, Niño-Vega G, Taylor JW**. 2006. Cryptic speciation and recombination in the fungus *Paracoccidioides brasiliensis* as revealed by gene genealogies. Mol Biol Evol **23:**65–73. http://dx.doi.org/10.1093/molbev/msj008.
6. **Teixeira MM, Theodoro RC, Nino-Vega G, Bagagli E, Felipe MSS**. 2014. *Paracoccidioides* species complex: ecology, phylogeny, sexual reproduction, and virulence. PLoS Pathog **10:**e1004397. http://dx.doi.org/10.1371/journal.ppat.1004397.
7. **Macoris SA, Sugizaki MF, Peraçoli MT, Bosco SM, Hebeler-Barbosa F,**

Simões LB, Theodoro RC, Trinca LA, Bagagli E. 2006. Virulence attenuation and phenotypic variation of *Paracoccidioides brasiliensis* isolates obtained from armadillos and patients. Mem Inst Oswaldo Cruz **101:** 331–334. http://dx.doi.org/10.1590/S0074-02762006000300019.

8. **Theodoro RC, Teixeira MdM, Felipe MS, dos Santos Paduan K, Ribolla PM, San-Blas G, Bagagli E**. 2012. Genus Paracoccidioides: species recognition and biogeographic aspects. PLoS One **7:**e37694. http://dx.doi.org/10.1371/journal.pone.0037694.
9. **Montoya AE, Alvarez AL, Moreno MN, Restrepo A, McEwen JG**. 1999. Electrophoretic karyotype of environmental isolates of *Paracoccidioides brasiliensis*. Med Mycol **37:**219–222. http://dx.doi.org/10.1080/j.1365-280X.1999.00217.x.
10. **Soares CM, Madlun EE, da Silva SP, Pereira M, Felipe MS**. 1995. Characterization of *Paracoccidioides brasiliensis* isolates by random amplified polymorphic DNA analysis. J Clin Microbiol **33:**505–507.
11. **Puccia R, McEwen JG, Cisalpino PS**. 2008. Diversity in *Paracoccidioides brasiliensis*. The PbGP43 gene as a genetic marker. Mycopathologia **165:**275–287. http://dx.doi.org/10.1007/s11046-007-9055-2.
12. **Desjardins CA, Champion MD, Holder JW, Muszewska A, Goldberg J, Bailão AM, Brigido MM, Ferreira ME, Garcia AM, Grynberg M, Gujja S, Heiman DI, Henn MR, Kodira CD, León-Narváez H, Longo LV, Ma LJ, Malavazi I, Matsuo AL, Morais FV, Pereira M, Rodriguez-Brito S, Sakthikumar S, Salem-Izacc SM, Sykes SM, Teixeira MM, Vallejo MC, Walter ME, Yandava C, Young S, Zeng Q, Zucker J, Felipe MS, Goldman GH, Haas BJ, McEwen JG, Nino-Vega G, Puccia R, San-Blas G, Soares CM, Birren BW, Cuomo CA**. 2011. Comparative genomic analysis of human fungal pathogens causing paracoccidioidomycosis. PLoS Genet **7:**e1002345. http://dx.doi.org/10.1371/journal.pgen.1002345.
13. **Muñoz JF, Gallo JE, Misas E, Priest M, Imamovic A, Young S, Zeng Q, Clay OK, McEwen JG, Cuomo CA**. 2014. Genome update of the dimorphic human pathogenic fungi causing paracoccidioidomycosis. PLoS Negl Trop Dis **8:**e3348. http://dx.doi.org/10.1371/journal.pntd.0003348.
14. **Muñoz JF, Gauthier GM, Desjardins CA, Gallo JE, Holder J, Sullivan TD, Marty AJ, Carmen JC, Chen Z, Ding L, Gujja S, Magrini V, Misas E, Mitreva M, Priest M, Saif S, Whiston EA, Young S, Zeng Q, Goldman WE, Mardis ER, Taylor JW, McEwen JG, Clay OK, Klein BS, Cuomo CA**. 2015. The dynamic genome and transcriptome of the human fungal pathogen *Blastomyces* and close relative *Emmonsia*. PLoS Genet **11:**e1005493. http://dx.doi.org/10.1371/journal.pgen.1005493.
15. **Bruen TC, Philippe H, Bryant D**. 2006. A simple and robust statistical test for detecting the presence of recombination. Genetics **172:** 2665–2681. http://dx.doi.org/10.1534/genetics.105.048975.
16. **Pritchard JK, Stephens M, Donnelly P**. 2000. Inference of population structure using multilocus genotype data. Genetics **155:**945–959.
17. **Neafsey DE, Barker BM, Sharpton TJ, Stajich JE, Park DJ, Whiston E, Hung CY, McMahan C, White J, Sykes S, Heiman D, Young S, Zeng Q, Abouelleil A, Aftuck L, Bessette D, Brown A, FitzGerald M, Lui A, Macdonald JP, Priest M, Orbach MJ, Galgiani JN, Kirkland TN, Cole GT, Birren BW, Henn MR, Taylor JW, Rounsley SD**. 2010. Population genomic sequencing of *Coccidioides* fungi reveals recent hybridization and transposon control. Genome Res **20:**938–946. http://dx.doi.org/10.1101/gr.103911.109.
18. **Parente-Rocha JA, Parente AF, Baeza LC, Bonfim SM, Hernandez O, McEwen JG, Bailão AM, Taborda CP, Borges CL, Soares CM**. 2015. Macrophage interaction with *Paracoccidioides brasiliensis* yeast cells modulates fungal metabolism and generates a response to oxidative stress. PLoS One **10:**e0137619. http://dx.doi.org/10.1371/journal.pone.0137619.
19. **Peres da Silva R, Heiss C, Black I, Azadi P, Gerlach JQ, Travassos LR, Joshi L, Kilcoyne M, Puccia R**. 2015. Extracellular vesicles from *Paracoccidioides* pathogenic species transport polysaccharide and expose ligands for DC-SIGN receptors. Sci Rep **5:**14213. http://dx.doi.org/10.1038/srep14213.
20. **Schwartz IS, Kenyon C, Feng P, Govender NP, Dukik K, Sigler L, Jiang Y, Stielow JB, Muñoz JF, Cuomo CA, Botha A, Stchigel AM, de Hoog GS**. 2015. 50 years of *Emmonsia* disease in humans: the dramatic emergence of a cluster of novel fungal pathogens. PLoS Pathog **11:** e1005198. http://dx.doi.org/10.1371/journal.ppat.1005198.
21. **Torres I, García AM, Hernández O, González A, McEwen JG, Restrepo A, Arango M**. 2010. Presence and expression of the mating type locus in *Paracoccidioides brasiliensis* isolates. Fungal Genet Biol **47:**373–380. http://dx.doi.org/10.1016/j.fgb.2009.11.005.
22. **Teixeira MdM, Theodoro RC, Derengowsky LdS, Nicola AM, Bagagli E, Felipe MS**. 2013. Molecular and morphological data support the existence of a sexual cycle in species of the genus *Paracoccidioides*. Eukaryot Cell **12:**380–389. http://dx.doi.org/10.1128/EC.05052-11.
23. **Matute DR, Quesada-Ocampo LM, Rauscher JT, McEwen JG**. 2008. Evidence for positive selection in putative virulence factors within the *Paracoccidioides brasiliensis* species complex. PLoS Negl Trop Dis **2:**e296. http://dx.doi.org/10.1371/journal.pntd.0000296.
24. **Valim CX, Basso LR, Jr, dos Reis Almeida FB, Reis TF, Damásio AR, Arruda LK, Martinez R, Roque-Barreira MC, Oliver C, Jamur MC, Coelho PS**. 2012. Characterization of PbPga1, an antigenic GPI-protein in the pathogenic fungus *Paracoccidioides brasiliensis*. PLoS One **7:**e44792. http://dx.doi.org/10.1371/journal.pone.0044792.
25. **Tamayo D, Muñoz JF, Lopez Á, Urán M, Herrera J, Borges CL, Restrepo Á, Soares CM, Taborda CP, Almeida AJ, McEwen JG, Hernández O**. 2016. Identification and analysis of the role of superoxide dismutases isoforms in the pathogenesis of *Paracoccidioides* spp. PLoS Negl Trop Dis **10:**e0004481. http://dx.doi.org/10.1371/journal.pntd.0004481.
26. **Ruiz OH, Gonzalez A, Almeida AJ, Tamayo D, Garcia AM, Restrepo A, McEwen JG**. 2011. Alternative oxidase mediates pathogen resistance in *Paracoccidioides brasiliensis* infection. PLoS Negl Trop Dis **5:**e1353. http://dx.doi.org/10.1371/journal.pntd.0001353.
27. **Sebghati TS, Engle JT, Goldman WE**. 2000. Intracellular parasitism by *Histoplasma capsulatum*: fungal virulence and calcium dependence. Science **290:**1368–1372. http://dx.doi.org/10.1126/science.290.5495.1368.
28. **Li L, Stoeckert CJ, Jr, Roos DS**. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res **13:**2178–2189. http://dx.doi.org/10.1101/gr.1224503.
29. **Yang Z**. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol **24:**1586–1591. http://dx.doi.org/10.1093/molbev/msm088.
30. **Yang Z, Nielsen R**. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol **17:**32–43. http://dx.doi.org/10.1093/oxfordjournals.molbev.a026236.
31. **Li H, Durbin R**. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics **25:**1754–1760. http://dx.doi.org/10.1093/bioinformatics/btp324.
32. **Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM**. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One **9:**e112963. http://dx.doi.org/10.1371/journal.pone.0112963.
33. **Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R, 1000 Genomes Project Analysis Group**. 2011. The variant call format and VCFtools. Bioinformatics **27:**2156–2158.
34. **Almeida AJ, Matute DR, Carmona JA, Martins M, Torres I, McEwen JG, Restrepo A, Leão C, Ludovico P, Rodrigues F**. 2007. Genome size and ploidy of *Paracoccidioides brasiliensis* reveals a haploid DNA content: flow cytometry and GP43 sequence analysis. Fungal Genet Biol **44:** 25–31. http://dx.doi.org/10.1016/j.fgb.2006.06.003.
35. **Feitosa LdS, Cisalpino PS, dos Santos MR, Mortara RA, Barros TF, Morais FV, Puccia R, da Silveira JF, de Camargo ZP**. 2003. Chromosomal polymorphism, syntenic relationships, and ploidy in the pathogenic fungus *Paracoccidioides brasiliensis*. Fungal Genet Biol **39:**60–69. http://dx.doi.org/10.1016/S1087-1845(03)00003-3.
36. **Farrer RA, Henk DA, MacLean D, Studholme DJ, Fisher MC**. 2013. Using false discovery rates to benchmark SNP-callers in next-generation sequencing projects. Sci Rep **3:**1512. http://dx.doi.org/10.1038/srep01512.
37. **Stamatakis A**. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics **22:**2688–2690. http://dx.doi.org/10.1093/bioinformatics/btl446.
38. **Drummond AJ, Suchard MA, Xie D, Rambaut A**. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol **29:**1969–1973. http://dx.doi.org/10.1093/molbev/mss075.
39. **Sharpton TJ, Stajich JE, Rounsley SD, Gardner MJ, Wortman JR, Jordar VS, Maiti R, Kodira CD, Neafsey DE, Zeng Q, Hung CY, McMahan C, Muszewska A, Grynberg M, Mandel MA, Kellner EM, Barker BM, Galgiani JN, Orbach MJ, Kirkland TN, Cole GT, Henn MR, Birren BW, Taylor JW**. 2009. Comparative genomic analyses of the human fungal pathogens *Coccidioides* and their relatives. Genome Res **19:**1722–1731. http://dx.doi.org/10.1101/gr.087551.108.
40. **Kloepper TH, Huson DH**. 2008. Drawing explicit phylogenetic networks

and their integration into SplitsTree. BMC Evol Biol **8:**22. http://dx.doi.org/10.1186/1471-2148-8-22.

41. **Patterson N, Price AL, Reich D**. 2006. Population structure and eigenanalysis. PLoS Genet **2:**e190. http://dx.doi.org/10.1371/journal.pgen.0020190.
42. **Verity R, Nichols RA**. 2016. Estimating the number of subpopulations (K) in structured populations. Genetics **203:**1827–1839 http://dx.doi.org/10.1534/genetics.115.180992.
43. **Hudson RR, Slatkin M, Maddison WP**. 1992. Estimation of levels of gene flow from DNA sequence data. Genetics **132:**583–589.
44. **Teixeira HC, Calich VL, Singer-Vermes LM, D'Imperio-Lima MR, Russo M**. 1987. Experimental paracoccidioidomycosis: early immunosuppression occurs in susceptible mice after infection with pathogenic fungi. Braz J Med Biol Res **20:**587–589.
45. **Imai T, Sano A, Mikami Y, Watanabe K, Aoki FH, Branchini ML, Negroni R, Nishimura K, Miyaji M**. 2000. A new PCR primer for the identification of *Paracoccidioides brasiliensis* based on rRNA sequences coding the internal transcribed spacers (ITS) and 5 × 8S regions. Med Mycol **38:**323–326. http://dx.doi.org/10.1080/mmy.38.4.323.326.
46. **Hanna SA, Monteiro da Silva JL, Giannini MJ**. 2000. Adherence and intracellular parasitism of *Paracoccidioides brasiliensis* in Vero cells. Microbes Infect **2:**877–884. http://dx.doi.org/10.1016/S1286-4579(00)00390-7.
47. **Teixeira MdM, Theodoro RC, Oliveira FF, Machado GC, Hahn RC, Bagagli E, San-Blas G, Soares Felipe MS**. 2014. *Paracoccidioides lutzii* sp. nov.: biological and clinical implications. Med Mycol **52:**19–28. http://dx.doi.org/10.3109/13693786.2013.794311.
48. **Hebeler-Barbosa F, Montenegro MR, Bagagli E**. 2003. Virulence profiles of ten *Paracoccidioides brasiliensis* isolates obtained from armadillos (Dasypus novemcinctus). Med Mycol **41:**89–96. http://dx.doi.org/10.1080/mmy.41.2.89.96.
49. **Machado GC, Moris DV, Arantes TD, Silva LR, Theodoro RC, Mendes RP, Vicentini AP, Bagagli E**. 2013. Cryptic species of *Paracoccidioides brasiliensis*: impact on paracoccidioidomycosis immunodiagnosis. Mem Inst Oswaldo Cruz **108:**637–643. http://dx.doi.org/10.1590/0074-0276108052013016.
50. **De Albornoz MB**. 1971. Isolation of *Paracoccidioides brasiliensis* from rural soil in Venezuela. Sabouraudia **9:**248–253. http://dx.doi.org/10.1080/00362177185190491.
51. **Restrepo-Moreno A, Schneidau JD, Jr.** 1967. Nature of the skin-reactive principle in culture filtrates prepared from *Paracoccidioides brasiliensis*. J Bacteriol **93:**1741–1748.
52. **Bustamante-Simon B, McEwen JG, Tabares AM, Arango M, Restrepo-Moreno A**. 1985. Characteristics of the conidia produced by the mycelial form of *Paracoccidioides brasiliensis*. Sabouraudia **23:**407–414. http://dx.doi.org/10.1080/00362178585380601.
53. **Corredor GG, Peralta LA, Castaño JH, Zuluaga JS, Henao B, Arango M, Tabares AM, Matute DR, McEwen JG, Restrepo A**. 2005. The naked-tailed armadillo Cabassous centralis (Miller 1899): a new host to *Paracoccidioides brasiliensis*. Molecular identification of the isolate. Med Mycol **43:**275–280. http://dx.doi.org/10.1080/13693780412331271090.
54. **Morais FV, Barros TF, Fukada MK, Cisalpino PS, Puccia R**. 2000. Polymorphism in the gene coding for the immunodominant antigen gp43 from the pathogenic fungus *Paracoccidioides brasiliensis*. J Clin Microbiol **38:**3960–3966.
55. **Ferreira MS, Freitas LH, Lacaz CdS, del Negro GM, de Melo NT, Garcia NM, de Assis CM, Salebian A, Heins-Vaccari EM**. 1990. Isolation and characterization of a *Paracoccidioides brasiliensis* strain from a dog-food probably contaminated with soil in Uberlândia, Brazil. J Med Vet Mycol **28:**253–256. http://dx.doi.org/10.1080/02681219080000311.
56. **Carrero LL, Niño-Vega G, Teixeira MM, Carvalho MJ, Soares CM, Pereira M, Jesuino RS, McEwen JG, Mendoza L, Taylor JW, Felipe MS, San-Blas G**. 2008. New *Paracoccidioides brasiliensis* isolate reveals unexpected genomic variability in this human pathogen. Fungal Genet Biol **45:**605–612. http://dx.doi.org/10.1016/j.fgb.2008.02.002.

Chapter 9

Design and standardization of a conventional and a real time PCR assay based on novel species-specific genomic regions of the fungal pathogen *Histoplasma capsulatum*

Design and standardization of a conventional and a real time PCR assay based on novel species-specific genomic regions of the fungal pathogen *Histoplasma capsulatum*

Juan Gallo[1,2,3], Isaura Torres[1,3,4], Lavanya Rishishwar[5,6,7], Frederick Vannberg[5], I. King Jordan[5,6,7], Juan G. McEwen[1,8], Oliver K. Clay[1,9]

[1] Corporación para Investigaciones Biológicas (CIB), Cellular and Molecular Biology Unit, Medellín, Colombia

[2] Universidad del Rosario, Doctoral Program in Biomedical Sciences, Bogotá, Colombia

[3] Universidad CES, School of Medicine, GenomaCES, Medellin, Colombia

[4]Institución Universitaria Colegio Mayor de Antioquia (IUCMA), Faculty of Health Sciences, Medellín, Colombia

[5] School of Biological Sciences, Georgia Institute of Technology, Atlanta, Georgia

[6] Applied Bioinformatics Laboratory, Atlanta, GA 30332, USA

[7] PanAmerican Bioinformatics Institute, Cali, Valle del Cauca, Colombia

[8] Universidad de Antioquia, Faculty of Medicine, Medellín, Colombia

[9] Universidad del Rosario, School of Medicine and Health Sciences, Bogotá, Colombia

Abstract

Histoplasmosis is a systemic fungal disease caused by the fungal pathogen *Histoplasma capsulatum* that causes significant morbidity and mortality in HIV/AIDS patients and can also affect immunocompetent individuals. Although more recently PCR assays and antigen-detection assays have been developed, conventional diagnosis has largely relied on culture, which can take up to several weeks. One molecular assay used for clinical detection amplifies a 100 kDa protein coding gene that is present in *H. capsulatum* using nested PCR. Our aim was to provide a proof of principle for rationally designing and standardizing a conventional and a corresponding real-time PCR assay using *Histoplasma*-specific genomic regions, found by screening aligned assemblies, that would only require one PCR run, instead of the two runs of a typical nested PCR assay. Using all available genomic sequences for *H. capsulatum* and closely related fungi, we tried search strategies for identifying genomic regions that are specific to this species and could therefore be used to design conventional and real-time PCR assays. In the first approach we searched for *Histoplasma*-specific regions within current assemblies with high intraspecies conservation via multiple sequence alignments of contigs/scaffolds. Our second approach utilized gene annotations to search for protein-coding genes that are present in *H. capsulatum* but are not present in other species via orthologous gene clusters. A third approach involves a sliding window Blast algorithm to search for sequences in *H. capsulatum* assemblies of any desired size and/or spacing that are conserved within and only within *H. capsulatum* strains. The genomes of sequenced *H. capsulatum* strains were

compared *in silico* to other closely related fungal genomes within the Onygenales order, as well as to some fungal outgroups and human. We were able to design two PCR assays, whose primer sets can be used with either conventional or real-time PCR. Both primer sets resulted in 100% analytical specificity *in vitro* and following a positive detection of 62/62 *H. capsulatum* isolates using purified DNA. Positive control plasmid 10 fold serial dilutions were used to further test the analytical sensitivity of the primer sets. We believe that assays that can be derived via the whole-genome screening approach we propose could detect the presence of *H. capsulatum* cheaper and faster than currently used nested PCR assays. The proof of principle presented here establishes a new way to develop diagnostic assays, based on whole-genome scanning, that should be useful also for detecting other pathogenic fungi.

Introduction

Histoplasmosis is a systemic fungal disease caused by the inhalation of conidia of the dimorphic fungus *Histoplasma capsulatum*, with cases reported worldwide. Histoplasmosis is one of the most frequent fungal infections affecting individuals with HIV/AIDS. Histoplasmosis causes significant morbidity and mortality in HIV-infected individuals, particularly in those countries with limited access to rapid diagnostics or antiretroviral therapies, with up to 40% mortality reported (Kauffman 2009, Scheel et al. 2009, Baddley et al. 2008, Deepe 2010). The initial pulmonary manifestations of histoplasmosis are often misdiagnosed as a bacterial or viral pneumonia or classified as another disease, e.g., tuberculosis, with time and effort spent looking for non-fungal infectious etiologies. Diagnosis has traditionally relied largely on conventional blood cultures, which are positive in only approximately 50% of cases and may take up to 6 weeks, thus delaying diagnosis and initiation of therapy (Kauffman 2007). A major limitation of the serological test is that in the presence of an active infection, they are negative in up to 50% of immunosuppressed patients, especially those with AIDS (Deepe 2010). Given the public health need to provide reliable, rapid and affordable diagnosis of histoplasmosis, there is strong motivation to develop new and rapid diagnostic methods with high sensitivity and specificity, for example by employing molecular techniques (Scheel and Gómez 2014, Gómez 2011, Nacher et al. 2013).

Fungal infections can be diagnosed on the basis of morphological, immunological, clinical, and histopathological information. Of these procedures, histopathology can provide important diagnostic information in a relatively short

period of time, but identification is typically only tentative unless complemented by specialized techniques such as immunofluorescence, or when the etiological agent has distinct unique structures. Gomori methenamine or Grocott staining are useful for *H. capsulatum* yeast identification although via these or other methods it is easily confused with other yeasts such as *Candida* spp., *Pneumocystes jirovecii*, *Cryptococcus neoformans* or other infectious agents such as *Leishmania* spp. and *Toxoplasma gondii* (Brandt, Gómez and Warnock 2011, Deepe 2010).

DNA-based diagnosis has not yet been established as a routine diagnostic tool for histoplasmosis, but is used in some reference laboratories (De Pauw et al. 2008, Kauffman 2009). One molecular assay used for the detection of histoplasmosis is a nested PCR assay based on a gene coding for a 100 kDa protein that is considered specific to *H. capsulatum* (Bialek et al. 2002).

A new line of interest has arisen as a result of increased reliability of fungal genome sequencing pipelines and assembly algorithms during the last decade. Reference sequences that were utilized when some of the currently available assays were designed and validated have since been updated, and for some fungi this has resulted in dramatic quality improvements in sequence and annotation (Muñoz et al. 2014). With recent availability of finished and draft genome assemblies, as well as unassembled raw sequence reads, new target regions can be identified for developing more accurate molecular diagnostic assays. Such regions should have sequences that are both specific to the fungus of interest (i.e., not present in other organisms) and conserved in all strains of the fungus that might be present in clinical contexts.

In general, properly designed and clinically validated assays should provide the laboratory technician and clinician with a definitive diagnosis of the fungal pathogen via a PCR assay that is easy to implement. Low income countries can be affected in the diagnosis of fungal pathogens as a result of inadequate infrastructure and high costs of importation, high costs of healthcare, and/or limited budgets of local healthcare facilities/research centers (Harris 1998), and application of molecular assays in clinical settings should not be limited to highly specialized reference laboratories. Although molecular biology equipment is costly, a local research center or small clinic may be able to easily acquire a thermal cycler for conventional PCR, or if the budget permits, the equipment needed for real time PCR.

With possible economic or time constraints in mind, we aimed to design a diagnostic method using either conventional PCR or real time PCR that does not require sequencing, and reasoned that an easily deployable and affordable assay would be beneficial for the public health sector. Considering that the development of molecular assay methods for the diagnosis of fungal infections has sometimes been guided by anecdotal reports rather than rational principles, the present work explores a new route. We describe here a molecular diagnostic approach that should be capable of a high level of analytical specificity and analytical sensitivity for detecting the important endemic fungus *H. capsulatum*, involving a novel strategy for the rational design of PCR assays that can be used where accurate whole genome sequences of multiple strains of the target species and its close relatives are available and can be aligned.

Materials and Methods

Finding of unique regions of *H. capsulatum*

The methodology to find genome regions that are unique to *H. capsulatum* is based on bioinformatic strategies explored by our group. These strategies utilize various open source whole genome alignment algorithms, such as NUCmer and PROmer from the MUMMER package (Delcher et al. 2003) and Blast (Altschul et al. 1990). We focused on closely related fungal species to demonstrate uniqueness of regions to *H. capsulatum*, including *Paracoccidioides brasiliensis, P. lutzii, Blastomyces dermatitidis, Emmonsia crescens, E. parva*, as well as outgroups within the order Onygenales, non-fungal pathogens, and human. For *H. capsulatum*, all available genome sequences, including sequences from diverse strains, were obtained from publicly available databases. The fungal species listed above must be considered for possible cross-reaction due to phylogenetic relationship and homology. Once the sequences were obtained, three approaches were considered.

The first approach used involved searching for *Histoplasma*-specific regions within current assemblies. The contigs/scaffolds of the reference assembly of *H. capsulatum* were aligned to contigs/scaffolds of the other fungal species, in particular with the closely related pathogen *Paracoccidioides* spp. where we could vouch for high sequence quality (Muñoz et al. 2014). The alignments were done using NUCmer and PROmer from the MUMMER package and Blast. Contigs/scaffolds from *H. capsulatum* that did not align to any contigs/scaffolds in these other fungal species were selected as potentially *Histoplasma*-specific contigs/scaffolds. A second-pass

verification for analytical sensitivity was done by aligning all available strain sequences of *H. capsulatum* to ensure intraspecies inclusion. The subregions within the contigs/scaffolds that passed these filters were retained as potential sites for primer design. Once candidate primer sequences were designed, we used an additional filtering step to check that they were not similar to known sequences of other pathogens or to human genomic DNA, via Blast versus the non-redundant (nr) NCBI database.

The second approach utilized gene annotations to search for protein-coding genes that are present in *H. capsulatum* but are not present in other fungal species. We used gene annotations from the closely related fungal pathogens in the Ajellomycetaceae family, which includes *H. capsulatum*, as well as other fungal species from the order Onygenales. This was done by searching orthologous gene clusters obtained using OrthoMCL (Li et al. 2003). Gene clusters that were represented only in *H. capsulatum* and not in the other species were selected as potential genes for primer design. An additional step was performed via Blast versus non-redundant databases to verify that the gene or genes selected are not present in any known sequences of other organisms. If the gene or genes have small regions that overlap with other organisms, or regions that are not conserved among *H. capsulatum* strains, these regions are flagged as sites to avoid when designing primers, and the regions that remain are used for primer design.

A third approach involves a pairwise alignment algorithm based on a sliding window Blast to search for sequences in *H. capsulatum* assemblies of any desired size and/or spacing that are conserved within and only within *H. capsulatum* strains. We used window sizes of 500 and 250 base pairs (bp). This screening generates a cluster of genomic regions based on sequence homology. *H. capsulatum* sequence segments that met the criteria of no similarity in the assemblies used for the query species, and that also met the criteria of being conserved in the strains of *H. capsulatum*, were considered as candidates for primer design.

DNA strains

Genomic DNA from fungal strain cultures listed in Table 1 were obtained from several fungal pathogen DNA collections maintained at the Corporación para Investigaciones Biológicas (CIB, Medellín, Colombia) or the Centers for Disease Control and Prevention (CDC, Atlanta, GA). Genomic DNA for microbial strains used in analytical specificity tests were also obtained from these collections and are listed in Table 1. The relative concentrations of the genomic DNA were determined with a NanoDrop ND1000 apparatus (Thermo Scientific, Wilmington, DE).

Table 1. Species used for *in vitro* testing of analytical specificity

Species	Isolate	Species	Isolate
Coccidioides immitis	CDC B6037, CDC B10637, CDC B10757, CDC B10813	*Candida guillermondi*	CIB Collection
Blastomyces dermatitidis	CDC B3591, CDC 26117, CDC 180017, CDC 26116, CDC 26114	*Candida tropicalis*	CIB Collection
Aspergillus fumigatus	CDC ATCC 1022 T	*Candida parapsilopsis*	CIB Collection
Aspergillus versicolor	CDC NRRL238, CDC NRRL239	*Candida glabrata*	CIB Collection
Aspergillus flavus	NRRL485, IFI 03-0139	*Chrysosporium keratinophilum*	CDC B1959, CDC B1980, CDC B3644, CDC B2705
Aspergillus terreus	CDC IBT14590, CDC 141	*Cryptococcus neoformans*	B8915, B9029
Aspergillus niger	IFI03-0052, ATCC1015	*Cryptococcus gattii*	B8558, B9300
Neosartorya fischeri	B6256	*Uncinocarpus reesii*	CDC CBS 121.77
Neosartorya pseudofischeri	B5571, B5573	*Pneumocystis jirovecii*	CDC 163
Paracoccidioides brasiliensis	CIB Pb18, CIB Pb03	*Mycobacterium tuberculosis*	CIB Collection
Paracoccidioides lutzii	CIB Pb01	*Mycobacterium avium*	CIB Collection
Candida albicans	CIB Collection	*Mycobacterium chelonae*	CIB Collection
		Mycobacterium fortuitum	CIB Collection

Primer design

The primers were designed using the *Histoplasma*-unique regions selected via the bioinformatic analysis and subsequently analyzed using OligoAnalyzer 3.1 using quality control guidelines provided by Integrated DNA Technologies, Inc. (IDT). Some quality control aspects that were checked include: primers must be between 20-23 nucleotides in length, ideal GC content of primers is between 40-60%, melting temperatures (T_m) of primers should be between 42-65°C, primers in a pair should

have T_m's within 2°C of each other, and secondary structures (i.e., hairpins) within primers and potential dimerization between the primers should be avoided.

The primers designed were subjected to a BLAST search against the GenBank sequence database, to avoid cross-homology with other microorganisms or the human genome. The primers that were selected are listed in Table 2. These primers were designed for use in the conventional and real time PCR assays.

Table 2. Primer details

Gene	Primer	Length (bp)	Temperature	GC%	Amplicon size (bp)	Primer Sequence
CFP4	Forward	23	57.1 °C	52.2	800	5'-GTGACATCTGGAGCAGCTGTTGA-3'
	Reverse	23	57.1 °C	52.2		5'-TCAACTCGGGCGCTCTGTCAAAA-3'
PPK	Forward	22	54.8 °C	50	400	5'-CTGGTAAATAGGCGCTGTCTTG-3'
	Reverse	22	54.8 °C	50		5'-AGCTCAGCATCGACCGAATGAA-3'

Conventional PCR assay

The uniqueness of genomic regions for *H. capsulatum* were computationally predicted to be unique and experimentally assessed by conventional PCR. Thermocycler conditions were standardized via a temperature gradient of 54°C - 60°C using T100 Bio-Rad Thermal Cycler. The amplification products were analyzed on agarose gel and visualized with ethidium bromide under UV light. The PCR conditions selected were as follows: an initial step of 95°C for 10 min, followed by 45 cycles of 95°C for 30 seconds, 60°C for 30 seconds, and 72°C for 1 min.

Real time PCR assay

Real-time PCR (RT-qPCR) was performed using SYBR Green Real-Time PCR Master Mix, according to the manufacturer's instructions (Thermo Fisher Scientific Inc, USA) and using conditions standardized via conventional PCR. The CFX96 Real-Time PCR Detection System (Bio-Rad, Headquarters Hercules, California, USA) was used to carry out the amplification. PCR reactions were performed in 20-µl final volume containing qPCR master mix 2x. Each experiment was carried out in triplicate. The real time PCR conditions were as follows: an initial step of 95°C for 10 min, followed by 45 cycles of 95°C for 30 seconds, 60°C for 30 seconds, and 70°C for 1 min, with a melting curve at 60°C to 95°C incremented 0.5°C each 0.05.

Determining analytical specificity of primers *in vitro*

The analytical specificity of the primer sets was evaluated by conventional PCR and corresponding real time PCR using purified DNA from different isolates of *H. capsulatum*, as well as from collections of other related fungal pathogens and *Mycobacterium tuberculosis* maintained by the Corporación para Investigaciones Biológicas (CIB) and the Centers for Disease Control and Prevention (CDC, Atlanta, GA). The isolates were tested at a concentration of 1 ng/µl. For analytical sensitivity tests, all strains of *H. capsulatum* were tested for amplification with our chosen primers. A total of 62 *H. capsulatum* isolates were used, including isolates from North America (*H. capsulatum* CDC/Thon and *H. capsulatum* G217B), Central and South America (*H. capsulatum* CIB 1980, *H. capsulatum* G184B, *H. capsulatum* CDC

3670/CDC2787), and from Africa (*H. duboisii* CDC5822/CDC5823). Isolates of other pathogens used for the analytical specificity are listed in Table 1.

Generation of positive-control plasmids

Positive-control plasmids were constructed for *H. capsulatum* using the primers designed from species-specific regions or genes as described above. The amplified targets were cloned into the pCR 2.1 vector using the pCR 2.1 TOPO TA cloning kit (Invitrogen Corporation Carlsbad, CA) according to the manufacturer's instructions. The plasmid construct was then purified using the PureLink® Quick Plasmid Miniprep Kit (Thermo Fisher Scientific Inc**).** The ligation reactions were transformed in TOP10 chemically competent cells. Colonies were selected in LB plates containing 50 μg/ml kanamycin (two for each transformation; Babady et al. 2011).

Determining diagnostic sensitivity

The *in vitro* analytical sensitivity test of the assay was determined by testing a dilution series of *H. capsulatum* control plasmids. A 10-fold serial dilution of the plasmid was performed in TE buffer (10^5 copies/μl serially diluted to 10 copies/μl) and was used to construct the standard curve for limit of detection (LOD). Cycle threshold (CT) values for each dilution series were determined in triplicates in 3 different experiments consisting of 3 different tubes corresponding to the specific dilution of the curve on 3 different days.

Results

In silico assay design

We implemented bioinformatic approaches to search for regions within the genome of *H. capsulatum* that are unique to this species (a criterion needed for high specificity), as well as likely to be present in all strains of these species (a criterion needed for high sensitivity), and that could therefore be used for the identification of *H. capsulatum* via PCR assays. These allowed us to find several regions that were optimal for primer design.

First, we searched for *H. capsulatum* scaffolds that were not similar to genomic regions of other species. In *H. capsulatum* we found scaffolds, i.e., large contiguous regions (HcG186AR contig 2.315 / supercontig 2.50, length=22,664 nt; HcG217B contig 171, length=13,544 nt; HcH143 contig 2.108 in supercontig 2.1, length=8,250 nt; HcH88 scaffold 455, length= 1,881 nt,; https://www.broadinstitute.org/fungal-genome-initiative/histoplasma-genome-project) that met the criteria of not aligning to other closely related fungal species and of being present in all of the *H. capsulatum* strains queried. It was not possible to design primers within these regions because of the low yield of perfect alignment stretches longer than 20 base pairs that were conserved among the *H. capsulatum* strain sequences we used. This low yield may be partly due to sequencing and/or assembly errors, e.g.., single nucleotide errors (SNEs) masquerading as true SNPs, reflecting limited quality of the assemblies currently available for *H. capsulatum*. We therefore

did not include the regions found in this approach in the design of our assays reported here, although future improvements of the *H. capsulatum* sequences may render this approach feasible.

We next used, as a second approach, a search for entire protein-coding genes that are unique to *H. capsulatum*. This second approach allowed us to design two promising PCR primer pairs for the detection of the species. We were able to design primer pairs for two of the genic regions obtained, which were located in the two genes coding for culture filtrate protein 4 (CFP4; HCAG_06604; Holbrook et al. 2014) and a predicted protein kinase (PPK; HCBG_02218). The decision to focus on these two regions was based on their gene annotation having a gene function associated to it, as many of the other genes had no name or putative function. The CFP4 gene had been previously described as having potential as a diagnostic exoantigen (Holbrook et al., 2014). The genes were found using OrthoMCL matrix results. The matrix produced was queried for genes that did not have any likely orthologs in other sequenced species that we screened (see Materials and Methods). We used strict filtering for the selection of the genes. Only genes that had no similarity to potentially orthologous genes in the species queried were selected for further analysis. Genes that met this criterion were then queried by Blastn (via the NCBI web server) against the nr database in order to check the uniqueness of the genes to *H. capsulatum*.

The implementation of our second approach yielded many regions with sufficient length for primer design. Although in theory any of the regions discovered could have been used, we specifically used the two genomic regions from non-trivially annotated genes as they seemed most likely to play a biological role. i.e., to be robustly species-specific regions because of some true functional differentiator contrasting with closely related fungi. The primer details and PCR conditions are listed in Table 2. For real time PCR contexts, we did not consider designs involving a sequence-specific probe for a region between the two primers, as our aim was to keep costs low without sacrificing specificity. Instead, we focused on the use of DNA intercalating dyes such as SYBR Green. PCR conditions were the same for conventional and real-time PCR.

Using a third approach, which does not depend on gene annotations or the lengths of the scaffolds, we confirmed the uniqueness of the regions selected by the first two search strategies. The use of our algorithm for this third approach permitted the discovery of other unique genomic regions that were not found using the other strategies, but they were not analyzed further in this study.

Analytical sensitivity and specificity of primers using fungal genomic DNA *in vitro*

We tested our assay designs with 62 *H. capsulatum* strains from the fungal DNA collections of the CDC and the CIB. The outcomes were positive for all of the 62 strains tested, i.e., the primer pairs had 100% analytical sensitivity *in vitro*. Specificity

tests were done using the species mentioned above. The amplification was 100% specific to *H. capsulatum* as no amplification was observed in other species.

The PPK primer pair resulted in PCR products with homogenous amplicon sizes in all the tested DNAs, weighing approximately 400 bp. The CFP4 assay resulted in PCR products amplicons weighing approximately 800 bp, although a few strains had amplicons reaching up to 1000 bp when testing in positive control plasmids. Considering that smaller fragment sizes (i.e., lower molecular weights) might often be present in preparations of DNA from a clinical sample, we recommend primarily the PPK primer pair instead of the CFP4 primer pair.

Limit of detection (LOD) using positive plasmid control

In order to test for the analytical limit of detection of the assays, 10 fold serial dilutions of positive control plasmids were performed from 1 ng down to 1 fg for both CFP4 and PPK plasmid controls. The dilutions were amplified via real time PCR using standard conditions in triplicates for a sample processed on day 1, day 2 and day 3 for both plasmid controls. Detection of the positive control plasmid was observed up to 1 fg for all attempts for both CFP4 and PPK plasmids. The LOD figure (Fig 1) shows very homogenous amplification curves for the PPK plasmid; the LOD figure for the CFP4 had higher variability within the day replicates. Although both regions showed low limits of detection, we recommend using the PPK primer pair based on the higher reproducibility observed. LOD was also tested via conventional PCR. A positive band was observed up to 1 fg for PPK (Fig 2).

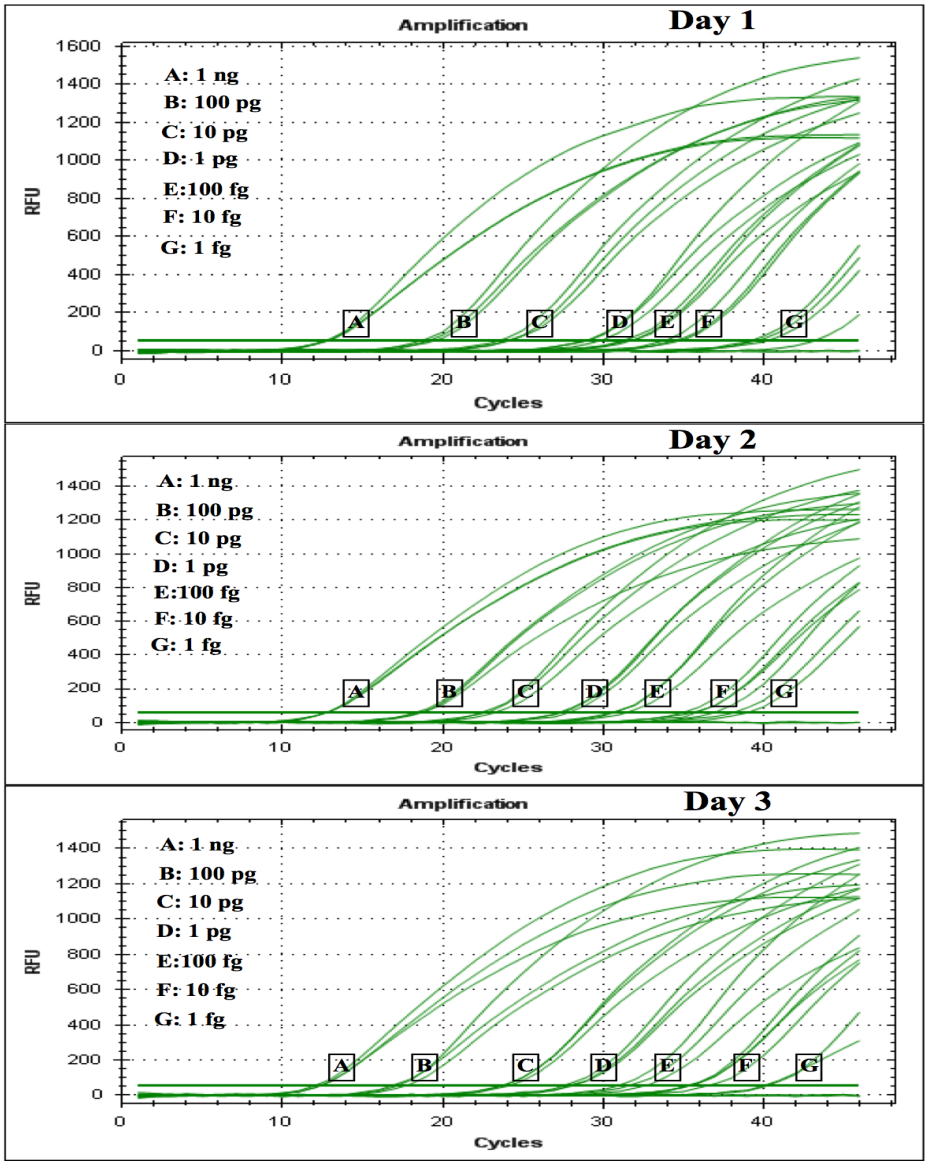


Figure 1. Quantification standard curve obtained by using serial dilution 1:10 of different concentrations of DNA construct from PPK genomic region.

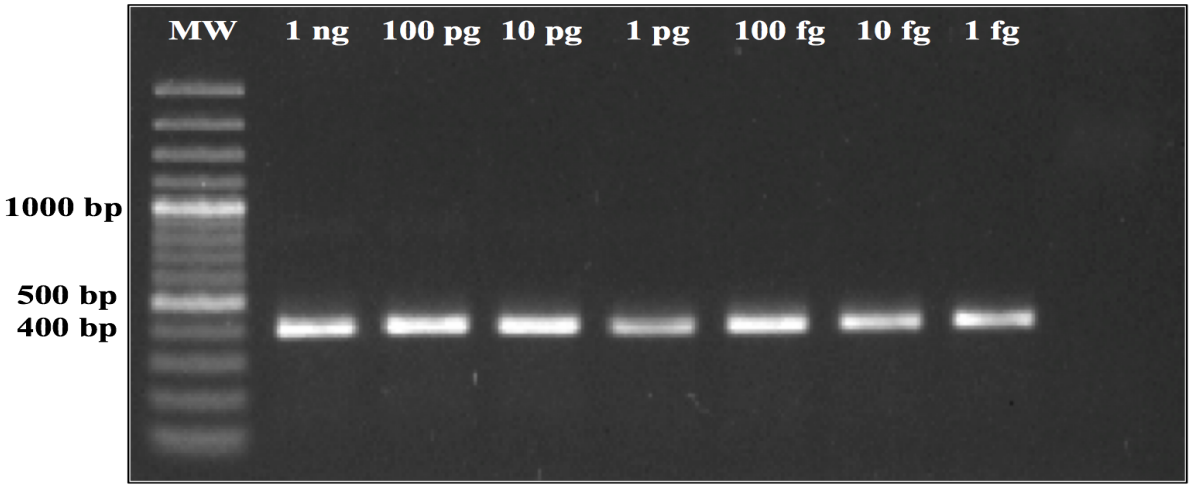


Figure 2. PPK control plasmid serial dilutions with no template control (NTC).

Discussion

We systematically designed and tested two primer pairs that could be used for conventional or real time PCR using similar conditions and the same primer sets. We specifically focused on a general design to work for both technologies due to the lack of specialized real time PCR equipment in standard molecular biology labs in some clinical facilities and countries. The regions in which the primer pairs were located correspond to two coding genes, culture filtrate protein 4 (*CFP4*) and a putative protein kinase, *PPK*. Due to the smaller PCR product size and more homogenous amplification curves during LOD testing, we selected the PPK primer pair for recommending of posterior assay testing. Although the CFP4 primer pair may be utilized, we believe that the PPK primer pair may be more efficient with the current design. The PPK primer pair was able to detect *in vitro* down to 1 fg of the control plasmid using both real time and conventional PCR designs. Other CFP4 primer pairs may allow reduction of the PCR product size and of the variability of the amplification curves.

We focused on the idea of creating primer sets that should work both for conventional and real time PCR, and that do not involve a nested PCR design. Nested PCR requires two amplification steps, which can increase complexity of the assay and laboratory processing times and costs as well as increasing the probability of contamination. Assays using primer pairs obtained via our screening methods should not need a secondary amplification step, and in the designs we tested the limits of

detection were low enough to be comparable to a previous nested PCR assay of Bialek et al. targeting a region of a gene for a 100kDa protein.

Since the PCR assay for the detection of *H. capsulatum* amplifying the 100kDa protein (Bialek 2002) has been used with success, we searched for the sequence of this gene in our OrthoMCL results. The results did not include this gene as a unique gene for *H. capsulatum*. Blast results showed that the sequence for the 100kDa gene as well as the primers used in the assay share homology with other closely related fungal species such as *B. dermatitidis* and *P. lutzii*. The 100 kDa assay may therefore succeed because of sufficient sequence differences in a gene that is however present in a number of related species, and not because of a strict absence of homologs of that gene in genera outside *Histoplasma*.

Since the PPK assay that we suggest here for further clinical testing was designed in a region coding for a protein kinase, we believe that this region should maintain evolutionary stability throughout strains of *H. capsulatum*. We confirmed in vitro analytical sensitivity of the primer pair via testing of DNA from 62 *H. capsulatum* strains of different lineages representing a large geographical diversity, and we observed amplification in all 62 isolates tested.

Experimental validations of PCR-based diagnostic tests typically involve three phases. The first phase is *in vitro* testing that the assay works well for pure fungal DNA of the target species, without any other DNA present (*in vitro* phylogenomic sensitivity testing), and that it does not amplify DNA of other species, again tested in isolation with only one species present at a time (*in vitro* phylogenomic specificity

testing). The second or intermediate phase involves contexts of the DNA that are not yet full clinical laboratory scenarios such as blood samples of infected patients, but that go beyond the first *in vitro* scenario; such contexts could include, for example, samples of human blood that are experimentally 'spiked' with fungal samples in the lab, i.e., whole-blood infection models (which are used also in other research contexts, e.g., Lehnert et al 2015; Hünniger et al, 2014; Dix et al. 2015). The third and final phase uses blood, tissues, body fluids or other samples of patients with demonstrated infection by the target species from clinical cohorts, together with corresponding negative control samples from individuals who are known not to be infected with the target pathogen but possibly with other pathogens.

Clinical samples were not available for testing the sensitivity or specificity of our assays in clinical contexts. To include a full clinical assay test in the present work would have led outside its aim and scope. Instead, we have presented here a proof of principle of the likely efficacy and practical utility of a conceptually simple new method for designing primer pairs for PCR assays to detect pathogenic target species in regions of the fungal tree that are represented by high-quality (`clinical grade') whole-genome sequences. In order to assess the validity of the results reported here for use in clinical laboratories, assays based on the primers pairs would need to be tested in clinical samples. In the future, we hope to collaborate in a cohort study and test our assays on clinical samples.

We were able to identify two promising target regions, representing two likely functional genes and their protein products, that were unique to *Histoplasma*, that were present in all *Histoplasma* isolates for which sequences or samples were available to us (i.e., with no false negatives observed *in silico* or *in vitro*), and that were absent in all other species for which we had access to sequences or samples (i.e., with no false positives observed *in silico* or *in vitro*). We also found a region containing an entire contig of a *H. capsulatum* reference genome sequence that was present and roughly conserved in this species and absent from other species, but the ambiguity between true variability and base-level errors that still is likely to exist in the current *Histoplasma* genome sequences suggested that assay designs in this region should await updates of these sequences. We are currently developing a more general, assembly- and annotation-independent method to screen for species-specific regions; a prototype implementation of this method again reported the two genic regions studied here, as well as other promising regions.

Rational, genome-sequence based approaches such as we advocate here, and that to our knowledge have so far been systematically explored only for viruses or bacteria (Slezak et al., 2003; Phillippy et al., 2007, 2009) but not for eukaryotes until this study, will allow researchers looking for target regions to exhaustively scan high-quality genome alignments for candidate primer set designs meeting a given set of criteria, and then rank them according to the expectation of their efficacy and robustness, e.g., when samples are slightly degraded and have low molecular weight (fragmented) fungal DNA. Using such an approach, we have identified two primer

pairs that could be used to detect the presence of *H. capsulatum*. To the best of our knowledge, one of the two primer pairs, from the PPK gene, should offer a quick and affordable method for detecting *H. capsulatum* using PCR technology. The approach we explored here should benefit substantially from genome sequence improvements that are underway. As an example, a recent update of three reference genomes of *Paracoccidioides*, a pathogen closely related to *Histoplasma*, using NGS resequencing and correction employing a Pilon pipeline, showed that dramatic improvement of base-level assembly and annotation quality is possible (Muñoz 2014, Walker et al. 2014). Future improvement of the *Histoplasma* and other fungal pathogen genome sequences along similar lines should allow the identification of additional molecular assay target sequences enabling more specific or more sensitive molecular assays for the detection of pathogens.

Acknowledgements

We thank the Mycotic Diseases Branch of the Centers for Disease Control and Prevention (CDC), Atlanta, GA for hospitality and sharing of laboratory facilities and materials, for the use of fungal strain collections, and for giving J.G. and I.T. the opportunity to complete the experimental work presented here. We especially thank Dr Anastasia Litvintseva for expert guidance and assistance, and for critical reading and helpful comments on the manuscript. J.G. acknowledges the partial funding of this work by Fulbright Colombia and by the Universidad del Rosario. The work described here was co-funded by Colciencias, Colombia via grant 1222-569-34875, *A gene atlas for human pathogenic fungi*.

References

1. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990). Basic local alignment search tool. J Mol Biol. 215:403-410.

2. Baddley JW, Sankara IR, Rodriquez JM, Pappas PG, Many WJ, Jr (2008). Histoplasmosis in HIV-infected patients in a southern regional medical center: poor prognosis in the era of highly active antiretroviral therapy. Diagn Microbiol Infect Dis. 62:151-156.

3. Bialek R, Feucht A, Aepinus C, Just-Nubling G, Robertson VJ, Knobloch J, et al. (2002). Evaluation of two nested PCR assays for detection of *Histoplasma*

capsulatum DNA in human tissue. J Clin Microbiol. 40:1644-1647.

4. Brandt ME, Gómez BL, Warnock D (2011). *Histoplasma, Blastomyces, Coccidioides*, and other dimorphic fungi causing systemic mycoses. In: Versalovic J and Warnock D. (eds). Manual of Clinical Microbiology, 10th ed. Vol. 2, Section VI, Chapter 120. ASM Press, Washington DC. pp.1902-1918.

5. de Pauw BE, Picazo JJ (2008). Present situation in the treatment of invasive fungal infection. Int J Antimicrob Agents 32 Suppl 2:S167-171.

6. Deepe, G. (2010) *Histoplasma capsulatum*. In: Mandell, 7th edition, Philadelphia, PA, principles and practices of infectious diseases edited by G.L. Mandell, J.E. Bennett and R. Dolin. Chapter 264, pp. 3315 ff.

7. Delcher AL, Salzberg SL, Phillippy AM (2003). Using MUMmer to identify similar regions in large sequence sets. Curr Protoc Bioinformatics,. Chapter 10, Unit 10.3. doi: 10.1002/0471250953.bi1003s00.

8. Dix A, Hunniger K, Weber M, Guthke R, Kurzai O, Linde J (2015). Biomarker-based classification of bacterial and fungal whole-blood infections in a genome-wide expression study. Front Microbiol. 6:171.

9. Gomez BL (2011). Histoplasmosis: Epidemiology in Latin America. Curr Fungal Infect. Rep 5:199-205.

10. Harris, E (1998). A Low Cost Approach to PCR: Appropriate Technology Transfer of Biomolecular Techniques, ed. Kadir, N., Oxford Univ. Press, New York..

11. Holbrook ED, Kemski MM, Richer SM, Wheat LJ, Rappleye CA (2014). Glycosylation and immunoreactivity of the *Histoplasma capsulatum* Cfp4 yeast-phase exoantigen. Infect. Immun. 82:4414-4425.

12. Hunniger K, Lehnert T, Bieber K, Martin R, Figge MT, Kurzai O (2014). A virtual infection model quantifies innate effector mechanisms and *Candida albicans* immune escape in human blood. PLoS Comput Biol. 10(2):e1003479.

13. Kauffman CA (2009). Histoplasmosis. Clin Chest Med. 30:217-225.

14. Lehnert T, Timme S, Pollmacher J, Hunniger K, Kurzai O, Figge MT (2015). Bottom-up modeling approach for the quantitative estimation of parameters in pathogen-host interactions. Front Microbiol. 6:608.

15. Li L, Stoeckert CJ, Jr., Roos DS (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13:2178-89.

16. Munoz JF, Gallo JE, Misas E, Priest M, Imamovic A, Young S, et al. (2014). Genome update of the dimorphic human pathogenic fungi causing paracoccidioidomycosis. PLoS Negl Trop Dis. 8(12):e3348.

17. Nacher M, Adenis A, Aznar C, Blanchet D, Vantilcke V, Demar M, et al (2014). How many have died from undiagnosed human immunodeficiency virus-associated histoplasmosis, a treatable disease? Time to act. Am J Trop Med Hyg. 90:193-194.

18. Phillippy AM, Mason JA, Ayanbule K, Sommer DD, Taviani E, Huq A, Colwell RR, Knight IT, Salzberg SL (2007). Comprehensive DNA signature discovery and validation. PLoS Comput Biol. 2007 3(5):e98.

19. Phillippy AM, Ayanbule K, Edwards NJ, Salzberg SL (2009). Insignia: a DNA signature search web server for diagnostic assay development. Nucleic Acids Res. 37:W229-234.

20. Scheel CM and Gómez BL (2014). Diagnostic methods for histoplasmosis: Focus on endemic countries with variable infrastructure levels. Curr Trop Med Rep. 1:129–137.

21. Scheel CM, Samayoa B, Herrera A, Lindsley MD, Benjamin L, Reed Y, et al. (2009). Development and evaluation of an enzyme-linked immunosorbent assay to detect *Histoplasma capsulatum* antigenuria in immunocompromised patients. Clin Vaccine Immunol. 16:852-858.

22. Slezak T, Kuczmarski T, Ott L, Torres C, Medeiros D, Smith J, Truitt B, Mulakken N, Lam M, Vitalis E, Zemla A, Zhou CE, Gardner S (2003). Comparative genomics tools applied to bioterrorism defence. Brief Bioinform. 4:133-149.

23. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 9(11):e112963.

Chapter 9 Appendix

Perl program implementing a pairwise alignment strategy for selecting unique genomic regions with diagnostic potential

Perl program implementing a pairwise alignment strategy for selecting unique genomic regions with diagnostic potential

Description of the algorithm

This program is used in the first step of a multi-step process for finding genomic regions of an organism that have diagnostic potential. In the first step, this organism's genome, the *query* genome, is compared with other genomes, the background or *subject* genomes. The aim of this first stage is to identify regions in the query genome that are not similar to any regions of any of the background or subject genomes. Candidate regions obtained in this first step must then afterwards be compared (a) against other strains of the query species or genus to ensure that the regions are also present there, and (b) possibly against additional other, more distant background species that could be present in a clinical context, including human, to ensure that the regions have no similarity with any parts of those genomes.

The user must first determine which genomic sequences will be included in the comparison. The sequences must be in Fasta format. The sequences should be in the same directory, as a preprocessing step should take place prior to the actual comparison.

In order to account for possible N's within the assemblies, all of the sequence scaffolds are initially queried for stretches of N's greater than 15, which are then removed. The program does not splice by joining the two flanking nucleotide regions. Instead, each segment of well-characterized nucleotides is saved as a new temporary

Fasta file. For example, if scaffold 1 has a length of 1000 bp, and a continuous stretch of 20 N's is found beginning at nucleotide position 450, the 20 N's will be spliced out of the sequence and two sub-scaffolds will be created; scaffold1.1, which will contain the nucleotides 1-449, and scaffold1.2, containing nucleotides 471-1000. This splicing out of long stretches of N's will prevent false positive mismatches in the sequence comparison steps downstream. The splicing process also renames the resulting Fasta files with generic names for the reading/interpretation of the fragments. A legend is created in a mapping file, which contains the information specifying from which original scaffolds the fragments were derived. This process is done for all sequences in the directory that will be used in the analysis by running the script *Break_and_rename.pl*.

The comparative similarity analysis (*do_pairwise.pl*) is based on a sliding window alignment using Blastn. A local *blast* version must be previously installed on the system and placed in the path. The fragments that were generated in the break and rename step are compared in an all versus all fashion. The default window size of the nucleotides to compare at a time is 1000 base pairs. The default moving-window step size is half of the window size, and can be modified by the user in the parameters.

The resulting *blast* output is in tabular format 6. This is key for various reasons. The alignment and similarity metrics provide us with needed information summarizing the comparison of the sequences analyzed. A matrix is created for each of the genomes used. Here one may extract the corresponding regions that are unique to the genome of interest. To do so, one can use basic Unix commands in order to

extract the sequence fragments that have no appreciable similarity (i.e., below the reporting threshold) to the other genomes in the comparative analysis. After applying those Unix commands, the user will have a complete set of sequences or coordinates of regions that are unique to the query genome, and not present or similar to any region in any of the background genomes.

Code

Break_and_rename.pl

```
#!/usr/bin/perl -w
use strict;
my @fasta = <*.fasta>;

open OUT, ">mapping.txt" or die "Cannot create output mapping file mapping.txt: $!\n";
print OUT "Species\tNewAccession\tOldAccession\n";
for(my $i=0; $i < @fasta; $i++){
	my $file = $fasta[$i];
	open FILE, "<$file" or die "Cannot open input file $file: $!\n";
	open NEW, ">temp.txt" or die "Cannot create output file temp.txt: $!\n";
	my $j=1;
	$file =~ s/\.fasta//;
	while(my $desc = <FILE>){
		my $seq = <FILE>;
		chomp $desc;
		chomp $seq;
		my @seqs = split(/N{15,}/, $seq);
		for(my $x = 0; $x < @seqs; $x++){
			print OUT "$file\tFUNG".$i."_".$j."-$x\t$desc\n";
			$seqs[$x] =~ s/\s+//g;
			print NEW ">FUNG".$i."_".$j."-$x\n$seqs[$x]\n";
		}
		$j++;
	}
	close FILE;
	close NEW;
	`mv temp.txt $file.fasta`;
}

close OUT;
```

Do_pairwise.pl

```
#!/usr/bin/perl -w
use strict;
use Getopt::Long;
```

```
my @fasta = <*.fasta>;

my $win  = 5000;
my $step = int($win/2);
my $createDb = 0;

my $usage = "$0 [-w=window size. Default = 5000]
                [-s=step size. Default = 2500]
                [-d (if database creation is to be skipped)]";


GetOptions ("d+"  => \$createDb,
            "w=i" => \$win,
            "s=i" => \$step);


if(! $createDb ){
    #Create database
    for(my $i=0; $i < @fasta; $i++){
        print STDERR "Creating blast database for $fasta[$i]\n";
        `makeblastdb -in $fasta[$i] -dbtype nucl -input_type fasta`;
    }
    print STDERR "Created blast databases\n";
}

for(my $i=0; $i < @fasta; $i++){ # Query Genome
    my %hits;
    my @windows;

    for(my $j=0; $j < @fasta; $j++){ # Subject Genome
        next if ($i == $j);
        print STDERR "Started comparing $fasta[$i] against $fasta[$j]\n";
        open IN, "<$fasta[$i]" or die "Cannot open input file $fasta[$i]: $!\n";
        my $desc = "";
        while(<IN>){
            if($_ =~ /^>/){
                $desc = $_;
                chomp $desc;
                next;
            }
            my $seq = $_;
            chomp $seq;
            my $len = length($seq);

            my $x = 0;
            for($x=0; $x < ($len-$win); $x = $x+$step){
                my $string = substr($seq, $x, $win);
                my $coord = "$x-".($x+$win);
                push(@windows, $desc.":$coord");
                if($string =~ /^N+$/){
                    $hits{$fasta[$j]}{$desc.":$coord"}{"cov"} = "NA";
                    $hits{$fasta[$j]}{$desc.":$coord"}{"pid"} = "NA";
```

```
                        } else{
                            #print STDERR "echo '$string' | blastn -db $fasta[$j] -query - -
max_target_seqs 1 -outfmt 6 -evalue 0.0005 | head -1 | awk 'BEGIN{OFS=\"\\t\"}{print \$4,\$3}'\n";
                            my ($cov, $pid) = split(/\s+/, `echo '$string' | blastn -db $fasta[$j] -query - -
max_target_seqs 1 -outfmt 6 -evalue 0.0005 | head -1 | awk 'BEGIN{OFS="\\t"}{print \$4,\$3}'`);
                            if(! defined $cov){
                                $cov = 0;
                                $pid = 0;
                            }
                            $cov /= 5;
                            $hits{$fasta[$j]}{$desc.":$coord"}{"cov"} = $cov;
                            $hits{$fasta[$j]}{$desc.":$coord"}{"pid"} = $pid;
                        }
                    }
                    my $string = substr($seq, $x, ($len-$x+1));
                    my $coord = "$x-$len";
                    push(@windows, $desc.":$coord");
                    my ($cov, $pid) = split(/\s+/, `echo '$string' | blastn -db $fasta[$j] -query - -
max_target_seqs 1 -outfmt 6 -evalue 0.0005 | head -1 | awk 'BEGIN{OFS="\\t"}{print \$4,\$3}'`);
                    if(! defined $cov){
                        $cov = 0;
                        $pid = 0;
                    }
                    $cov /= 5;
                    $hits{$fasta[$j]}{$desc.":$coord"}{"cov"} = $cov;
                    $hits{$fasta[$j]}{$desc.":$coord"}{"pid"} = $pid;
                }
                close IN;
        }

        open OUT, ">$fasta[$i]-blastHits.tsv" or die "Cannot create output file $fasta[$i]-blastHits.tsv:
$!\n";
        print OUT "Window";
        for(my $j=0; $j < @fasta; $j++){
            next if($i == $j);
            my $file = $fasta[$j];
            $file =~ s/\.fasta//;
            print OUT "\t$file-Coverage\t$file-PercIdent";
        }
        print OUT "\n";


        for(my $w = 0; $w < @windows; $w++){
            print OUT "$windows[$w]";
            for(my $j=0; $j < @fasta; $j++){ # Subject file
                print OUT
"\t".$hits{$fasta[$j]}{$windows[$w]}{"cov"}."\t".$hits{$fasta[$j]}{$windows[$w]}{"pid"};
            }
            print OUT "\n";
        }
        close OUT;}
```

Chapter 10

Design and analytical validation of novel primer pairs for the detection of *Paracoccidioides* spp.

Design and analytical validation of novel primer pairs for the detection of *Paracoccidioides* spp.

Juan Gallo[1,2,3], Isaura Torres[1,3,4], Lavanya Rishishwar[5,6,7], Frederick Vannberg[5], I. King Jordan[5,6,7], Juan G. McEwen[1,8], Oliver K. Clay[1,9]

[1] Corporación para Investigaciones Biológicas (CIB), Cellular and Molecular Biology Unit, Medellín, Colombia

[2] Universidad del Rosario, Doctoral Program in Biomedical Sciences, Bogotá, Colombia

[3] Universidad CES, School of Medicine, GenomaCES, Medellin, Colombia

[4]Institución Universitaria Colegio Mayor de Antioquia (IUCMA), Faculty of Health Sciences, Medellín, Colombia

[5] School of Biological Sciences, Georgia Institute of Technology, Atlanta, Georgia

[6] Applied Bioinformatics Laboratory, Atlanta, GA 30332, USA

[7] PanAmerican Bioinformatics Institute, Cali, Valle del Cauca, Colombia

[8] Universidad de Antioquia, Faculty of Medicine, Medellín, Colombia

[9] Universidad del Rosario, School of Medicine and Health Sciences, Bogotá, Colombia

Abstract

Paracoccidioidomycosis (PCM) is a human systemic granulomatous mycosis caused by thermally dimorphic fungi from the *Paracoccidioides* genus. The disease is prevalent in Latin America and triggers a serious clinical condition. Consequently, rapid diagnosis and treatment are crucial to prevent progression of the disease, which can result in death in the absence of therapy. Currently, there are several methods for the detection of *Paracoccidioides brasiliensis*. However, many of these tests still present challenges in terms of cost, accessibility, reproducibility and efficiency. Via a genome-wide scanning we identified candidate target sequences, and designed and analytically validated 3 genus-specific conventional and real time PCR primer sets with potential for the rapid and economical detection of *Paracocidioides* species. All three primer sets show promising results as 100% analytical specificity was obtained with as well as amplification of isolated DNA from 92 *Paracoccidioides* spp. strains. The lowest level of detection was that of 1 fg using a unique gene shared amongst the species of the genus. using These primer sets show promising evidence for use in detecting *Paracoccidioides* spp. in clinical or environmental samples.

Introduction

The fungal genus *Paracoccidioides* is a thermally dimorphic Ascomycota that causes paracoccidioidomycosis (PCM), a chronic, systemic and progressive mycosis endemic in Latin America, where up to ten million people can be infected (Brummer et al., 1993, Restrepo and Tobon., 2010). The *Paracoccidioides* genus contains two species, *Paracoccidioides brasiliensis* and *Paracoccidioides lutzii*. The infection is acquired through inhalation of conidia produced by the mycelial phase of the fungus, which at body temperatures grows as multibudding yeast (McEwen et al., 1987). PCM must be treated promptly, as *Paracoccidioides* infections may be fatal. Like most other invasive mycoses, this fungus is difficult to eliminate from the human body, and prompt detection and treatment are usually decisive to avoid long-term deleterious injuries and consequences (Greer and Restrepo, 1977). Currently, isolation of *Paracoccidioides* spp. in culture constitutes evidence of active disease, but this test is positive in only 85% of cases. Growth in culture is required, taking between 20 - 30 days, followed by microscopic examination and/or immunologic tests such as immunodiffusion and complement fixation to identify the microorganism (Restrepo and Tobon., 2010).

During the last decades, molecular assays for PCM diagnosis have been developed to provide alternative strategies for the diagnosis of PCM that can replace the often time-consuming conventional methods. However, many of these tests still present challenges in terms of cost, accessibility and efficiency. DNA probes have been developed for the rapid identification of *P. brasiliensis* in mycelial and yeast

cultures as well as in biopsies from oral cavity lesions of PCM patients and guinea pigs inoculated with the fungus (Sandhu et al., 1997; De Brito et al., 1999). Furthermore, several molecular biological tools have been applied to detect specific *P. brasiliensis* DNA sequences, such as polymerase chain reaction (PCR), nested PCR, PCR- enzyme immunoassay, real time PCR (RT-PCR) and loop-mediated isothermal amplification (LAMP) (Bialek et al., 2000; Gomes et al., 2000; Motoyama et al., 2000; Lindsley et al., 2001; Semighini et al., 2002; San-Blas et al., 2005). In the latter studies, various DNA templates were used to detect *P. brasiliensis* from distinct sources such as: DNA from clinical and environmental isolates; sputum and cerebrospinal fluid from PCM patients and tissues samples from mice infected with conidia by intranasal inoculation The Pb*GP43* and ribosomal DNA genes have been the most commonly targeted *P. brasiliensis* sequences for the detection of the fungus in clinical samples. On the other hand, the Pb*P27* gene has been applied used to screen for *P. brasiliensis* in clinical and environmental isolates, artificially contaminated soils and in tissues of armadillos naturally infected with the fungus, although it has not been used for the evaluation of human tissue samples (Diez et al., 1999, Rocha-Silva et al 2016). Moreover, microsatellites in the genome of *P. brasiliensis* have been detected and proposed as genetically associated elements with the potential to discriminate clinical isolates in accordance with virulence profiles (Nascimento et al., 2004).

Several PCR assays have been reported previously for the detection of *Paracocidioides brasiliensis*, but these studies used the detection of 18S, 5.8S, 28S and their spacer regions ITS1 and ITS2 as well as Pb*GP43* and Pb*P27* genes. The

Pb*GP43* and Pb*P27* genes share homology with other phylogenetically related species and therefore could cause false positive results (Teles and Martins., 2011). There are difficulties in cultivating the *Paracoccidioides* species, and molecular assays have shown success in the detection of *P. brasiliensis* even when microscopic observation and antibody detection fail (Teles and Martins., 2011). The previous molecular detection designs were optimized for detecting *P. brasiliensis*, but not for *P. lutzii*. Our group recently published the genome updates of *P. brasiliensis* and *P. lutzii* (Muñoz et al., 2014). The availability of curated genome assemblies allowed us to search for novel genomic regions that are unique to the genus *Paracoccidioides*, yet absent in other species, permitting the design of highly specific primer pairs for the molecular detection of these pathogens. We present 3 primer pairs that we designed on the basis of such a search, and found that they are 100 % specific to the genus using conventional and/or real time PCR technologies.

Materials and methods

The methodology used in this study was very similar to the one described in chapter 9. For simplicity of this thesis, we will only describe the major differences in methodology with respect to chapter 9. The methodology for finding the unique regions for *Paracoccidioides* spp. used the 3 strategies mentioned described in chapter 9, with the only modification being the change of query genomes to *P. brasiliensis* and *P. lutzii* and including the *H. capsulatum* genomes in the subject genomes list. The DNA strains used for the analytical validation are listed in Table 1 of chapter 9, with

the only modification being the inclusion of the *H. capsulatum* strains to the table. We analyzed 92 *Paracoccidioides* spp. isolates from a variety of sources (geographic, clinical, environmental, and phylogenetic species). Genomic DNA from fungal strain cultures listed in Table 1 were obtained from several fungal pathogen DNA collections maintained at the Corporación para Investigaciones Biológicas (CIB; Medellin, Colombia) or the Centers for Disease Control and Prevention (CDC; Atlanta, GA). Positive-control plasmids were constructed from the selected unique regions of Pb18 isolates using the different primer pairs, taking into account the conserved regions between the three reference genomes (Pb01, Pb03 and Pb18). The genome sequences of *Paracoccidioides* are available in GenBank, with accession numbers *P. lutzii* Pb01 (ABKH00000000), *P. brasiliensis* Pb03 (ABHV00000000), and *P. brasiliensis* Pb18 (ABKI00000000)) (Dejardins et al., 2011; Muñoz et al., 2014).

Table 1. *Paracoccidioides* spp. isolates used in the study

Isolates	Country of origin-Source	Species
T1F1	Brazil, environmental	*P. brasiliensis*
T3B6	Brazil, environmental	*P. brasiliensis*
T4B14	Brazil, environmental	*P. brasiliensis*
T7F6	Brazil, enviromental	*P. brasiliensis*
T8B1	Brazil, environmental	*P. brasiliensis*
T9B1	Brazil, environmental	*P. brasiliensis*
T10B1	Brazil, environmental	*P. brasiliensis*
T5LN1	Brazil, environmental	*P. brasiliensis*
T13LN1	Brazil, environmental	*P. brasiliensis*
T15LN1	Brazil, environmental	*P. brasiliensis*
IBIÁ	Brazil, environmental	*P. brasiliensis*
Uberlandia	Brazil, environmental	*P. brasiliensis*
BT60	Brazil, clinical	*P. brasiliensis*

BT84	Brazil, clinical	*P. brasiliensis*
14 - 121A	Brazil, clinical	*P. brasiliensis*
Pb18	Brazil, clinical	*P. brasiliensis*
B339	Brazil, clinical	*P. brasiliensis*
P149	Colombia, clinical	*P. brasiliensis*
P159	Colombia, clinical	*P. brasiliensis*
P163	Colombia, clinical	*P. brasiliensis*
ATCC60855	Colombia, clinical	*P. brasiliensis*
P141	Colombia, clinical	*P. brasiliensis*
P196	Colombia, clinical	*P. brasiliensis*
P204	Colombia, clinical	*P. brasiliensis*
P202	Colombia, clinical	*P. brasiliensis*
P68	Colombia, clinical	*P. brasiliensis*
P72	Colombia, clinical	*P. brasiliensis*
P46	Colombia, clinical	*P. brasiliensis*
P161	Colombia, clinical	*P. brasiliensis*
ATCC76533	Colombia, clinical	*P. brasiliensis*
H0054-1-45	Colombia, clinical	*P. brasiliensis*
H0054-1-47	Colombia, clinical	*P. brasiliensis*
P206	Colombia, clinical	*P. brasiliensis*
P151	Colombia, clinical	*P. brasiliensis*
CIB44197	Colombia, environmental	*P. brasiliensis*
CIB40392	Colombia, environmental	*P. brasiliensis*
Pb300	Venezuela, environmental	*P. brasiliensis*
U1	Antarticta, environmental	*P. brasiliensis*
A1	Argentina, clinical	*P. brasiliensis*
A2	Argentina, clinical	*P. brasiliensis*
A3	Argentina, clinical	*P. brasiliensis*
A4	Argentina, clinical	*P. brasiliensis*
A5	Argentina, clinical	*P. brasiliensis*
A6	Argentina, clinical	*P. brasiliensis*
A7	Argentina, clinical	*P. brasiliensis*
A8	Argentina, clinical	*P. brasiliensis*
P1	Paraguay, clinical	*P. brasiliensis*
P2	Paraguay, clinical	*P. brasiliensis*
P164	Colombia, clinical	*P. brasiliensis*
P165	Colombia, clinical	*P. brasiliensis*
P166	Colombia, clinical	*P. brasiliensis*
P168	Colombia, clinical	*P. brasiliensis*
P169	Colombia, clinical	*P. brasiliensis*
Pb 73	Colombia, clinical	*P. brasiliensis*
Pb381	Venezuela, clinical	*P. brasiliensis*
Pb 304	Venezuela ,clinical	*P. brasiliensis*

15632	Brazil, clinical	*P. brasiliensis*
4154	Brazil, clinical	*P. brasiliensis*
15601	Brazil, clinical	*P. brasiliensis*
Pb 89	Brazil, clinical	*P. brasiliensis*
Pb76	Brazil, clinical	*P. brasiliensis*
Pb 78	Brazil, clinical	*P. brasiliensis*
P174	Colombia, clinical	*P. brasiliensis*
P175	Colombia, clinical	*P. brasiliensis*
P178	Colombia, clinical	*P. brasiliensis*
Mg4	Brazil, clinical	*P. brasiliensis*
924	Brazil, clinical	*P. brasiliensis*
SS	Brazil, clinical	*P. brasiliensis*
1017	Brazil, clinical	*P. brasiliensis*
D01	Brazil, clinical	*P. brasiliensis*
D02	Brazil, clinical	*P. brasiliensis*
PbS1	Brazil, clinical	*P. brasiliensis*
Pb01	Brazil, clinical	*P. brasiliensis*
Pb2	Venezuela, clinical	*P. brasiliensis*
Pb3	Brazil, clinical	*P. brasiliensis*
Pb4	Brazil, clinical	*P. brasiliensis*
Pb5	Brazil, clinical	*P. brasiliensis*
Pb6	Brazil, clinical	*P. brasiliensis*
Pb7	Brazil, clinical	*P. brasiliensis*
Pb8	Brazil, clinical	*P. brasiliensis*
Pb9	Brazil, clinical	*P. brasiliensis*
Pb10	Perú, clinical	*P. brasiliensis*
Pb11	Brazil, clinical	*P. brasiliensis*
Pb13	Brazil, clinical	*P. brasiliensis*
Pb14	Brazil, clinical	*P. brasiliensis*
Pb15	Venezuela, clinical	*P. brasiliensis*
P179	Colombia, clinical	*P. brasiliensis*
Pb305	Venezuela, clinical	*P. brasiliensis*
Pb Bolivia	Bolivia, clinical	*P. brasiliensis*
Pb01	Brazil, clinical	*P. lutziii*
Pl1578	Brazil, clinical	*P. lutziii*
ED01	Brazil, clinical	*P. lutziii*
PlEE	Brazil, clinical	*P. lutziii*

Results

In silico assay design

We implemented bioinformatic approaches to search for regions within the genus *Paracoccidioides* that are unique to this genus (a criterion needed for high specificity), as well as likely to be present in all species within this genus (a criterion needed for high sensitivity), and that could therefore be used for the identification of *P. brasiliensis* and *P. lutzii* via PCR assays. We were able to find several regions that were optimal for primer design. We designed 3 candidate primer pairs within genomic regions of *P. brasiliensis* and *P. lutzii* that allow for the molecular detection of *Paracoccidioides* spp. These regions include two coding genes and one non coding region.

First, we searched for *Paracoccidioides* spp. scaffolds that were not similar to genomic regions of other species. Using the reference strains Pb18 and Pb03 of *P. brasiliensis* and Pb01 representing *P. lutzii*, the only scaffold that was present in all three reference strains was the scaffold named Supercontig 2.59 of the Pb03 assembly. This scaffold contains a non coding genomic region of the *Paracoccidiodies* spp. that is unique to the genus. The corresponding scaffolds in the two other reference strains are shown in Figure 1A. This scaffold met the criteria of not aligning to other closely related fungal species and of being present in all of the *Paracoccidioides* strains. As compared to chapter 9, where we were not able to design primer pairs using this approach due to the long stretches of low similarity regions within the species of *H. capsulatum*, the updated genomes of *Paracoccidioides* spp. allowed for the design of

highly specific primer pairs. For convenience we will call this identified unique region *GROP* (Genomic region of *Paracoccidioides*). This example serves as a proof of principle that the genome updates significantly improved regions of the assembly that could otherwise have been misclassified as variable between strains and species due to sequencing and/or assembly errors, e.g.., single nucleotide errors (SNEs) masquerading as true SNPs.

We next used, as a second approach, a search for entire protein-coding genes that are unique to *the Paracoccidioides* spp. This second approach allowed us to design two promising PCR primer pairs for the detection of the genus. The candidate genes were found using OrthoMCL matrix results. The matrix produced was queried for genes that did not have any likely orthologs in other sequenced species that we screened, previously described in chapter 9. We used strict filtering for the selection of the genes. Only genes that had no similarity to potentially orthologous genes in the species queried were selected for further analysis. Genes that met this criterion were then queried by Blastn (via the NCBI web server) against the nr database in order to check the uniqueness of the genes to *Paracoccidioides* spp. We found 282 genes that met the criteria of being unique using strict filtering, yet this list was limited to 5 genes with a described putative biological function. Although in theory any of the regions discovered could have been used, we specifically used the two genomic regions from non-trivially annotated genes as they seemed most likely to play a biological role i.e., to be robustly species-specific regions because of some true functional differentiator contrasting with closely related fungi. These 5 genes are listed

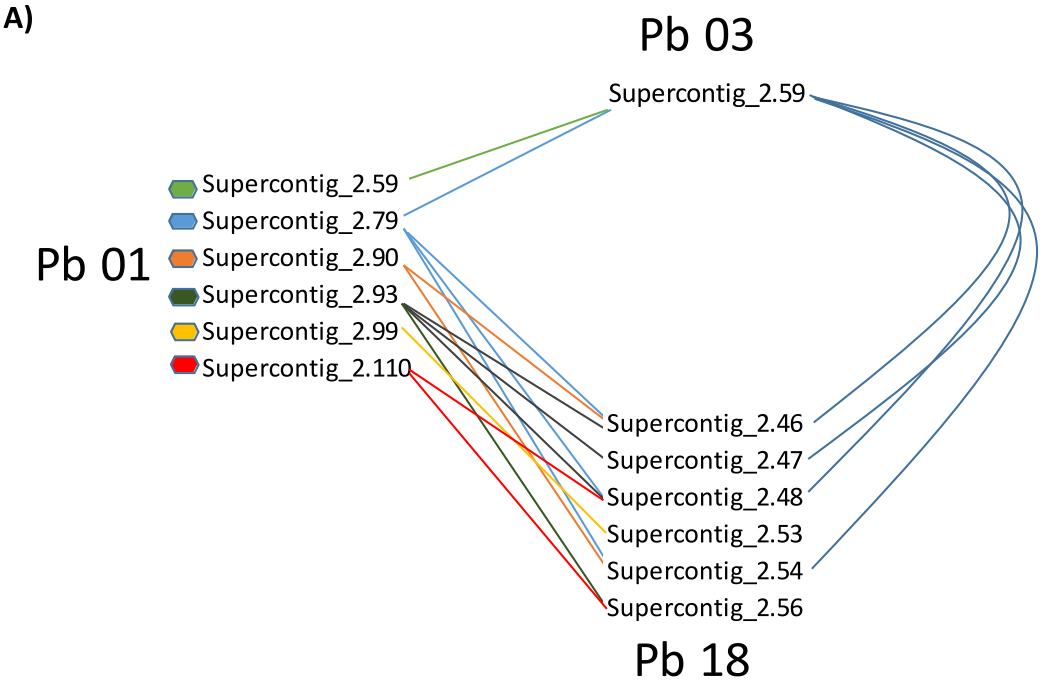
in Figure 1B. Of these 5 genes, we designed primer pairs for 2 genes. The first gene for which we designed primers encodes a protein having a known biological function, 2-dehydropantoate 2-reductase (PADG_00213, XM_015844381.1). For convenience, we will call this gene *2DROP* (2 dehydropantoate 2-reductase of *Paracoccidioides*). The second gene for which we designed primers also has a biological function and encodes a dipeptidase (PADG_07542, XM_015845235.1). For convenience, we will call this gene *DPOP* (Dipeptidase of *Paracoccidioides*). For real time PCR contexts, we did not consider designs involving a sequence-specific probe for a region between the two primers, as our aim was to keep costs low without sacrificing specificity. Instead, we focused on the use of DNA intercalating dyes such as SYBR Green. The primer details are listed in Table 2. The PCR conditions used are the same as described chapter 9. PCR conditions were the same for conventional and real-time PCR. The BLAST search of the primer and target sequences for *Paracoccidioides* spp did not yield any cross- reacting sequences. In addition, testing of nucleic acids from phylogenetic related fungi, others pathogens such as *Mycobacterium* spp, and other potentially cross-reacting microbes and human demonstrated no cross- reactivity with these organisms.

Using a third approach, which does not depend on gene annotations or the lengths of the scaffolds, we confirmed the uniqueness of the regions selected by the first two search strategies. The use of our algorithm for this third approach permitted the discovery of other unique genomic regions that were not found using the other strategies, but they were not analyzed further in this study.

Table 2. Primer sequences

Gene name	Primer	Sequence (5' - 3')
2DROP	Forward	TTCTAAGGAGCCGTTATGCTGT
2DROP	Reverse	CAACTCCATTGGCCTTCCATTC
DPOP	Forward	ATGCATGAACTGAAGACGCCACC
DPOP	Reverse	GAGAAATTGCCGGAGACTTTGAG
GROP	Forward	CTTCTGAGAAAACTGCTAAGATG
GROP	Reverse	CTAAAGAGTTCACTGTCTGCTGT

A)



B)

Gene Name	Pb01	Pb03	Pb18
2-dehydropantoate 2-reductase	PAAG_01565	PABG_03619	PADG_00213
alpha-amylase B	PAAG_02687	PABG_03167	PADG_01713
amine oxidase	PAAG_02065	PABG_02316	PADG_11154
dipeptidase	PAAG_04312	PABG_04087	PADG_07542
nuclear distribution protein PAC1	PAAG_11104	PABG_06688	PADG_02266

Figure 1. *Paracoccidioides* spp. regions used for the design of primer pairs. (A) Scaffolds used in the search for unique genomic regions of *Paracoccidioides* spp using the reference strains Pb18, Pb03 and Pb01. (B) List of 5 gene orthologs with an assigned biological function found in *Paracoccidioides* spp.

Analytical sensitivity and specificity of primers using fungal genomic DNA *in vitro*

We tested our primer designs using 92 *Paracoccidioides* spp. strains from the fungal DNA collections of CIB. The outcomes were positive for all of the 92 strains tested, i.e., the primer pairs had 100% analytical sensitivity *in vitro*. Specificity tests were done using the species listed in Table 1 in chapter 9 with the modifications described in Materials and Methods. The amplification was 100% specific to the genus *Paracoccidioidies* as no amplification was observed in other species. The amplicon sizes of the primer sets were: DPOP primer pair, 600 bp, 2DROP, 1000 bp and GROP, 300 bp, which were expected by design and confirmed by agarose gel electrophoresis. All PCR products showed homogenous amplicon sizes in all the tested DNAs and were confirmed by Sanger sequencing.

Limit of detection (LOD) using positive plasmid control

In order to test for the analytical limit of detection of the assays, 10 fold serial dilutions of positive control plasmids were performed from 1 ng/µl down to 1 fg/µl for all three primer sets using plasmid controls. The dilutions were amplified via real time PCR using standard conditions in triplicates for a sample processed on day 1, day 2 and day 3 for all plasmid controls.

For *2DROP*, the LOD was 1 fg/µl using real time PCR. The amplification curves are clearly separated from the no template control and primer dimers (Figure 2A). When visualized via agarose gel electrophoresis, the bands are also visible down

to 1 fg/µl (Figure 2B). All amplicons share the same melting curve temperature of 82 ^{o}C (Figure 2C), indicating high specificity of the region amplified across 92 strains from across Latin America. Primer dimer presence was not observed during late cycles of amplification.

For *DPOP*, the LOD was 10 fg/µl using real time PCR (Figure 3A). The amplification curves were clearly separated until 100 fg/µl. For lower dilutions which are not clearly separated in the amplification curves, the use of melting temperature data allows one to discern low levels of amplification from primer dimer noise. No amplification curves were observed for the 1 fg/µl dilution. Primer dimers are observed in late cycles above 40. All amplicons had a melting curve temperature of 81.5^{o}C, while primer dimers had a melting curve temperature of 74 ^{o}C (Figure 3C). The melting curve temperatures for all 92 strains showed marked homogeneity and reproducibility, indicating a strong conservation of this gene in the *Paracoccidioides* spp. When visualized via agarose gel electrophoresis, the bands are visible down to only 10 pg/µl (Figure 3B). This may be due to the amount of DNA that was amplified not being visible via the gel, but confirmed in the real time PCR run.

For *GROP*, the LOD was 100 fg/µl using real time PCR (Figure 4A). All amplicons had melting curve temperatures of 75.5^{o}C (Figure 4C), and when visualized via agarose gel electrophoresis, the bands are also visible down to only 100 fg/µl (Figure 4B). The homogenous amplification observed also indicates that this non coding genomic region is highly conserved within the diverse strains of *Paracoccidioides* spp.

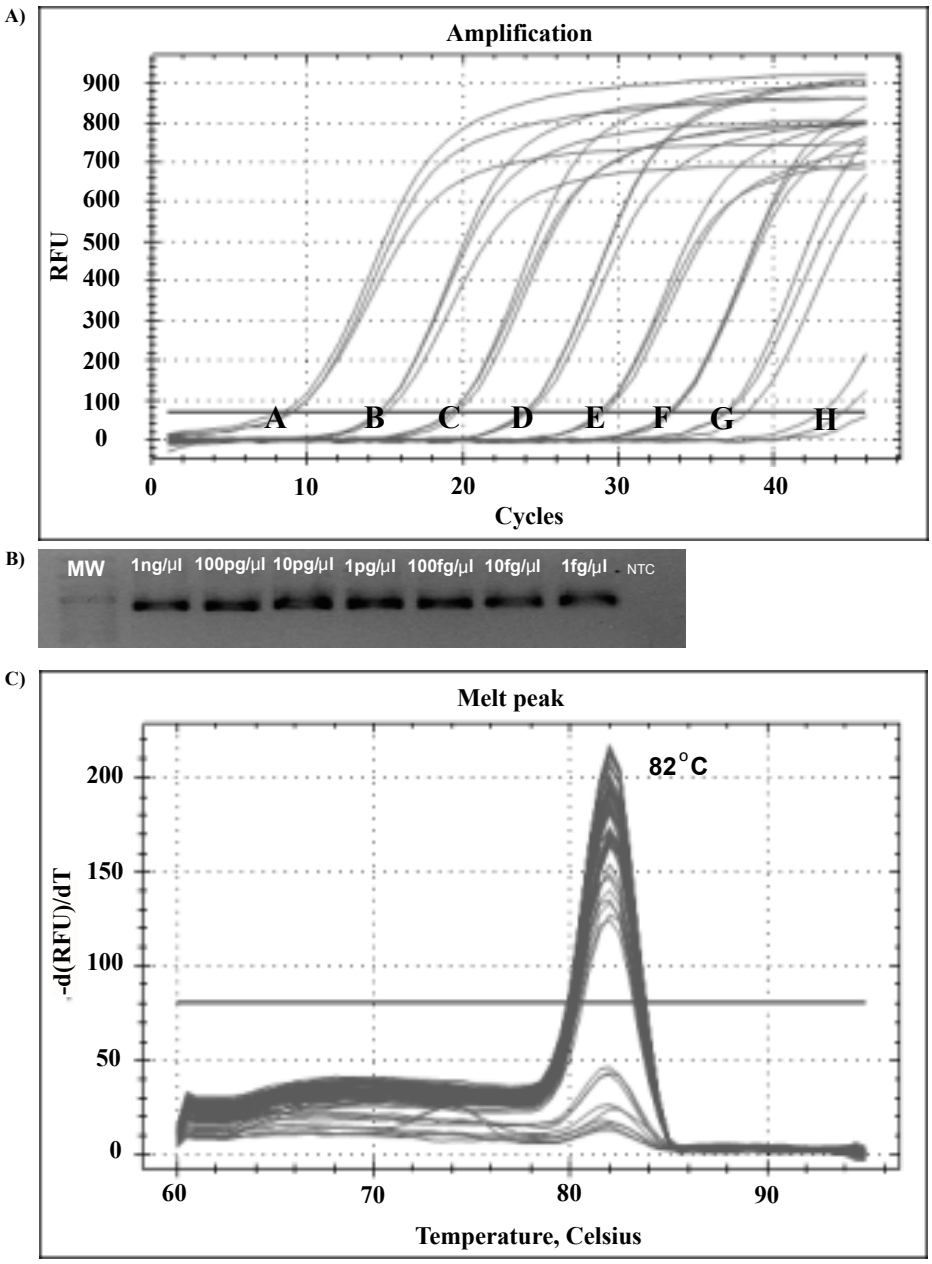


Figure 2. Real time PCR data of 2DROP. (A) Real time PCR amplification curves using the serial dilutions of positive plasmid control of 2DROP. (B) 2% agarose gel electrophoresis of PCR products obtained from real time PCR run shown in A. (C) Melting curve analysis of real time PCR run shown in A.

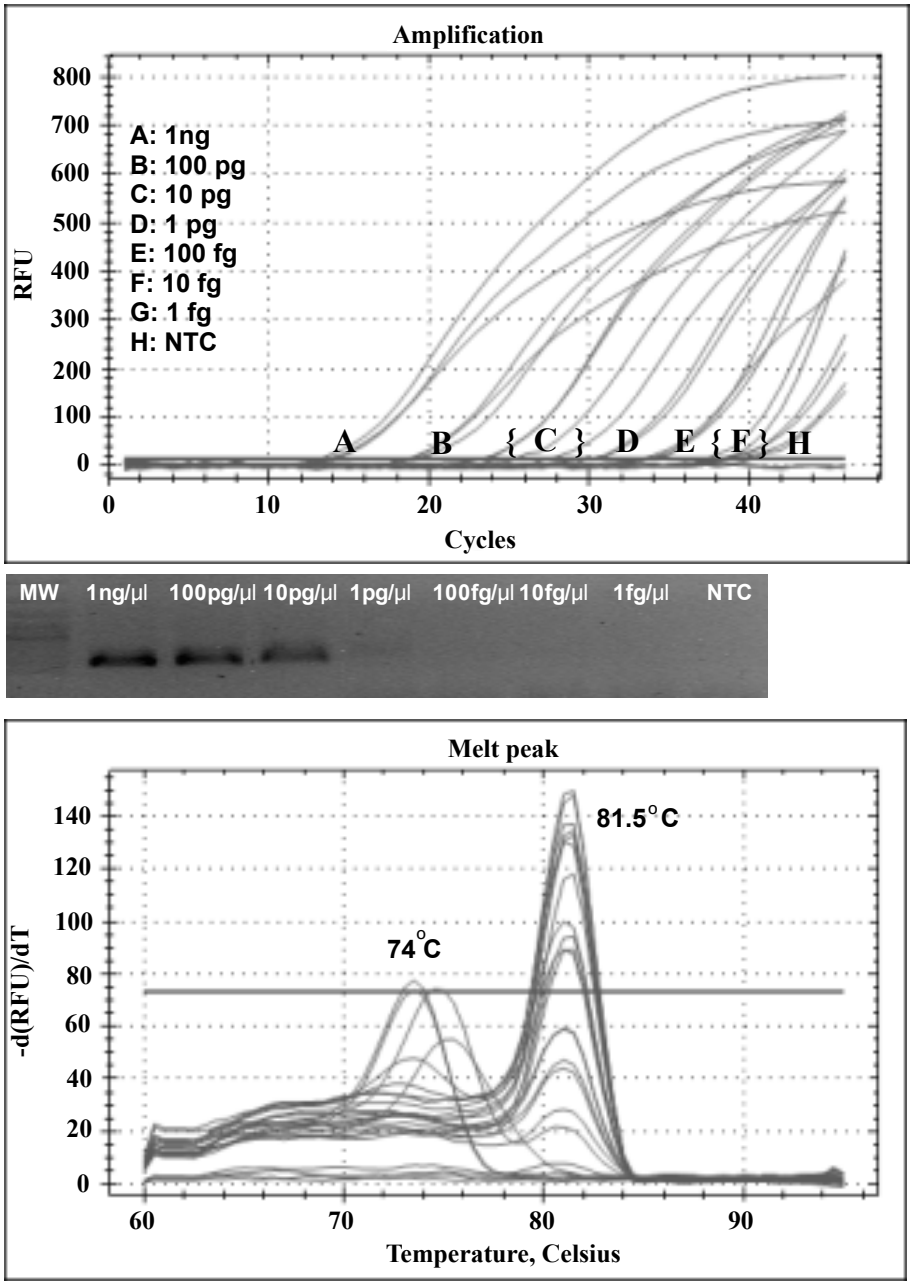


Figure 3. Real time PCR data of DPOP. (A) Real time PCR amplification curves using the serial dilutions of positive plasmid control of DPOP. (B) 2% agarose gel electrophoresis of PCR products obtained from real time PCR run shown in A. (C) Melting curve analysis of real time PCR run shown in A.

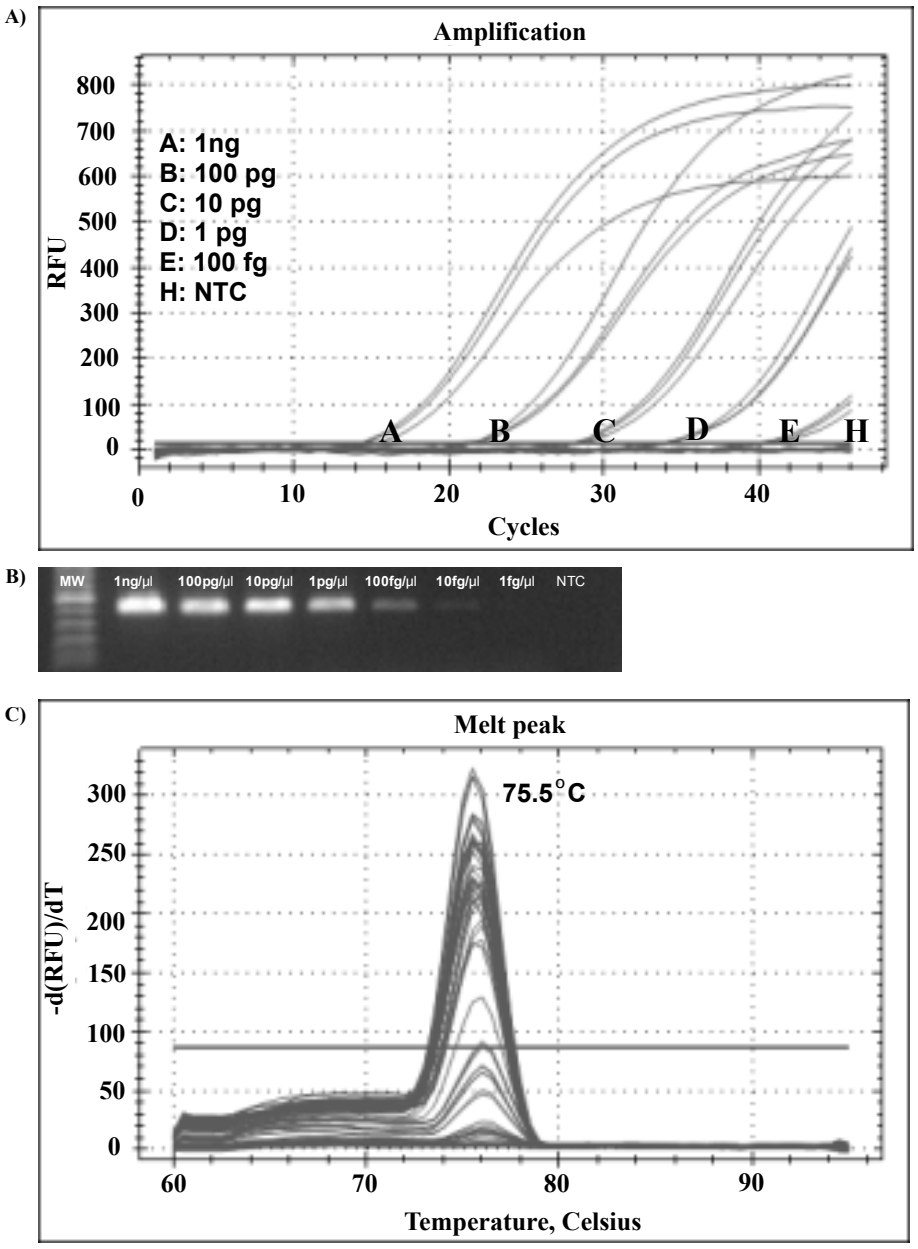


Figure 4. Real time PCR data of GROP. A) Real time PCR amplification curves using the serial dilutions of positive plasmid control of GROP. B) 2% agarose gel electrophoresis of PCR products obtained from real time PCR run shown in A. C) Melting curve analysis of real time PCR run shown in A.

Discussion

The primer pairs we designed and analytically tested in this study showed promising results for posterior validation using clinical or environmental samples. Validation of these primer sets in clinical or environmental samples is still pending. We observed analytical sensitivity and specificity of 100% (92/92) using DNA from *Paracoccidoides* spp. cultured isolates for 2DROP, DPOP and GROP primer sets. All three primer pair designs were purposely designed to work for conventional as well as real time PCR without the need of probes. The results suggest that for either conventional and/or real-time PCR technologies there is potential of the method for detecting *Paracoccidioides* spp.

To our knowledge, no study has shown the limit fragment size for the detection of fungal DNA in clinical or soil samples. Hence, we designed amplicon sizes which vary from 300 bp to 1000 bp. The use of fluorescent DNA intercalators with large amplicons should increase fluorescence signals in the presence of low levels of DNA, as larger amplicons should emit more signal throughout the amplification process. This is observed in the amplification curves of *2DROP*, which has the largest amplicon size (1000 bp) and the lowest LOD (1 fg/µl). Ideally, one should test all three of our primer pairs in order to determine the clinical and/or environmental sensitivity of the primer sets, taking into account the amplicon sizes. Our design of the *GROP* primer pair also serves as a proof of principle for the use of non coding regions of the genome containing no known genes in molecular detection, a less conventional

approach for molecular identification.

Although publications use variations of PCR to detect PCM, most primers detect only *P. brasiliensis* using either molecular beacons or nested PCR targeting the ITS regions. ITS is typically used as a general amplification primer for fungi and is highly variable in fungal species (Buitrago et al., 2009, Koishi et al., 2010, Dias et al., 2012, Pitz et al., 2013). The use of *PbGP43* for molecular detection of *P. brasiliensis* has been questioned by some authors to be included in all isolates of the *Paracoccidioides* genus (Dias et al., 2012). A recent publication uses Pb27 to detect PCM, but this gene can often have high similarity in some regions of its sequence with other fungal pathogens of the Onygenales (Rocha-Silva et al., 2016). Transposable elements have also been used as a possible marker for the detection of PCM, but require the interpretation of band patterns for the identification of species within the genus, and no unique band pattern exists for the genus itself (Alves et al., 2015). To our knowledge, this is the first report of a conventional PCR and corresponding real-time PCR assay for the detection of pathogens of the genus *Paracoccidioides* based on unique regions.

Overall, the development of these and other molecular tools may play an important role in PCM detection in the cases where cross-reactivity with other fungal species may interfere with conventional diagnosis (Bialek et al., 2000), as well as potentially facilitating epidemiological and ecological studies of *Paracoccidioides* spp.

Acknowledgements

We thank the Mycotic Diseases Branch of the Centers for Disease Control and Prevention (CDC), Atlanta, GA for hospitality and sharing of laboratory facilities and materials, for the use of fungal strain collections, and for giving J.G. and I.T. the opportunity to complete the experimental work presented here. We especially thank Dr Anastasia Litvintseva for expert guidance and assistance. J.G. acknowledges the partial funding of this work by Fulbright Colombia and by the Universidad del Rosario. The work described here was co-funded by Colciencias, Colombia via grant 1222-569-34875, A gene atlas for human pathogenic fungi.

References

1. Brummer E, Castaneda E, Restrepo A. Paracoccidioidomycosis: an update. Clin Microbiol Rev. 1993 Apr;6(2):89-117.
2. Restrepo A and Tobón A, *Paracoccidioides brasiliensis*, chapter 268. In Mandell, seveth edition, 2010, philadelphia, PA, principles and practices of infectious diseases edited by Gerald L. Mandell, Jhon E Bennett 7th ed.
3. McEwen JG, Bedoya V, Patino MM, Salazar ME, Restrepo A. Experimental murine paracoccidiodomycosis induced by the inhalation of conidia. J Med Vet Mycol. 1987 Jun;25(3):165-75.
4. Greer DL, Restrepo AM. La epidemiologia de la paracoccidioidomicosis. Bol Of Sanit Panam. 1977 sept; 82:428-45.
5. Sandhu GS, Aleff RA, Kline BC, da Silva Lacaz C. Molecular detection and

identification of *Paracoccidioides brasiliensis*. J Clin Microbiol. 1997 Jul;35(7):1894-6.

6. De Brito T, Sandhu GS, Kline BC, Aleff RA, Sandoval MP, Santos RT, et al. In situ hybridization in paracoccidioidomycosis. Med Mycol. 1999 Jun;37(3):207-11.
7. Bialek R, Ibricevic A, Aepinus C, Najvar LK, Fothergill AW, Knobloch J, et al. Detection of *Paracoccidioides brasiliensis* in tissue samples by a nested PCR assay. J Clin Microbiol. 2000 Aug;38(8):2940-2.
8. Gomes GM, Cisalpino PS, Taborda CP, de Camargo ZP. PCR for diagnosis of paracoccidioidomycosis. J Clin Microbiol. 2000 Sep;38(9):3478-80.
9. Motoyama AB, Venancio EJ, Brandao GO, Petrofeza-Silva S, Pereira IS, Soares CM, et al. Molecular identification of *Paracoccidioides brasiliensis* by PCR amplification of ribosomal DNA. J Clin Microbiol. 2000 Aug;38(8):3106-9.
10. Lindsley MD, Hurst SF, Iqbal NJ, Morrison CJ. Rapid identification of dimorphic and yeast-like fungal pathogens using specific DNA probes. J Clin Microbiol. 2001 Oct;39(10):3505-
11. Semighini CP, de Camargo ZP, Puccia R, Goldman MH, Goldman GH. Molecular identification of *Paracoccidioides brasiliensis* by 5' nuclease assay. Diagn Microbiol Infect Dis. 2002 Dec;44(4):383-6.
12. San-Blas G, Nino-Vega G, Barreto L, Hebeler-Barbosa F, Bagagli E, Olivero de Briceno R, et al. Primers for clinical detection of *Paracoccidioides*

brasiliensis. J Clin Microbiol. 2005 Aug;43(8):4255-7.

13. Diez S, Garcia EA, Pino PA, Botero S, Corredor GG, Peralta LA, et al. PCR with *Paracoccidioides brasiliensis* specific primers: potential use in ecological studies. Rev Inst Med Trop Sao Paulo. 1999 Nov-Dec;41(6):351-8.
14. Nascimento E, Martinez R, Lopes AR, de Souza Bernardes LA, Barco CP, Goldman MH, et al. Detection and selection of microsatellites in the genome of *Paracoccidioides brasiliensis* as molecular markers for clinical and epidemiological studies. J Clin Microbiol. 2004 Nov;42(11):5007-14.
15. Teles FR, Martins ML. Laboratorial diagnosis of paracoccidioidomycosis and new insights for the future of fungal diagnosis. Talanta. 2011 Oct 15;85(5):2254-64.
16. Muñoz JF, Gallo JE, Misas E, Priest M, Imamovic A, Young S, et al. Genome update of the dimorphic human pathogenic fungi causing paracoccidioidomycosis. PLoS Negl Trop Dis. 2014 Dec;8(12):e3348.
17. Desjardins CA, Champion MD, Holder JW, Muszewska A, Goldberg J, Bailao AM, et al. Comparative genomic analysis of human fungal pathogens causing paracoccidioidomycosis. PLoS Genet. 2011 Oct;7(10):e1002345.
18. Buitrago MJ, Merino P, Puente S, Gomez-Lopez A, Arribi A, Zancopé-Oliveira RM, et al. Utility of real-time PCR for the detection of *Paracoccidioides brasiliensis* DNA in the diagnosis of imported paracoccidioidomycosis. Med Mycol. 2009 Dec;47(8):879–82.
19. Koishi AC, Vituri DF, Dionízio Filho PSR, Sasaki AA, Felipe MSS, Venancio

EJ. A semi-nested PCR assay for molecular detection of *Paracoccidioides brasiliensis* in tissue samples. Rev Soc Bras Med Trop. 2010 Nov;43(6):728–30.

20. Dias L, de Carvalho LF, Romano CC. Application of PCR in serum samples for diagnosis of paracoccidioidomycosis in the southern Bahia-Brazil. Hotez PJ, editor. PLoS Negl Trop Dis. Public Library of Science; 2012;6(11):e1909.
21. Pitz A de F, Koishi AC, Tavares ER, Andrade FG de, Loth EA, Gandra RF, et al. An optimized one-tube, semi-nested PCR assay for *Paracoccidioides brasiliensis* detection. Rev Soc Bras Med Trop. SBMT; 2013 Nov;46(6):783–5.
22. Rocha-Silva F, Gomes LI, Góes AM, Graciele-Melo C, Caligiorne RB. Real Time Polymerase Chain Reaction (rt-PCR): A New Patent to diagnostic purposes for paracoccidioidomycosis. Recent Pat Endocr Metab Immune Drug Discov. 2016 Sep 5.
23. Alves FL, Ribeiro MA, Hahn RC, de Melo Teixeira M, de Camargo ZP, Cisalpino PS, et al. Transposable elements and two other molecular markers as typing tools for the genus *Paracoccidioides*. Med Mycol. 2015 Feb 1;53(2):165–70.

Chapter 11

Towards multiple-SNP motif analyses of loci associated with phenotypic traits

Toward multiple-SNP motif analyses of loci associated with phenotypic traits

Juan E. Gallo [a,b,c] , Elizabeth Misas [a,d] , Juan G. McEwen, MD, PhD [a,e] , Oliver K. Clay, PhD [a,f]

[a] Cellular and Molecular Biology Unit, Corporación para Investigaciones Biológicas, Medellín, Colombia

[b] Doctoral Program in Biomedical Sciences, Universidad del Rosario, Bogotá, Colombia

[c] Universidad CES, School of Medicine, GenomaCES, Medellín, Colombia

[d] Institute of Biology, Universidad de Antioquia, Medellín, Colombia

[e] Faculty of Medicine, Universidad de Antioquia, Medellín, Colombia

[f] School of Medicine and Health Sciences, Universidad del Rosario, Bogotá, Colombia

Address for Correspondence

Oliver K. Clay, PhD
Cellular and Molecular Biology Unit
Corporación para Investigaciones Biológicas
Carrera 72A # 78B-141
Medellín, Colombia
Tel: +57 4 403 5990
Fax: +57 4 441 5514
E-mail: oliver.clay@gmail.com

Article type: Letter to the Editor

Brief title: SNP motif analysis of trait-associated loci

Total word count: 362 words for main text and references (+ 83 words for figure title and legend)

Funding: This work was funded by COLCIENCIAS grant 221356934877.

Disclosures: The authors (J.E.G., E.M., J.G.M., O.K.C.) report that they have no relationships relevant to the contents of this paper to disclose.

A recent JACC article (**1**), and other studies (e.g., **2**), are greatly improving our overview of candidate genetic co-determinants of cardiovascular risk or blood pressure, also raising curiosity about sources of consistently strong (e.g., $p < 10^{-15}$) observed associations. In 12q24, a SNP (rs3184504; **1**) in one of 6 loci associated with coronary artery disease was also strongly associated with blood pressure (**2**); a web summary (**3**) noted its associations with 17 diseases or disease-related characteristics.

From 1000 Genomes (**4**), we extracted 15-SNP, < 30-kb haplotype-motif views around rs3184504 for individual and pooled world populations. The world view reveals strong population structure of the 15-SNP motifs (Fig. 1, $N = 2 \times 2504$ haplotypes). At least two of the giant components and their master motifs (H1-H4; encoding of ref. **4**) appear also in individual populations, also some (e.g., GBR) that are not considered appreciably admixed or structured. Two dominating consensus motifs (H1+H2+H4 and H3) account for 83% of the 5008 haplotypes worldwide.

Statistical significance can occur also where individual SNPs have no adaptive significance; if one drug is administered to cats and another to horses, efficacy differences may just reflect cat-horse differences (**5**). In three (boxed) columns (SNPs) in Fig. 1 with lowest p, allele-0 appears exclusively in the null motif H3 (red) and two lower-frequency variants of its component (blue), thus best exposing an underlying motif-component structure that explains observable variation and covariation.

SNP-by-SNP associations can be profitably complemented by (~ 30 kb) *k*-SNP motif associations where two or more giant components dominate, even if the reasons for their presence are uncertain.

References

1. Webb TR, Erdmann J, Stirrups KE, et al. Systematic evaluation of pleiotropy identifies 6 further loci associated with coronary artery disease. J Am Coll Cardiol 2017;69:823-36.

2. Ehret GB, Ferreira T, Chasman DI, et al. The genetics of blood pressure regulation and its target organs from association studies in 342,415 individuals. Nat Genet 2016; 48:1171-84.

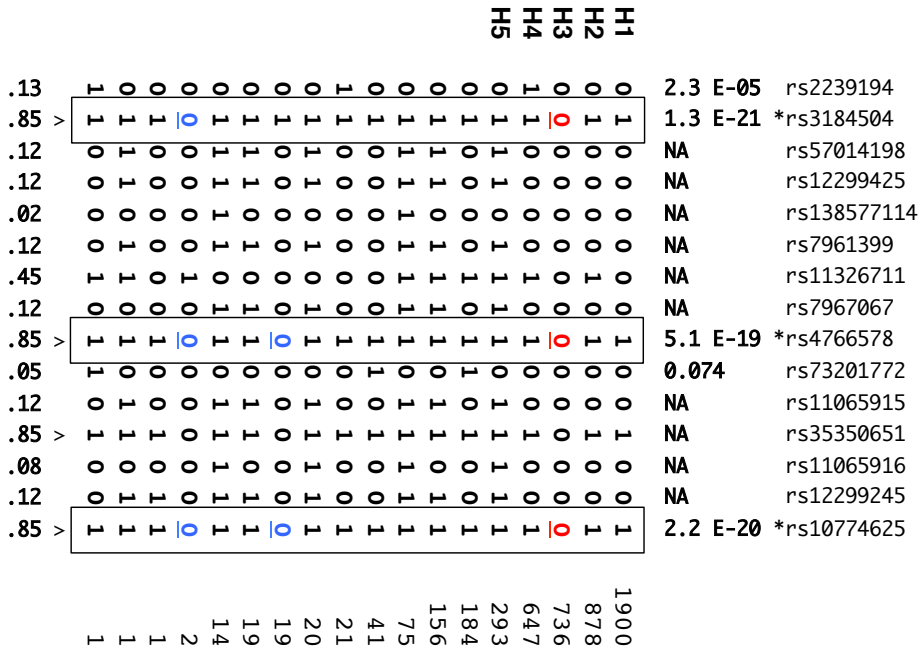
3. www.sciencedaily.com/releases/2017/02/170221081945.htm

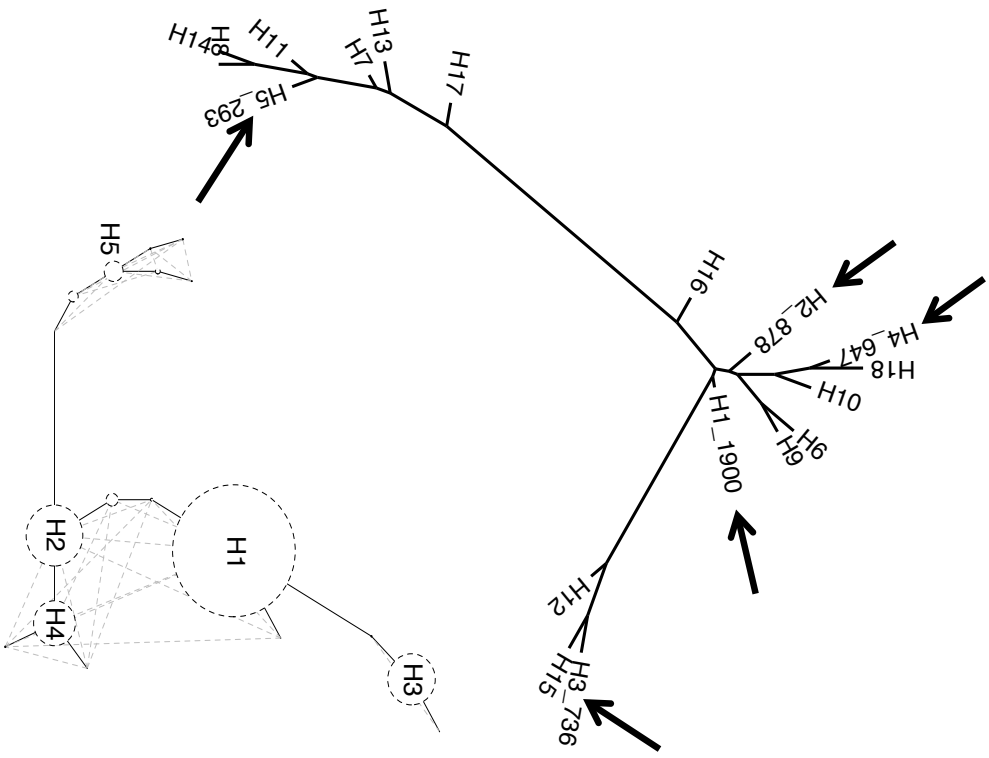
4. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, et al. A global reference for human genetic variation. Nature 2015;526:68-74.

5. Price T. Correlated evolution and independent contrasts. Philos Trans R Soc Lond B Biol Sci. 1997;352:519-29.

Legend to the Figure

Figure 1. **Human local common-SNP haplotype motifs/clusters including 12q24 SNP rs3184504.** Each row of the matrix is a motif, i.e., a vector of biallelic SNP states observed in 1000 Genomes. Asterisks denote SNPs (columns) with diastolic blood pressure associations $p < 10^{-15}$ in ref. **2** (top: *p* values, SNP rs numbers; right: motif instances observed; bottom: frequencies of allele 1). Motif networks are plotted from difference (Hamming distance) matrices by programs Phylip/neighbor (top) and R/pegas (bottom); arrows indicate the five most frequent motifs (H1-H5).


H5 H4 H3 H2 H1
.13 1 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 2.3 E-05 rs2239194
.85 > 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1.3 E-21 *rs3184504
.12 0 1 0 0 1 1 0 1 0 0 1 1 0 1 0 0 0 0 0 NA rs57014198
.12 0 1 0 0 1 1 0 1 0 0 1 1 0 1 0 0 0 0 0 NA rs12299425
.02 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 NA rs138577114
.12 0 1 0 0 1 1 0 1 0 0 1 1 0 1 0 0 0 0 0 NA rs7961399
.45 1 1 0 1 0 0 0 0 0 0 1 1 1 1 1 0 1 0 0 NA rs11326711
.12 0 0 0 0 1 1 0 1 0 0 1 1 0 1 0 0 0 0 0 NA rs7967067
.85 > 1 1 1 0 1 1 0 1 1 1 1 1 1 1 1 0 1 1 5.1 E-19 *rs4766578
.05 1 0 0 0 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0.074 rs73201772
.12 0 1 0 0 1 1 0 1 0 0 1 1 0 1 0 0 0 0 0 NA rs11065915
.85 > 1 1 1 0 1 1 1 0 1 1 1 1 1 1 1 0 1 0 NA rs35350651
.08 0 0 0 0 1 0 0 1 0 0 1 0 0 1 0 0 0 0 0 NA rs11065916
.12 0 1 1 0 1 1 1 0 1 0 0 1 1 0 1 0 0 0 0 NA rs12299245
.85 > 1 1 1 0 1 1 1 0 1 1 1 1 1 1 1 0 1 1 2.2 E-20 *rs10774625
1 1 1 2 14 19 19 20 21 41 75 156 184 293 647 736 878 1900


H14
H8
H11
H13
H7
H17
H5_293
H5
H16
H2_878
H4_647
H18
H10
H6
H9
H1_1900
H12
H15
H3_736
H2
H1
H4
H3

Chapter 12

Current state of cardiovascular genomics in Colombia

+Model
RCCAR-379; No. of Pages 2
ARTICLE IN PRESS

Rev Colomb Cardiol. 2016;xxx(xx):xxx.e1–xxx.e2

SOCIEDAD COLOMBIANA DE CARDIOLOGIA Y CIRUGIA CARDIOVASCULAR

Revista Colombiana de
Cardiología

www.elsevier.es/revcolcar

EDITORIAL

Current state of cardiovascular genomics in Colombia

La actualidad de la genómica clínica en el área cardiovascular en Colombia

Juan Esteban Gallo

Scientific Director, Genoma CES, Medellín, Colombia

Received 18 October 2016; accepted 21 October 2016

Clinical genomics has advanced exponentially worldwide during the last decade due to innovation in sequencing techniques, which has made procedures increasingly faster and more economical. The current scenario is one in which the physician has the latest generation diagnostic aids in molecular and sequencing techniques at his/her disposal. What follows is a brief review of diagnostic aids based on genetic knowledge of the patient, and their current status in the Colombian market.

Prior to new generation sequencing technologies, the alternative for obtaining a patient's genetic information was traditional sequencing methods, based on Sanger's technology. In order to obtain the sequence of a specific gene, the target gene of interest was first selected. The physician had to have an idea of which of these genes could be involved in his/her patient's disease, a decision which can be relatively complex to make. Sequencing of any gene can be cumbersome due to the large difference in size and number of exons among genes, which is relevant for understanding the way in which genes are sequenced using traditional methods. For genetic studies of one gene, first, all coding parts of the gene are amplified, using the PCR (polymerase chain reaction) technique. For example, the SCN5A gene has 28 exons with an approximate size of 80 kb. This entails many PCR amplification cycles, where each exon must be amplified independently, thus increasing costs and labor time of molecular biology personnel. In cases in which sequencing of a single gene has clinical validity, taking SCN5A as an example, the complexity of the laboratory processes pales in comparison to the great benefit for the patients of knowing the possible genetic cause of their disease or cardiovascular event. However, since cardiovascular conditions are not classical Mendelian diseases, in most cases it is impossible to make an easy decision regarding which specific gene to study.

The era of new generation sequencing expanded a bit more the genetic tests that can be run simultaneously. Sequencing of panels of genes associated with certain diseases may be more useful when making decisions, compared with the information provided by the sequencing of a single gene. There are many gene panels on the market, which in the case of cardiovascular diseases may range from a dozen to several hundred genes. Many of these are commercial kits with a fixed list of genes to be sequenced. The availability of these panels varies in the reference laboratories, depending on the sequencing technology implemented. In economic terms, they can be a bit more expensive than single gene sequencing, but the clinical usefulness increases with the amount of information obtained. The literature has shown that in certain cases the gene involved in the pathology is not included in the complete lists available for study in these panels, due to the fact that the human genome is still being studied, and the association of genes with diseases has not yet concluded.

In addition, there are other panels which are very effective and useful. These do not sequence genes associated with diseases, but rather focus on genes whose

DOI of original article:
http://dx.doi.org/10.1016/j.rccar.2016.10.044
E-mail address: jegallo@ces.edu.co

http://dx.doi.org/10.1016/j.rccar.2016.10.045

Please cite this article in press as: Gallo JE. Current state of cardiovascular genomics in Colombia. Rev Colomb Cardiol. 2016. http://dx.doi.org/10.1016/j.rccar.2016.10.045

functions are related to metabolism and drug interactions. These tests are known as pharmacogenomic tests, which provide information regarding how a patient will react to taking a medication: whether favorably, as a non-responder, or dose-dependent. In the cardiovascular field, there are many medications which are included in these tests, among which are antiarrhythmics, anticoagulants, antianginals, antiplatelets, beta-blockers, calcium channel blockers, phosphodiesterase inhibitors, and statins.

These results, together with genomic tests of genes associated with diseases, lead to an understanding of the reasons why specific patients do not respond adequately to certain treatments. An example of this would be those who do not respond to cholesterol medications, for whom there can be various scenarios:

1. The patients have gene mutations associated directly with body processes which process cholesterol, such as the LDL and Apo B genes. These patients do not respond to the medication since their genetics do not allow the regulation of cholesterol levels.
2. The patients do not have mutations in the genes associated with body processes, but they do have mutations in the cytochrome P450 genes involved in drug metabolism, causing an inadequate response to treatment.
3. The patients have a combination of mutations 1 and 2. In this case, there is much evidence to consider new generation medications which may be more effective, but the genetic causes of high cholesterol levels affect the results.

Furthermore, what should be done when studying orphan cardiovascular diseases which do not have enough clinical genomics studies, and for which there are no available gene panels to aid in diagnosis? The most inclusive test today with clinical usefulness is whole exome sequencing. This technique is based on sequencing all the coding regions of the human genome. At last report, the human genome codes for approximately 20,000 genes. This complete picture provides greater clarity on how the body as a whole maintains a functional balance through its protein structures, which allows the physician to treat his/her patient with precise and personalized medicine.

The implementation of whole exome sequencing provides, in a relatively economical fashion, all the genetic information which will code for proteins whose functions will be involved in the maintenance, performance and healthy functioning of the human body. Being the most inclusive does not mean that it is the most employed test. Many physicians still order single gene sequencing tests.

Perhaps the myth regarding how expensive panel or whole exome tests can be, or the lack of coverage under the compulsory health plan, may be the reasons why single gene tests are still ordered. On the other hand, there is a great cost-benefit in sequencing all an individual's genes, *versus* one gene at a time. With regard to costs, single gene sequencing can cost approximately one to three million pesos, and more in some cases. The cost of exome sequencing is close to six million pesos. To simplify the comparison, let us suppose that the same user has the SCN5A sequencing done for two million pesos. If we compare the cost-benefit of obtaining information on one gene for two million pesos, and a complete exome for six million pesos, the sequencing of each gene using the exome (taking into account that the human genome has approximately 20,000 genes) would be 300 pesos. This cost-benefit exercise clarifies that whole exome sequencing continues to be a better option when genetic tests are needed. As far as coverage by POS [compulsory health insurance coverage], in 2015, the Ministry of Health and Social Protection issued resolution 5592, where the Health Benefits Plan was comprehensively updated, funded by the Captitation Payment Unit of the General Health Social Security System. Codes 90.8.4.02 to 90.8.4.39 mention system coverage for molecular/genetic/genomic tests. In particular, codes 90.8.4.12 (molecular study of diseases), 90.8.4.19 (mitochondrial DNA genetic studies), 90.8.4.20 (molecular gene studies), 90.8.4.22 (molecular exon studies), and 90.8.4.24 (molecular study of mutations), fit within the interpretation that the whole exome sequencing could be implemented for these tests.

In conclusion, the use of whole exome sequencing is a test that may have sufficient clinical validity when issuing a complete diagnosis. The potential benefits for the patients outweigh the risks. The presence of pathogenic genomic variants does not mean a 100% probability of developing some medical problem, since other still undescribed factors may exist. For many diseases which may be treatable, this information will allow the patient and physician to implement timely prevention processes.

Please cite this article in press as: Gallo JE. Current state of cardiovascular genomics in Colombia. Rev Colomb Cardiol. 2016. http://dx.doi.org/10.1016/j.rccar.2016.10.045

Chapter 13
Hipercolesterolemia familiar heterocigota en manejo con anti-PCSK9

+Model
RCCAR 405 1–8
ARTICLE IN PRESS

Rev Colomb Cardiol. 2017;xxx(xx):xxx–xxx

SOCIEDAD COLOMBIANA DE CARDIOLOGIA Y CIRUGIA CARDIOVASCULAR

Revista Colombiana de
Cardiología

www.elsevier.es/revcolcar

CARDIOLOGÍA DEL ADULTO – PRESENTACIÓN DE CASOS

Hipercolesterolemia familiar heterocigota en manejo con anti–PCSK9

Q1 Mauricio Duque*, María C. Gaviria, Juanita González y Juan E. Gallo

Cardiología y Electrofisiología, Genoma CES, Universidad CES, Medellín, Colombia

Recibido el 15 de diciembre de 2016; aceptado el 12 de marzo de 2017

PALABRAS CLAVE
Hipercolesterolemia familiar;
Enfermedad cardiovascular;
Enfermedad coronaria;
Inhibidores

Resumen

Introducción: la hipercolesterolemia familiar representa un factor de riesgo sustancial para padecer enfermedad coronaria prematura, arterial periférica y valvular. Se han descrito dos formas según su alteración genética y cigocidad, así como, tres mutaciones genéticas asociadas. Pese a que el tratamiento con estatinas se considera la primera línea, algunos pacientes no alcanzan metas, de modo que se han utilizado los inhibidores del PCSK9 como nueva estrategia.

Métodos y materiales: se expone el caso de una paciente de 42 años con hipercolesterolemia familiar heterocigota tratada con inhibidores del PCSK9. Se describen los criterios y estudios genéticos utilizados para realizar el diagnóstico, la cronología de tratamientos que recibió y los exámenes de laboratorio anteriores y posteriores al inicio del evolocumab. Adicionalmente se hace una revisión de tema acerca de la hipercolesterolemia familiar y su tratamiento con inhibidores del PCSK9.

Conclusiones: la hipercolesterolemia familiar es una enfermedad que ocasiona graves consecuencias cardiovasculares. Los inhibidores de PCSK9 se han convertido en una alternativa prometedora para aquellos que no responden a las terapias convencionales. Se requieren estudios que corroboren o contradigan los beneficios y eventos adversos encontrados hasta el momento en que los pacientes se someten a estas nuevas terapias para así ofrecer un tratamiento ideal y oportuno.

* Autor para correspondencia.
Correo electrónico: mauricioduquemd@gmail.com (M. Duque).

http://dx.doi.org/10.1016/j.rccar.2017.03.002

Cómo citar este artículo: Duque M, et al. Hipercolesterolemia familiar heterocigota en manejo con anti–PCSK9. Rev Colomb Cardiol. 2017. http://dx.doi.org/10.1016/j.rccar.2017.03.002

+Model
RCCAR 405 1–8
ARTICLE IN PRESS

KEYWORDS
Familial hypercholesterolaemia;
Cardiovascular disease;
Coronary disease;
Inhibitors

Heterozygous familial hypercholesterolaemia being managed with anti–PCSK9

Abstract
Introduction: Familial hypercholesterolaemia is a substantial risk factor for suffering premature coronary, peripheral arterial, and valular disease. There are two forms described, depending on their genetics and zygosity, as well as three associated genetic mutations. Although treatment with statins is considered first line, some patients do not reach targets, as such that that PCSK9 inhibitors have been used as a new strategy.
Materials and method: A case is presented of a 42 year-old patient with heterozygous familial hypercholesterolaemia treated with PCSK9 inhibitors. The criteria and genetic studies used to make a diagnosis are described, as well as the chronology of the treatments that have been received and the laboratory results before and after starting with evolocumab. A review has also been made of the subject of familial hypercholesterolaemia and its treatment with PCSK9 inhibitors.
Conclusions: Familial hypercholesterolaemia is a diseases that may have serious cardiovascular consequences. PCSK9 inhibitors have become a promising alternative for those who do not respond to conventional therapies. Studies are required that can corroborate or contradict the benefits and adverse effects found up until now in patients subjected to these new therapies in order to offer an ideal and appropriate treatment

Introducción

La hipercolesterolemia familiar es un trastorno genético autosómico dominante que se caracteriza por un aumento importante de los niveles de colesterol de baja densidad (c-LDL, su sigla en inglés), y a su vez genera un incremento significativo en el riesgo cardiovascular, acompañado de sus patologías y complicaciones consecuentes.

Dentro de la hipercolesterolemia familiar se enmarcan dos tipos: la homocigota y la heterocigota. En ambos casos, la enfermedad coronaria aparece antes que en el resto de la población.

Se ha calculado que entre 14 y 34 millones de personas en el mundo padecen hipercolesterolemia familiar, sin embargo, en menos del 1% de los afectados se logra establecer el diagnóstico[1], lo cual implica que muchos pacientes no reciban las medidas terapéuticas adecuadas para tratar una enfermedad que suele ser muy agresiva para adultos jóvenes y en edad media.

Se sabe que existen tres mutaciones genéticas asociadas con la hipercolesterolemia familiar: la del receptor de LDL (R-LDL), la de la apolipoproteína B (ApoB) y la de la proproteína convertasa subtilina/kenina 9 (PCSK9). Gracias a que se han reconocido dichas mutaciones, se han podido desarrollar terapias farmacológicas que buscan un impacto en aquellos pacientes que no responden a los medicamentos de uso habitual para hipercolesterolemias sin origen genético. Dentro de esta nueva gama de medicamentos están los inhibidores de PCSK9, anticuerpos monoclonales que buscan influenciar la reducción de niveles de c-LDL.

A continuación se presenta el caso de una paciente joven con hipercolesterolemia familiar heterocigota, diagnosticada mediante estudio genético, con importantes compromisos cardiovasculares y no respuesta a tratamientos farmacológicos convencionales, razón por la que se decidió iniciar terapia con inhibidores de PCSK9.

Caso

Paciente femenina, de 42 años, raza mestiza, madre de dos hijas, con antecedente de dislipidemia diagnosticada a los 32 años y en tratamiento con estatinas, sin historia de consumo de sustancias psicoactivas, alcohol, tabaquismo, sobrepeso u obesidad. Como antecedentes familiares relató enfermedad coronaria y dislipidemia en madre, hermana y tíos maternos (entre ellas dos mujeres); estos últimos murieron de infarto agudo de miocardio antes de los 50 años. Su madre tuvo el primer infarto agudo de miocardio a los 55 años y es diabética insulino-requiriente. Ambas hijas, adolescentes, tienen dislipidemia documentada.

Fue valorada por primera vez por el grupo CES Cardiología en 2014, fecha en la que consultó para continuar manejo de su dislipidemia, dolor en el miembro superior derecho y dolor precordial de varios meses de evolución.

Debido al cuadro clínico y sus antecedentes personales, se hicieron varios estudios e intervenciones. Se llevó a arteriografía selectiva de miembros superiores en la que se documentó síndrome del opérculo torácico bilateral con compromiso severo de la extremidad superior derecha. También se realizó arteriografía coronaria en la que se evidenció compromiso severo del *ostium* de la primera rama diagonal (*ramus* intermedio), y así mismo se hizo angioplastia con doble balón e implantación de un *stent* medicado en el tercio medio de la arteria coronaria descendente anterior y primera rama diagonal. La arteria coronaria derecha presentaba enfermedad aterosclerótica difusa sin repercusión hemodinámica.

Cómo citar este artículo: Duque M, et al. Hipercolesterolemia familiar heterocigota en manejo con anti–PCSK9. Rev Colomb Cardiol. 2017. http://dx.doi.org/10.1016/j.rccar.2017.03.002

+Model
RCCAR 405 1–8
ARTICLE IN PRESS

Tabla 1 Criterios de ''Dutch lipid clinic network'' para hipercolesterolemia familiar, aplicados al caso

Criterios que cumple la paciente	Puntaje
Infarto agudo de miocardio en madre de la paciente a los 55 años por enfermedad coronaria. Requirió manejo con *bypass* coronario.	1
Hija de la paciente con hipercolesterolemia: Hija #1: 10 años, c-LDL* 141 mg/dl.	1
Menor de 18 años con c- LDL en percentil superior para edad y género (hija #1).	1
Paciente con enfermedad coronaria prematura (mujer menor de 60 años): implante de 2 *stents* medicados	2
c-LDL entre 191 y 250 mg/dl	3
Total	**8**

* c-LDL: colesterol de baja densidad.

Debido a estos hallazgos y al cuadro clínico sugestivo de hipercolesterolemia familiar, se aplicaron los criterios de la escala holandesa ''Dutch lipid clinic network'' para hipercolesterolemia familiar (tabla 1), la cual arrojó un puntaje de 10 y apoyó el diagnóstico definitivo con base en los siguientes criterios: familiar en primer grado con enfermedad coronaria, familiar en primer grado con dislipidemia, menores de 18 años con anormalidad en el c-LDL, enfermedad coronaria prematura, enfermedad vascular periférica y niveles actuales de c-LDL entre 191-250 mg/dl. Se estableció que el riesgo cardiovascular global por la escala de Framingham era del 39%.

Cabe anotar que el diagnóstico confirmatorio de hipercolesterolemia familiar debe realizarse con un estudio genético, prueba de oro para estos casos. En concordancia, este se solicitó al grupo GENOMACES, quienes reportaron tres mutaciones asociadas a dislipidemia, ubicadas en los genes APOA5, APOE y SCARB1. Adicionalmente, la paciente tiene otras mutaciones genéticas, sin embargo, estos cambios no se ha reportado asociación clínica en las bases de datos genómicas: no obstante, otros cambios genéticos dentro del mismo gen se han relacionado directamente con las dislipidemias. Dichas mutaciones se encuentran en los genes: INSRR (1 mutación, 1q23.1), APOB (1 mutación, 2p24.1), GCKR (2p23.3), ABCA12 (2q35), CELSR3 (3p21.31), MTTP (4q23), ABCA13 (7p12.3), LDLRAD3 (11p13), CELA1 (12q13.13), LMF1 (16p13.3), APOBR (16p11.2), ABCA10 (17q24.3), INSR (19p13.2) y CELSR1 (22q13.31).

Respecto al tratamiento de su dislipidemia, previo al diagnóstico genético recibió diversos manejos: inicialmente fue tratada con dieta, ejercicio aeróbico (300 minutos por semana de moderada intensidad, según las indicaciones de la Sociedad Americana del Corazón[2]) y tratamiento farmacológico hipolipemiante intensivo (tabla 2), sin cambios significativos en los niveles de c-LDL (valores de LDL por encima de 200 mg/dl). No se reportó intolerancia a las estatinas, pero dada la intensidad del tratamiento se hicieron mediciones de creatinina quinasa (CK), que arrojaron resultados dentro de los límites normales.

Pese a todas las medidas farmacológicas y no farmacológicas mencionadas, no se logró controlar su perfil de lípidos con la medicación de uso común: estatinas de alta intensidad, tratamiento combinado con ezetimibe y colestiramina (tabla 3). En consecuencia, se decidió iniciar manejo con evolocumab 420 mg subcutáneo con frecuencia mensual. Se ordenaron paraclínicos básicos previos al inicio de dicho medicamento para documentar la existencia o no de enfermedades reumatológicas, incluidos anticuerpos nucleares extractables (ENA), anticuerpos antinucleares (ANAS), proteína c reactiva y factor reumatoide, todos estos con resultados negativos. En la tabla 4 se muestra la evolución de los valores de laboratorio encontrados en la paciente a partir del inicio de evolocumab. Cabe anotar que hasta la fecha ha recibido 5 dosis de 210 mg quincenales, y no ha presentado eventos adversos; así mismo ha permanecido en estrecho seguimiento por parte del grupo de CES Cardiología.

Revisión de tema

La hipercolesterolemia familiar es un desorden genético autosómico dominante que se caracteriza por niveles de c-LDL elevados en plasma, lo que ocasiona mayor riesgo para padecer enfermedad coronaria aterosclerótica prematura[1,3]. Este trastorno fue reconocido por primera vez en 1938 por el noruego Carl Muller, quien evidenció que existía una relación entre el nivel de c-LDL, los xantomas tendinosos y las lesiones coronarias[4].

Como consecuencia del aumento del c-LDL circulante se generan unos cambios del endotelio que llevan a lesiones ateroscleróticas, enfermedad coronaria temprana, enfermedad arterial periférica y enfermedad valvular (principalmente estenosis aórtica)[5,6]. además de acumulación de colesterol en la piel, que conduce a la formación de xantomas, particularmente en superficies tendinosas (aquiliana y extensor de los dedos), arco presenil por depósitos en la córnea y xantelasmas por depósitos de colesterol alrededor de los ojos[7,8]. Los xantomas son patognomónicos de la enfermedad y deben hacer sospechar el diagnóstico a cualquier edad[7,9]; por su parte, los xantelasmas se asocian más con enfermedad coronaria y mortalidad, de manera independiente del nivel de colesterol plasmático[10].

Dentro de la hipercolesterolemia familiar se describen dos presentaciones según su alteración genética y cigocidad: hipercolesterolemia familiar heterocigota y homocigota. En la primera se clasifican niveles de colesterol entre 350-550 mg/dl, que se han relacionado con la aparición de enfermedad coronaria en hombres menores de 55 años y mujeres menores de 60 años[7,11,12]. En la segunda, los niveles de c-LDL alcanzan valores entre 650 y 1.000 mg/dl[13] y se relacionan con muerte por causas cardiovasculares en menores de 30 años.

Adicionalmente, se ha estimado que estos pacientes pueden padecer su primer evento coronario 20 años antes que la población general (42 años vs. 64 años)[14]. En un análisis de pacientes con enfermedad heterocigota hecho por el grupo ''Simon Broome'' en 1980, era preestatinas, se documentó un aumento de mortalidad de hasta cien veces, a causa de enfermedad coronaria en jóvenes entre 20-39 años vs. población general[15].

En cuanto a su epidemiología, se ha calculado que entre 14 a 34 millones de personas en el mundo sufren hipercolesterolemia familiar, no obstante, a menos del 1% se

Cómo citar este artículo: Duque M, et al. Hipercolesterolemia familiar heterocigota en manejo con anti-PCSK9. Rev Colomb Cardiol. 2017. http://dx.doi.org/10.1016/j.rccar.2017.03.002

+Model
RCCAR 405 1–8
ARTICLE IN PRESS

Tabla 2 Cronología de tratamientos farmacológicos y no farmacológicos recibidos por la paciente

	2012	2013	2014	2015	2016
Dieta	x	x	x	x	x
Ejercicio[*]	x	x	x	x	x
Hipolipemiantes					
Lovastatina - mg	-	20	20	-	-
Atorvastatina - mg	-	20	40	-	-
Rosuvastatina - mg	-	-	20	40	40
Ezetimibe - mg	-	-	-	10	10
Colestiramina - g	-	-	-	4	4
Adherencia al tratamiento integral -%	100	100	100	100	100
c-LDL[a] *más bajo identificado - mg/dl*	245	-	211,8	207	220,7

[*] 5 veces por semana durante 2 horas cada día.
[a] c-LDL: colesterol de baja densidad.

les realiza un diagnóstico adecuado[1]. La prevalencia de la forma heterocigota en europeos es de 1 en 500 (0,20%)[13], sin embargo, un estudio más reciente de 69.016 individuos con hipercolesterolemia, familiar sugiere una prevalencia mayor: 1 por cada 200 habitantes[16]. Se han reconocido poblaciones con mayor prevalencia, entre estos africanos[1,7], canadienses, franceses (1 en 270), libaneses (1 en 85) y judíos Ashkenazi (1 en 72)[14,17]. Por su parte, la forma homocigota de la enfermedad afecta 1 de cada 1′000.000 de personas[17].

Según la etiología, se sabe que esta enfermedad envuelve tres mutaciones genéticas diferentes: del R-LDL, de la ApoB y de la PCSK9[18]. En última instancia, todas estas alteraciones genéticas generan disminución de la degradación del c-LDL plasmático mediante diferentes mecanismos. Así, por ejemplo, los R-LDL, que están en su mayoría en la superficie hepática, son responsables de remover el c-LDL circulante. Se han documentado más de 1.288 mutaciones asociadas a este receptor, 79% de las cuales pueden generar hipercolesterolemia familiar, lo que constituye la etiología más común de la enfermedad[19].

De otra parte, la ApoB es una proteína que actúa como ligando entre el c-LDL y el R-LDL; si hay mutaciones en la misma se impide la unión y, por tanto, su degradación. Esta mutación es causa del 5% de las hipercolesterolemias familiares.

El PCSK9 es el encargado de la eliminación de los R-LDL por medio de los lisosomas. Las mutaciones que aumentan la acción de esta proteína llevan a niveles disminuidos del R-LDL en los hepatocitos y en consecuencia a hiperlipidemia. Esta mutación contribuye al 1% de las hipercolesterolemias familiares[7,20].

En el caso expuesto, las tres principales mutaciones encontradas se relacionan con enfermedades que claramente aumentan el riesgo cardiovascular:

La mutación en el gen ApoA5 (cromosoma 11q23.3) se asocia con hipertrigliceridemia familiar e hiperlipoproteinemia familiar tipo 5[21]. La mutación en el gen ApoE, ubicado en el cromosoma 19q13.32, se relaciona con aterosclerosis e hiperlipoproteinemia familiar tipo 3[22]. La mutación del SCARB1, cromosoma 12q24.31, se relaciona con niveles bajos de proteínas de alta densidad (c-HDL)[23].

Tabla 3 Exámenes de laboratorio antes del inicio de los inhibidores del PCSK9

	2009	2012	2014		2015		2016	
			Inicia l	Final	Inicia l	Final	Inicia l	Fi na l
Perfil lipídico								
Colesterol total - mg/dl	254	301	258	257	270	299	283	270
c–HDL[*] - mg/dl	52	40	31	34	47	-	44,5	32
c–LDL[a] - mg/dl	187	245	214	211,8	207	237,8	220,7	222,6
Triglicéridos - mg/dl	75	78	64	56	80	76	100,9	77
Apolipoproteína B- mg/dl	-	-	143	-	-	-	202,9	122
Apolipoproteína A1- mg/dl	-	-	130	-	-	-	129	133
Lipoproteína - mg/dl	-	-	-	-	-	-	-	10,1
Perfil tiroideo								
Hormona estimulante de tiroidea - mU/ml	-	-	3,16	-	3,21	2,5	-	3,58
Tiroxina libre - ng/dl	-	-	-	-	-	-	-	0,98
Creantina quinasa total - Ui/L	-	-	-	-	-	-	-	66,41

[*] c-HDL: colesterol de alta densidad.
[a] C-LDL: colesterol de baja densidad.

Cómo citar este artículo: Duque M, et al. Hipercolesterolemia familiar heterocigota en manejo con anti–PCSK9. Rev Colomb Cardiol. 2017. http://dx.doi.org/10.1016/j.rccar.2017.03.002

+Model
RCCAR 405 1–8
ARTICLE IN PRESS

Tabla 4 Exámenes de laboratorio posteriores al inicio de los inhibidores del PCSK9 (evolocumab)

	A las 4 semanas	A las 8 semanas
Perfil lipídico		
Colesterol total - mg/dl	215,6	181
c–HDL* - mg/dl	51	48
c–LDL[a] - mg/dl	149,8	121,6
Triglicéridos - mg/dl	-	57
Perfil hepático		
Aspartato aminotransferasa - U/L	-	19
Alanino aminotransferasa - U/L	-	13
Fosfatasa alcalina - U/L	-	50,4
Gamma gutamil transferasa U/L	14	16
Perfil renal		
Creatinina - mg/dL	0,68	0,68
Nitrógeno ureico - mg/dL	11,67	13,63
Perfil tiroideo		
Hormona estimulante de tiorides - mU/mL	2,14	2,82
Tiroxina libre (T4) - ng/dL	0,95	

* c-HDL: colesterol de alta densidad.
[a] C-LDL: colesterol de baja densidad.

Hasta la fecha la hipercolesterolemia familiar es subdiagnosticada y, por ende, subtratada. Se estima que solo se reconoce el 20% de los casos[13]. Un diagnóstico adecuado debe incluir, por tanto, una combinación entre historia familiar, signos clínicos y concentración de c-LDL. Así mismo, es pertinente excluir todas las causas secundarias de hiperlipidemia (diabetes mellitus, hipotiroidismo, enfermedad hepática y renal, medicamentos, sedentarismo, entre otras)[4]. Para ello existen dos herramientas diagnósticas frecuentemente utilizadas: los criterios de ''Dutch Lipid Clinic Netwok'' y los de ''Simon Broome'', ambos combinan el nivel de LDL en plasma, historia familia y marcadores genéticos[24].

Otra forma de diagnóstico es la tamización con un individuo como caso índice, necesaria debido a la alta prevalencia, mortalidad y morbilidad de la enfermedad; es de anotar que es más efectiva cuando se realiza en familiares en primer grado de consanguinidad del caso índice[25]. Los casos índices deben ser sometidos a estudios genéticos, y la mutación hallada es la que se estudia en los familiares. Este método identifica el 50% de los casos totales de hipercolesterolemia familiar[26]. Se debe aclarar que hasta un 40% de pacientes diagnosticados con criterios clínicos no presentan una mutación genética identificable; en consecuencia, estos individuos tienden a presentar menores niveles de c-LDL y a tener mejor pronóstico[24,27].

El objetivo principal del tratamiento de la hipercolesterolemia familiar es la reducción de los eventos cardiovasculares y la mortalidad. Hasta la fecha, éste se basa en los niveles de c-LDL y no en las mutaciones genéticas específicas[3]. La Asociación Americana de Lípidos (NLA) y el Instituto Nacional para la Salud y Excelencia Clínica (NICE) del Reino Unido, recomiendan una reducción de la concentración de c-LDL mayor al 50% del nivel previo al tratamiento[13,28].

Dentro de las opciones farmacológicas, la primera línea de tratamiento para la hipercolesterolemia familiar heterocigota son las estatinas de alta intensidad (atorvastatina 80 mg/día o la rosuvastatina 40 mg/día)[29], las cuales disminuyen la enfermedad coronaria hasta un 80% si se inician como tratamiento preventivo en la adultez temprana[30]. La respuesta inicial debe ser monitorizada uno a tres meses después del inicio. Si el objetivo de disminución de más del 50% del valor inicial de c-LDL no se ha logrado después de tres meses, el paso a seguir es la adición de un segundo medicamento, en cuyo caso el más usado es el ezetimibe, que disminuye los niveles de c-LDL un 15-20% adicional y aumenta la cantidad de R-LDL en la superficie hepática[3,4,31]. Si pese a la doble terapia, tres meses después, el objetivo no se cumple, se agrega un tercer medicamento (PCSK9 o secuestrador de sales biliares)[3].

En pacientes con intolerancia a las estatinas, la combinación entre ezetimibe, niacina, secuestradores de ácidos biliares y posiblemente inhibidores PCSK9 representan la única alternativa de tratamiento[24]. Contrario a lo anterior, un estudio en 1.249 pacientes con hipercolesterolemia familiar heterocigota evidenció que de 96% de los casos tratados con estatinas, solo el 47% lograba la meta (reducción del 50% del c-LDL inicial) y sólo el 21% alacanzaba niveles de c-LDL menores de 100 mg/dl. Casi un tercio (27%) de los pacientes que no cumplían las metas ya tenían prescripción de terapia combinada con ezetimibe[4,32]. Por todo ello, en los últimos 5 años se han intentado crear nuevas tecnologías en busca de medicamentos que controlen de manera más eficiente la hipercolesterolemia familiar y sus complicaciones subsecuentes. En 2012 y 2013, respectivamente, surgieron el lomitapide (inhibidor de la proteína de transferencia de triglicéridos microsomal) y el mipomersen (inhibidor de oligonucleótido antisentido de Apo B100) para hipercolesterolemia familiar homocigota. En 2015 se

Cómo citar este artículo: Duque M, et al. Hipercolesterolemia familiar heterocigota en manejo con anti–PCSK9. Rev Colomb Cardiol. 2017. http://dx.doi.org/10.1016/j.rccar.2017.03.002

+Model
RCCAR 405 1–8
ARTICLE IN PRESS

Tabla 5 Características de los inhibidores del PCSK9 aprobados para aplicación en seres humanos

	Alirocumab	Evolocumab
Vía de administración	Subcutánea	Subcutánea
Biohabilidad	75%	82%
Indicaciones	Pacientes con Hipercolesterolemia familiar heterocigota que se encuentren con máximas dosis de estatinas y control dietario.	Pacientes mayores de 12 años con hipercolesterolemia familiar homocigota que no logran alcanzar las metas del c-LDL con las dosis máximas de estatinas tolerables y dieta.
	Pacientes con enfermedad cardiovascular aterosclerótica que requieran reducción adicional de c- LDL o que no toleran las estatinas.	Pacientes con enfermedad cardiovascular aterosclerótica que requieren reducción adicional del c-LDL Pacientes intolerantes a las estatinas
Dosis	150 mg cada 15 días	420 mg cada 4 semanas (eficacia similar con dosis de 140 mg cada 2 semanas)
Resultados hasta la fecha	Reducción del c-LDL del 50%	Reducción promedio del c-LDL del 51% al adicionarse con estatinas.
Efectos adversos	Nasofaringitis (10,5%), infecciones de vía aérea superior (9,3%), influenza (7,5), dolor lumbar (6,22%).	Nasofaringitis (11,3%), reacciones en el sitio de inyección (7.2%), influenza (5,7%), infección urinaria (4,8%)
En estudio	Eventos cardiovasculares a largo plazo (estudio ODYSSEY)	Eventos cardiovasculares a largo plazo (estudio FOURIER)

aprobaron dos inhibidores de PCSK9 (alirocumab y evolocumab) para el tratamiento de la hipercolesterolemia familiar heterocigota[33].

Los efectos clínicos del PCSK9 fueron reconocidos inicialmente por Abifadel et al., quienes en 2003, describieron en dos familias francesas mutaciones que potencian la función del gen de PCSK9. Estos pacientes tenían c-LDL elevado en asociación con aumento de enfermedad coronaria[34]. Posteriormente, estudios en animales identificaron la función del PCSK9, molécula sintetizada en el hígado que en circulación se une al R-LDL hepático y al complejo PCSK9/R-LDL, que es endocitado e internalizado en el lisosoma donde es sometido a degradación. Este proceso reduce la capacidad de remover el c-LDL circulante y se traduce en niveles aumentados del mismo[35,36]. El estudio ''Dallas Heart Study'' encontró individuos con mutaciones que disminuían la función del gen PCSK9 llevando a 28% menos c-LDL circulante que la población general[37]. Estos estudios plantearon la posibilidad de que la inhibición farmacológica del PCSK9 podría disminuir el c-LDL en pacientes con hipercolesterolemia.

Así pues, se ha logrado la inhibición del PCSK9 mediante la creación de anticuerpos monoclonales humanizados que no atraviesan la barrera hematoencefálica y que son capaces de aumentar la remoción del c-LDL circulante[33]. Al ser inyectado el anticuerpo anti-PCSK9 se une al PCSK9 circulante agotándolo rápidamente, y en consecuencia se genera menor degradación de R-LDL en compartimentos lisosomales, aumento de R-LDL en la superficie hepática y disminución de los niveles de c-LDL[35]. Se han creado tres medicamentos de este tipo: alirocumab, evolocumab y bococizumab. Los dos primeros ya están aprobados por la FDA y la EMA para uso en seres humanos. En la tabla 5 se resumen algunas características de estos dos medicamentos[38,39].

El evolocumab es un anticuerpo monoclonal humano completo tipo IgG2. Los estudios hechos hasta el momento estiman que este fármaco logra reducir en un 60% la concentración de c-LDL cuando se administra en las dosis probadas por estudios de fase III[40–42], con una disminución de los niveles circulantes de PCSK9 del 85-95% después de una semana de la administración[43].

El primer estudio de fase III que evalúo el evolocumab en pacientes con hipercolesterolemia familiar heterocigota fue el RUTHERFORD–2, multicéntrico, doble ciego, aleatorizado, placebo controlado, en el que se incluyeron 331 pacientes: un grupo recibió 210 mg de evolocumab cada 2 semanas (420 mg mensuales), y el otro placebo. Luego de 12 semanas, el grupo que habían recibido dosis quincenales de evolocumab presentaba una disminución del c-LDL entre el 59-60% comparado con el placebo. También se demostró una disminución del 15% en los triglicéridos y un incremento del 7% en el c-HDL. Al parecer, la respuesta al evolocumab es independiente del tipo de mutación que cause la hipercolesterolemia familiar heterocigota[42]. Para el caso expuesto se calcularon los valores porcentuales en los que aumentaron o disminuyeron los parámetros básicos del perfil lipídico respecto a los valores inmediatamente previos al inicio de la aplicación de evolocumab (tabla 6).

Existen reportes de efectos adversos generados por la administración del evolocumab. En los estudios de fase II y III se informa un abandono al tratamiento entre el 1,9 al 2,3% por eventos adversos[44]. Los más reportados hasta la fecha son: nasofaringitis (5,9% en el grupo de evolocumab vs. 4,8% en el grupo control), infecciones de vía respiratoria superior (3,2 vs. 2,7%), cefalea (3,0 vs. 3,2%), dolor de espalda (3,0 vs. 2,7%) y mialgias (2,5 vs. 2,6%). Entre los efectos adversos de interés se encuentran reacciones en el sitio de punción (3,3 vs. 3,0%), elevación de creatinina quinasa más de 5 veces el límite superior de normalidad (0,7 vs. 0,7%), elevaciones de ALT y/o AST más de 3 veces al límite superior de normalidad (0,4 vs. 1%) y alteraciones neurocognitivas (amnesia, delirium, desorientación y alteraciones en la memoria) (0,1 vs. 0,3%) las cuales están en estudio en el EBBINGHAUS[44,45]. Entre 2-8% de los pacientes en los estudios descontinuaron el medicamento, sin claridad sobre la

Cómo citar este artículo: Duque M, et al. Hipercolesterolemia familiar heterocigota en manejo con anti-PCSK9. Rev Colomb Cardiol. 2017. http://dx.doi.org/10.1016/j.rccar.2017.03.002

+Model
RCCAR 405 1–8
ARTICLE IN PRESS

Tabla 6 Comportamiento de los parámetros básicos del perfil lipídico de la paciente posterior al inicio de los inhibidores del PCSK9

	A las 4 semanas	A las 8 semanas
Colesterol total - mg/dl	↓20,14	↓32,96
c-HDL* - mg/dl	↑59,3	↑50
c-LDL[a] - mg/dl	↓32,7	↓44,92
Triglicéridos - mg/dl	-	↓25,97

* c-HDL: colesterol de alta densidad.
[a] c-LDL: colesterol de baja densidad.

causa de ésta. Así mismo, por la composición del inhibidor de PCSK9 (anticuerpo humanizado) se pueden generar anticuerpos contra el inhibidor de PCSK9, pero hasta ahora no se han demostrado con frecuencia y no han disminuido la eficacia del medicamento[33].

Conclusiones

Pese a que la hipercolesterolemia familiar no se presenta en un gran porcentaje de la población, merece ser sospechada y diagnosticada, a raíz de las consecuencias cardiovasculares significativas que genera. Por tanto, su detección temprana permite impactar en intervenciones oportunas.

En la actualidad, los inhibidores del PCSK9 se han convertido en una alternativa prometedora para aquellos que no responden a las terapias convencionales y cursan con enfermedades tan severas como la hipercolesterolemia familiar. Por ahora, dichos inhibidores, especialmente, el evolocumab en el tratamiento de la hipercolesterolemia familiar heterocigota, parecen ser una terapia prometedora en términos de disminución de valores de colesterol total, c-LDL y triglicéridos y aumento de c-HDL. Faltan estudios y seguimientos muy estrechos y objetivos en el tiempo, que corroboren o contradigan los beneficios y eventos adversos encontrados hasta el momento en los pacientes que se someten a estas nuevas opciones terapéuticas.

Responsabilidades éticas

Protección de personas y animales. Los autores declaran que para esta investigación no se han realizado experimentos en seres humanos ni en animales.

Confidencialidad de los datos. Los autores declaran que en este artículo no aparecen datos de pacientes.

Derecho a la privacidad y consentimiento informado. Los autores declaran que en este artículo no aparecen datos de pacientes.

Financiación

Ninguna.

Conflicto de intereses

Ninguno.

Bibliografía

1. Nordestgaard BG, Chapman MJ, Humphries SE, Ginsberg HN, Masana L, Descamps OS, et al. Familial hypercholesterolaemia is underdiagnosed and undertreated in the general population: guidance for clinicians to prevent coronary heart disease: consensus statement of the European Atherosclerosis Society. Eur Heart J [Internet]. 1 de diciembre de 2013 [Acceso 7 Nov 2016];34(45). Disponible en: http://search.ebscohost.com/login.aspx?direct=true&db=mdc&AN=23956253&site=eds-live
2. Strath SJ, Kaminsky LA, Ainsworth BE, Ekelund U, Freedson PS, Gary RA, et al. Guide to the assessment of physical activity: clinical and research applications: a scientific statement From the American Heart Association. Circulation. 2013;128:2259–79.
3. Gidding SSM, Ann Champagne MR, de Ferranti SDM, Defesche J, Ito MKP, Knowles JW, et al. The agenda for familial hypercholesterolemia: a scientific statement from the American Heart Association. [Miscellaneous Article]. Circulation. 2015;132:2167–92.
4. Hovingh GK, Davidson MH, Kastelein JJP, O'Connor AM. Diagnosis and treatment of familial hypercholesterolaemia. Eur Heart J. 2013;34:962–71.
5. Kroon AA, Ajubi N, van Asten WN, Stalenhoef AF. The prevalence of peripheral vascular disease in familial hypercholesterolaemia. J Intern Med. 1995;238:451–9.
6. Kolansky DM, Cuchel M, Clark BJ, Paridon S, McCrindle BW, Wiegers SE, et al. Longitudinal evaluation and assessment of cardiovascular disease in patients with homozygous familial hypercholesterolemia. Am J Cardiol. 2008;102:1438–43.
7. Cartier JL, Goldberg AC. Familial hypercholesterolemia: advances in recognition and therapy. Prog Cardiovasc Dis. 2016;59:125–34.
8. Nemati MH, Astaneh B. Optimal management of familial hypercholesterolemia: treatment and management strategies. Vasc Health Risk Manag. 2010;6:1079–88.
9. Oosterveer DM, Versmissen J, Yazdanpanah M, Hamza TH, Sijbrands EJG. Differences in characteristics and risk of cardiovascular disease in familial hypercholesterolemia patients with and without tendon xanthomas: a systematic review and meta-analysis. Atherosclerosis. 2009;207:311–7.
10. Christoffersen M, Frikke-Schmidt R, Schnohr P, Jensen GB, Nordestgaard BG. Tybjærg-Hansen A. Xanthelasmata, arcus corneae, and ischaemic vascular disease and death in general population: prospective cohort study. BMJ. 2011;343:d5497.
11. Goldstein JL, Schrott HG, Hazzard WR, Bierman EL, Motulsky AG. Hyperlipidemia in coronary heart disease. II. Genetic analysis of lipid levels in 176 families and delineation of a new inherited disorder, combined hyperlipidemia. J Clin Invest. 1973;52:1544–68.
12. Raal FJ, Santos RD. Homozygous familial hypercholesterolemia: current perspectives on diagnosis and treatment. Atherosclerosis. 2012;223:262–8.

Cómo citar este artículo: Duque M, et al. Hipercolesterolemia familiar heterocigota en manejo con anti-PCSK9. Rev Colomb Cardiol. 2017. http://dx.doi.org/10.1016/j.rccar.2017.03.002

13. Goldberg AC, Hopkins PN, Toth PP, Ballantyne CM, Rader DJ, Robinson JG, et al. Familial hypercholesterolemia: screening, diagnosis and management of pediatric and adult patients: clinical guidance from the National Lipid Association Expert Panel on Familial Hypercholesterolemia. J Clin Lipidol. 2011;5 3 Suppl:S1–8.
14. Stone NJ, Levy RI, Fredrickson DS, Verter J. Coronary artery disease in 116 kindred with familial type II hyperlipoproteinemia. Circulation. 1974;49:476–88.
15. Risk of fatal coronary heart disease in familial hypercholesterolaemia. Scientific Steering Committee on behalf of the Simon Broome Register Group. BMJ. 1991; 303(6807):893-6.
16. Benn M, Watts GF, Tybjaerg-Hansen A, Nordestgaard BG. Familial hypercholesterolemia in the danish general population: prevalence, coronary artery disease, and cholesterol- lowering medication. J Clin Endocrinol Metab. 2012;97:3956–64.
17. Austin MA, Hutter CM, Zimmern RL, Humphries SE. Genetic causes of monogenic heterozygous familial hypercholesterolemia: a HuGE prevalence review. Am J Epidemiol. 2004;160:407–20.
18. Soutar AK, Naoumova RP. Mechanisms of disease: genetic causes of familial hypercholesterolemia. Nat Clin Pract Cardiovasc Med. 2007;4:214–25.
19. Usifo E, Leigh SEA, Whittall RA, Lench N, Taylor A, Yeats C, et al. Low-density lipoprotein receptor gene familial hypercholesterolemia variant database: update and pathological assessment. Ann Hum Genet. 2012;76:387–401.
20. Cuchel M, Bruckert E, Ginsberg HN, Raal FJ, Santos RD, Hegele RA, et al. Homozygous familial hypercholesterolaemia: new insights and guidance for clinicians to improve detection and clinical management. A position paper from the Consensus Panel on Familial hypercholesterolaemia of the European Atherosclerosis Society. Eur Heart J. 2014;35:2146–57.
21. APOA5[gene] - ClinVar - NCBI [Internet]. [Acceso 21 Feb 2017]. Disponible en: https://www.ncbi.nlm.nih.gov/clinvar/?term=APOA5%5Bgene%5D
22. APOE[gene] - ClinVar - NCBI [Internet]. [Acceso 21 Feb 2017]. Disponible en: https://www.ncbi.nlm.nih.gov/clinvar/?term=APOE%5Bgene%5D
23. SCARB1[gene] - ClinVar - NCBI [Internet]. [Acceso 21 Feb 2017]. Disponible en: https://www.ncbi.nlm.nih.gov/clinvar/?term=SCARB1%5Bgene%5D
24. Hughes DP, Viljoen A, Wierzbicki AS. Familial hypercholesterolaemia in the era of genetic testing. Curr Cardiol Rep. 2016;18:42.
25. Starr B, Hadfield SG, Hutten BA, Lansberg PJ, Leren TP, Damgaard D, et al. Development of sensitive and specific age- and gender-specific low-density lipoprotein cholesterol cutoffs for diagnosis of first-degree relatives with familial hypercholesterolaemia in cascade testing. Clin Chem Lab Med. 2008;46:791–803.
26. Wonderling D, Umans-Eckenhausen MAW, Marks D, Defesche JC, Kastelein JJP, Thorogood M. Cost-effectiveness analysis of the genetic screening program for familial hypercholesterolemia in The Netherlands. Semin Vasc Med. 2004;4:97–104.
27. Clarke REJ, Padayachee ST, Preston R, McMahon Z, Gordon M, Graham C, et al. Effectiveness of alternative strategies to define index case phenotypes to aid genetic diagnosis of familial hypercholesterolaemia. Heart Br Card Soc. 2013;99:175–80.
28. Qureshi N, Humphries SE, Seed M, Rowlands P, Minhas R, NICE Guideline Development Group. Identification and management of familial hypercholesterolaemia: what does it mean to primary care? Br J Gen Pract J R Coll Gen Pract. 2009;59:773–6.
29. Smilde TJ, van Wissen S, Wollersheim H, Trip MD, Kastelein JJ, Stalenhoef AF. Effect of aggressive versus conventional lipid lowering on atherosclerosis progression in familial hypercholesterolaemia (ASAP): a prospective, randomised, double-blind trial. Lancet. 2001;357:577–81.
30. Versmissen J, Oosterveer DM, Yazdanpanah M, Defesche JC, Basart DCG, Liem AH, et al. Efficacy of statins in familial hypercholesterolaemia: a long term cohort study. BMJ. 2008;337:a2423.
31. Cannon CP, Blazing MA, Giugliano RP, McCagg A, White JA, Theroux P, et al. Ezetimibe Added to Statin Therapy after Acute Coronary Syndromes. N Engl J Med. 2015;372:2387–97.
32. Pijlman AH, Huijgen R, Verhagen SN, Imholz BPM, Liem AH, Kastelein JJP, et al. Evaluation of cholesterol lowering treatment of patients with familial hypercholesterolemia: A large cross-sectional study in The Netherlands. Atherosclerosis. 2010;209:189–94.
33. Wierzbicki AS, Grant P. Drugs for hypercholesterolaemia–from statins to pro-protein convertase subtilisin kexin 9 (PCSK9) inhibition. Clin Med. 2016;16:353–7.
34. Abifadel M, Varret M, Rabès JP, Allard D, Ouguerram K, Devillers M, et al. Mutations in PCSK9 cause autosomal dominant hypercholesterolemia. Nat Genet. 2003;34:154–6.
35. Stein EA, Mellis S, Yancopoulos GD, Stahl N, Logan D, Smith WB, et al. Effect of a monoclonal antibody to PCSK9 on LDL cholesterol. N Engl J Med. 2012;366:1108–18.
36. Shimada YJ, Cannon CP. PCSK9 (Proprotein convertase subtilisin/kexin type 9) inhibitors: past, present, and the future. Eur Heart J. 2015;36:2415–24.
37. Cohen J, Pertsemlidis A, Kotowski IK, Graham R, Garcia CK, Hobbs HH. Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9. Nat Genet [Internet]. 1 de febrero de 2005 [Acceso 16 Nov 2016];37(2). Disponible en: http://search.ebscohost.com/login.aspx?direct=true&db=mdc&AN=15654334&site=eds-live
38. Merchán A, Ruiz ÁJ, Campo R, Prada CE, Toro JM, Sánchez R, et al. Hipercolesterolemia familiar: artículo de revisión. Rev Colomb Cardiol. 2016;23:4–26.
39. Gupta S. Development of proprotein convertase subtilisin/kexin type 9 inhibitors and the clinical potential of monoclonal antibodies in the management of lipid disorders. Vasc Health Risk Manag. 2016;12:421–33.
40. Koren MJ, Lundqvist P, Bolognese M, Neutel JM, Monsalvo ML, Yang J, et al. Anti-PCSK9 monotherapy for hypercholesterolemia: the MENDEL-2 randomized, controlled phase III clinical trial of evolocumab. J Am Coll Cardiol. 2014;63:2531–40.
41. Stroes E, Colquhoun D, Sullivan D, Civeira F, Rosenson RS, Watts GF, et al. Anti-PCSK9 antibody effectively lowers cholesterol in patients with statin intolerance: the GAUSS-2 randomized, placebo-controlled phase 3 clinical trial of evolocumab. J Am Coll Cardiol. 2014;63:2541–8.
42. Raal FJ, Stein EA, Dufour R, Turner T, Civeira F, Burgess L, et al. PCSK9 inhibition with evolocumab (AMG 145) in heterozygous familial hypercholesterolaemia (RUTHERFORD- 2): a randomised, double-blind, placebo-controlled trial. Lancet. 2015;385:331–40.
43. Blom DJ, Hala T, Bolognese M, Lillestol MJ, Toth PD, Burgess L, et al. A 52-week placebo- controlled trial of evolocumab in hyperlipidemia. N Engl J Med. 8 de mayo de. 2014;370:1809–19.
44. Keating GM. Evolocumab: a review in hyperlipidemia. Am J Cardiovasc Drugs Devices Interv. 2016;16:67–78.
45. Toth PP, Sattar N, Genest J, Descamps OS, Dent R, Djedjos C, et al. A comprehensive safety analysis of 6026 patients from phase 2 and 3 short and long term clinical trials with evolocumab (AMG 145). J Am Coll Cardiol [Internet]. 2015 [Acceso 17 Nov 2016];65(10_S). Disponible en: http://content.onlinejacc.org/article.aspx?articleid=2198685

Cómo citar este artículo: Duque M, et al. Hipercolesterolemia familiar heterocigota en manejo con anti–PCSK9. Rev Colomb Cardiol. 2017. http://dx.doi.org/10.1016/j.rccar.2017.03.002

Chapter 14

Concluding remarks and perspectives

Concluding remarks and perspectives

This thesis has taught me the valuable skill of understanding genomic content and structure using NGS data and how these play an important role in clinical applications. The initial work on fungal genome assemblies marked an important milestone for our line of research. We were able to attain a deep level of understanding of the bioinformatic limits of using NGS reads and address complex tasks involved in choosing de novo assemblies. One of the main reasons why we were interested in improving our understanding of the structure and content of genomic assemblies was to be able to provide the scientific community with high quality novel genome sequence assemblies of the fungal pathogens *B. dermatitidis*, *Paracoccidioides* spp., *Emmonsia crescens* and *Emmonsia parva,* as well as their annotations, which are presented in the work we published. We were also able to update the genome assemblies and annotations of three reference strains of *Paracoccidioides* spp, which were published and made publically available to the scientific community. These results have provided the medical mycology community with new and updated fungal genomes that should facilitate comparative genomics projects requiring high quality of assemblies and annotations, such as those addressing gene presence/absence, SNP analysis of multiple species and/or strains and evolutionary analyses. For example, the sequences of the *Emmonsia* spp. and *Blastomyces dermatitidis* and *Paracoccidioides* spp., described and presented in this thesis, were used in a recent publication where the authors present evidence to justify the introduction of a novel taxon within the Ajellomycetaceae (Dukik et al,.2017).

Another contribution for medical mycology and clinical laboratories was to devise and test ways to systematically design primer sets for the detection of fungal pathogens

from aligned genome sequences, and to analytically validate primers we obtained for *H. capsulatum* and *Paracoccidioides* spp. Here we provide the scientific community with primer sets that can be tested using clinical samples in order to validate their application in clinical settings. The algorithm we created will be made available to the scientific community so that other groups can utilize this tool to design primer pairs for the molecular detection of other pathogens. The algorithm is now being used by collaborators to design primer pairs for the detection of *Sporothrix* spp., with promising results. We feel that this contribution may help the scientific community bridge gaps between NGS sequencing and molecular detection assays that can be implemented in laboratories.

As newer sequencing technologies are developed and costs decrease, the use of genomics in the clinical setting will become more frequent. For the detection of pathogens, more information of unique genomic signatures will be required, and the generation of large databases will permit the rapid differential diagnosis of pathogens. Such molecular detection protocols would be of high importance also in tropical areas such as Colombia where many of the world's important pathogens are endemic and cause a high burden of health care costs.

Human genomics in Colombia is only recently starting to become incorporated into clinical settings. The results we have obtained, but not yet published, in our study of 9p21.3 and its association with cardiovascular disease, blood pressure / hypertension and type 2 diabetes are highly significant for understanding the genetic risk component also of other diseases associated with this locus, including cancers. The results, when completed, will provide valuable information of public health relevance, and improved understanding of such genetic risk will help in decision making, for example in preventative health programs for individuals with a higher risk of developing cardiovascular disease.

A whole exome approach on which we are working for mutation analysis of various diseases is also showing promise. If most coding regions of the human genome could be analyzed by a single test, this would provide clinicians with a top-down approach that could aid in the correct diagnosis and treatment of patients with genetic diseases. We have now tested our approach in more than 90 patients with various genetic diseases including cardiovascular, neurological and lysosomal diseases and cancer. We have seen that is it possible to accurately detect pathogenic mutations as well as variants of unknown significance (VUS; work in progress).

Our analyses for a patient with familial hypercholesterolemia presented in chapter 13 shows a clear example of how genomics in the clinical setting is bridging gaps in precision medicine in Colombia. The patient, who was affected by mutations in genes related to the cholesterol processing pathway that were the most probable cause of her disease, was also a non-responder to conventional cholesterol lowering drugs.

Genomics is forecasted to grow in the clinical setting in the upcoming years. We foresee the future of genomics in medicine heading in a direction of not only using NGS data for diagnostic purposes but also gene editing of pathogenic mutations. Metagenomic sequencing of clinical samples is becoming more frequent. Important advances in clinical genomics such as liquid biopsy for tumor sample characterization will be highly useful for non-biopsiable or other regions of the body that are not easy to reach. As an example, circulating tumor DNA extraction methods from peripheral blood are now allowing for genomic analyses of tumors without highly invasive procedures. For brain stem tumors, circulating tumor DNA is obtained directly from cerebrospinal fluid allowing tumor classification.

It is our hope that the results, insights and resources contributed in this thesis will help expand the use of clinical genomics for the detection of microorganisms and human chronic disease in the near future.

References

1. Dukik K, Muñoz JF, Jiang Y, Feng P, Sigler L, Stielow JB, et al. Novel taxa of thermally dimorphic systemic pathogens in the Ajellomycetaceae (Onygenales). Mycoses. 10 ed. 2017 May;60(5):296–309.