Candidate Gene Discovery in Autoimmunity by Using Extreme Phenotypes, Next Generation Sequencing and Whole Exome Capture

Angad Johar [a], Juan-Manuel Anaya [b*], Dan Andrews [c], Hardip R. Patel [d], Matthew Field [c], Chris Goodnow [c], Mauricio Arcos-Burgos [a*]

[a] Genomics and Predictive Medicine, Genome Biology Department, John Curtin School of Medical Research, ANU College of Medicine, Biology & Environment, The Australian National University, Canberra, ACT, Australia.

[b] Centre for Autoimmune Diseases Research (CREA), School of Medicine and Health Sciences, Universidad del Rosario, Bogota, Colombia.

[c] Immunogenomics and Bioinformatics Group, Immunology Department, John Curtin School of Medical Research, ANU College of Medicine, Biology & Environment, The Australian National University, Canberra, ACT, Australia;

[d] Genome Discovery Unit, Genome Biology Department, John Curtin School of Medical Research, ANU College of Medicine, Biology & Environment, The Australian National University, Canberra, ACT, Australia.

[#] **Corresponding authors**

Tel.: +61 2 61259396, E-mail address: Mauricio.Arcos-Burgos@anu.edu.au (M Arcos-Burgos)

Tel.: +57 3212339828, E-mail address: juan.anaya@urosario.edu.co (J-M. Anaya)

The authors have indicated they have no financial interest to disclose.

Keywords: Next Generation Sequencing, Multiple Autoimmune Syndrome, Whole Exome Sequencing, And Whole Genome Sequencing.

Abstract

Whole exome sequencing (WES) is a widely used strategy for detection of protein coding and splicing variants associated with inherited diseases. Many studies have shown that the strategy has been broad and proficient due to its ability in detecting a high proportion of disease causing variants, using only a small portion of the genome. In this review we outline the main steps involved in WES, the comprehensive analysis of the massive data obtained including the genomic capture, amplification, sequencing, alignment, curating, filtering and genetic analysis to determine the presence of candidate variants with potential pathogenic/functional effect. Further, we propose that the Multiple Autoimmune Syndrome (MAS), an extreme phenotype of autoimmune disorders, is a very well suited trait to tackle genomic variants of major effect underpinning the lost of self-tolerance.

1. Introduction

A growing body of evidence support the involvement of rare variants (population allele frequency < 1%) in the aetiology of common diseases. It is possible that much of the genetic control of common diseases is due to rare and pathogenic variants with a major effect on the phenotype [1-3]. The detection of these rare genome variants harboured in coding regions has shown to be successfully achievable using extreme phenotypes and pedigrees segregating exceptional phenotypes [1-4].

Whole exome sequencing (WES) is a cost effective technique that employs high throughput capture by hybridisation techniques, using exon specific oligonucleotides to enrich only protein coding sequences that can be later used for sequencing [5]. The WES is rapidly becoming the first-line approach for monogenic disorders, and an alternative one for dissecting extreme phenotypes of complex inherited conditions [4,6]. Its rationale is based on the fact that gene variants located in exons are more likely to be pathogenic than those located in introns or between genes [1,4,7].

The WES of pedigrees is a highly effective approach for identifying homozygous, compound heterozygous, novel, germinal, and *de novo* rare coding sequence variants. This is because multiple rare sequence variants occurring within a specific gene (or within a gene family or pathway) are extremely implausible events [6,7]. This concept might be applicable, with some restrictions, to WES of sporadic cases of very unequivocal and conspicuous phenotypes [2].

Even though some estimates suggest that the success rate of uncovering variants that account for Mendelian disorders via exome sequencing is only 25%, this strategy is still arguably efficient as disease-causing protein coding variation can be located using less than 2% of the human genome [4,7]. The comparatively low cost of this technology means that exome databases are rapidly expanding [7]. Therefore, there is a large abundance of this type of data given the access to large numbers of publicly available exome sequences which allows the comparison of frequencies, as well as the identification of *de novo* variants and the matching of cases and controls by ethnicity to avoid genetic stratification. Several manuscripts have reported the identification of candidate genes for several Mendelian and complex traits [2,8-11].

Nevertheless, one must be aware of the fact that exome sequencing has deficiencies. This is because hybridisation probes are not available for all annotated exons within the gold standard databases. Also, exome sequencing will not be able to detect mutations in non-coding DNA that alter gene function by various regulatory mechanisms and enhancer effects. Such variants (in recent times) are emerging as important contributors to genetic disease and they occur in >98% of the human genome, which is missed by exome capture [12]. For sequencing these non-coding regions (either intronic or inter-genic) it would be necessary whole genome sequencing (WGS).

Polyautoimmunity is defined as the presence of more than one AD in a single patient [13]. When three or more ADs coexist, this condition is called multiple autoimmune syndrome (MAS), which represents the best example of polyautoimmunity as well as the effect of a single genotype on diverse autoimmune phenotypes [14].

MAS identifies a form of extreme autoimmune disorder, frequently clustering in families [14]. The MAS running in families often displays Mendelian segregation ratios and consequently represents a very potential powerful tool for identifying major genes commonly underpinning the development of autoimmunity [14]. Several pedigrees clustering polyautoimmunity and autoimmune syndromes that were ascertained from probands affecteds with MAS has been described [14].

We think that these pedigrees as well as sporadic cases of MAS would be critical for dissecting genes of major effect conferring susceptibility to autoimmunity. Given that we have already demonstrated the existence of major effects and the potential location of these MAS loci, in this review we present a comprehensive and practical strategy of the use of WES to map genes implicated in extreme autoimmune phenotypes (i.e., polyautoimmunity and MAS).

2. Whole Exome Capture and Sequencing

In general, the next generation sequencing process works by fragmenting genomic DNA using sonication or mechanical shearing. The formed ends are adenylated and adaptor oligonucleotides are added to the ends of these adenylated fragments. These adaptors are short oligonucleotides of known sequences for universal priming of both amplification and sequencing steps [5]. Commonly, the fragments are enriched for specific genes of interest (targeted sequencing), or for all coding regions (whole exome capture for WES), in a physical capture step. That enrichment is not needed for WGS and all fragments are sequenced [5].

Prior to the sequencing process, fragments are separated and clonally amplified by PCR. To do that, single strands of these fragments are then hybridised to oligo-primers on the genome analyser flow cell, beginning the process of cluster generation. After hybridisation the oligo-primers are then extended by polymerases, generating complementary strands bound to the same surface. The double stranded DNA is then denatured and only newly synthesised strand remains as the original template is removed after washing. The single strand molecule, still bound to the flow cell, then hybridises to adjacent oligo-primers and amplification produces a double strand 'bridged' molecule.

Strands are then denatured and the bridge amplification is repeated, eventually generating millions of fragments. Removal of the reverse strands then leaves the forward strands for sequencing primers to hybridise to the oligo sequence on the template strand. During each cycle polymerases then extend the newly synthesised strand and in doing so incorporate one of 4 fluorescently labelled terminator nucleotides (ddNTP), that each have a blocking group on the 3 prime end. Excitation of the fluorescent dye by a laser enables identification of the base. Cleavage of the dye and the blocking group nucleotide from the fluoresecent label, allows continuation of subsequent sequencing cycles (Fig. 1).

In terms of accuracy, Sanger sequencing is still considered by many as the gold standard amongst sequencing technologies that are currently available. Unlike NGS, the Sanger method doesn't consist of mass cluster generation steps whilst samples are being prepared for sequencing. NGS is more likely to generate accumulated errors due to numerous amplification cycles in the bridge PCR phase [15]

Sanger sequencing also has the advantage of longer and more contiguous read lengths than most NGS technologies, thereby increasing its capability to accurately align reads and identify SNPs even in the presence of long repeat sequences [16]. Consequently it is

imperative that we perform resequencing using the Sanger method, in order to check the validity of the chosen candidate genes and their variants.

Another problem observed with NGS is that it is prone to sequence specific errors that result of secondary structure formation on single stranded DNA molecules bound to the genome analyser flow cell [17] (Nakamura et al 2011). Sequence-specific error profile of Illumina sequencers. Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, Shiwa Y, Ishikawa S, Linak MC, Hirai A, Takahashi H, Altaf-Ul-Amin M, Ogasawara N, Kanaya S. Nucleic Acids Res. 2011 Jul;39(13):

Examples of these include inverted repeats and GGC repetitive sequences. Inverted repeats inhibit bidirectional nucleotide elongation on the complementary strand for several cycles prior to the position of where the repeat sequence is present. This is known as lagged sequence contamination. With each chain reversible termination cycle, there is a delay in the addition of fluorescently labelled nucleotides to the synthesised strand. As the strand continues to grow, the secondary structure's hydrogen bonds become destabilized and the reversible termination cycles for each sequence eventually resumes at normal speed. However by this stage of the sequencing phase, it is likely that many reads proximal to the repeat region will falsely report a variant base due to the effects of lagged sequence contamination in previous cycles [17]. (Nakamura et al 2011, see above).

GGC repeat sequences have a similar effect, but it is hypothesised that this sequence repeat preferentially binds to DNA polymerase to inhibit the synthesis process. However the exact mechanism is not well understood. Nevertheless it was also found in the same study [17] (Nakumara et al 2011, see above), that 10% sequence specific errors were not associated with these inverted repeats or GGC repeats. Thus other secondary structures could also contribute to sequencing errors and inaccurate variant calls. This provides extra incentive to conduct resequencing as a means of follow up SNP validation. Also it is essential to look towards improved NGS technologies and advanced variant calling software to mitigate these problems.

3. Bioinformatics Analysis

DNA sequence of each individual obtained from exome capture is processed through a variant calling pipeline containing custom PERL scripts developed by the Immunogenomics Lab Bioinformatics team. Sequence reads are aligned to the latest version of the reference human genome, using the alignment algorithm (BWA or Burrrows Wheel Aligner) developed by [18]. The output alignment file and its index are compressed into a binary file. Often during the PCR amplification stage of library preparation, duplicate reads are produced as a result of amplification bias, i.e. multiple copies of the same DNA fragment can be placed on different primers along the flow cell. Assuming that these reads contain the same sequence content and align to the same region, they can be selectively removed by the pipeline's algorithms. This is done to ensure that variants are not called as a result of these sequencing artefacts [19] (Fig. 2).

The output of the BWA alignment step (which includes the alignment index and alignments in binary format) is used as a template by a set of bioinformatics utilities (known as SAMtools) for alignment viewing and variant calling [20]. Variants are filtered by the pipeline in accordance to the following criteria:

- Variants in the sequenced exomes that overlap dbSNP variants are annotated along with the dbSNP population frequency.
- The SNP score is an indication of the level of confidence that a variant is present in a given nucleotide position. An arbitrary cut-off score of 40 is set for this purpose.
- The variants are then overlapped with the ENSEMBL database, in order to identify which variants fall within coding exons and splice sites (defined as 10bp outside the exon boundary). Based on previous empirical data, it has been indicated that splice mutations within 10bp from exons have a lower density of SNPs. This suggests that there may be high levels of sequence conservation in these genomic regions as a result

of selection pressure [21] (Fairbrother et al 2004). Therefore, splice variants 1-10 bases from exon boundaries are also considered as potentially causative mutations, along with non-synonymous SNPs.

Genetic variants are filtered and prioritised using a heuristic system using several tools calculating the amino acid substitution effects on the structure and function of the protein, for all non-synonymous variants within genes. Using many algorithms from externally available databases, each variant is scored and classified, based on the protein region in which the amino acid substitution has taken place. Some of these algorithms include:

3.1. Uniprot

When a query amino acid sequence of a protein is submitted to Uniprot (http://www.uniprot.org), the software can search the uniprot database annotations, which indicate whether the substitution took place at a transmembrane domain, carbohydrate molecule, lipid side chain, etc. [22].

3.2. PSIC (Position Specific Independent Counts) profiles

PSIC computes the likelihood ratio that any given variant amino acid is likely to be found at a particular position in the protein vs. the likelihood of observing the same amino acid at any other position (http://www.imb.ac.ru/PSIC) [23]. This calculation is based on the sequence conservation of protein sequences determined after PolyPhen2 aligns to and identifies homologous sequences from BLAST (Basic Local Alignment Search Tool) search of the UniRef 100 database. PSIC profiles are generated for sequence homologues that are longer than 50 residues in length and have a sequence identity of 30-94%. Large differences in the PSIC probability profile scores between the wild type amino acid and the variant amino acid, indicate that substitutions in a given protein region are rare. This indicates that amino acid

sequences in the protein have a high level of conservation, suggesting that substitutions in these regions are likely to be deleterious [22,23].

3.3. **Ramachadran Plots**.

These plots measure the change in the dihedral angle within sections of the 3D structure of the protein as a result of an amino acid substitution. The change in dihedral angles (as a result of amino acid substitutions) in these plots are measured the C' (carbonyl carbon)-N-C^a (alpha carbon)-C' backbone and the N-C^a –C'-N backbone in the protein's secondary structure [24]. Information about the 3D protein structures can be obtained from the Dictionary of Secondary Structure Proteins (DSSP) [25].

3.4. PolyPhen (http://genetics.bwh.harvard.edu/pph2/) **and SIFT (Sorting Intolerant from Tolerant) scores (http://sift.bii.a-star.edu.sg).**

These tools are similar in the sense that are designed to achieve the same outcome, and rely on sequence conservation. They use a normalized position specific scoring matrix, i.e. it calculates probability that the amino acid, will change states to another particular amino acid, based on the level of conservation in the protein family. Those with low transition probabilities, due to highly conserved residues are predicted to be deleterious [26]. However, unlike PolyPhen, SIFT doesn't implement information from the 3D or secondary structures of proteins.

Thus using all these incorporated criteria, we scored the amino acid substitution and then prioritised as benign, tolerated possibly damaging, deleterious or probably damaging. Variants with 'benign' and 'tolerated; amino acid changes can be excluded during the filtration process.

3.5. **Variant Population Frequency Obtained from the dbSNP Database.**

The population frequencies of variants (SNPs and INDELs) annotated in the variant calling pipeline were obtained from the dbSNP archives. As mentioned earlier, this database contains known genetic variation, obtained from genome sequencing and variant haplotyping of individuals from different ethnic groups. Hence the dbSNP population frequencies for all SNPs and INDELs obtained from the variant calling pipeline were used as a quantitative guide for identifying variants that were rare and common amongst the worldwide population. Given the nature of the design that uses extreme phenotypes, common variants are discarded, as the effect size provided by them must be small.

4. **Network Building Algorithms and Pathway Analysis for Filtered Gene List**

After applying the filtration steps described above, the refined gene list was then used as an input source for functional network and pathway analysis algorithms as implemented in Metacore®. In it, the network and pathway analysis algorithms are available through a web interface, and the software suite also includes a manually curated gene ontology database. The algorithm to build these networks incorporates our input gene list into a single dense network. This is known as a 'global network', which is then divided into biologically functional sub-networks. Within these networks, each node (connected by 2 or more genes) is represented by a subset of gene ontology processes. These can be used in a heuristic manner to identify genes with important functions in autoimmunity.

The ontology terms for a gene (which is connected to other genes via network nodes) within the networks are prioritized by the likelihood of overrepresentation that are calculated based on the size of the intersection between the input gene list and the network process in question. In other terms, the statistical probability that a given number of genes from the input gene list would randomly overlap with a particular GeneGo ontology process.

The only nodes used for statistical evaluations only include those with direct physical interactions with our input dataset (i.e. our gene list generated from the filtration strategies described above). This is done to help minimize artefacts in the statistical analysis, which can arise from genes in the database, which may be in the same network, but have no functional connection or interaction with any gene from the input list.

5. Conclusions

Whole-exome capture and sequencing analysis is a time and resource-intense endeavor. Currently, we employ software that allows rapid selection of any genetic variant according to variant type, novelty (via screening public and private databases), and predicted protein effect. However, linking these results to phenotypic manifestations in a particular person is currently performed by a mixture of manual analysis using a number of additional databases (e.g., Human Genome Mutation Database, OMIM, PubMed, and UCSC, among others). We built on existing analytic tools in order to rapidly detect and annotate genomic variants associated with human disease. We are aware that analytical criteria for filtering need to be flexible and up-to-date; therefore, we undertook a systematic upgrade and iterative processes of the databases evaluation by considering each filter.

JuanMa: Can you introduce something about MAS and its importance.

Take Home Messages

1. Based on data from recent studies, exome capture can be considered as a feasible and effective strategy to detect potentially causative variants in autoimmunity.
2. Variant detection is enhanced by Next Generation Sequencing technology , which provides enhanced sequencing chemistry and parallelization of DNA samples.

3. In addition to improvements in sequencing, more powerful bioinformatics tools are available to determine the association with and potential functional significance of candidate genetic variants in autoimmunity.

References

1. Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat Rev Genet. 2010 Jun;11(6):415-25.
2. Paz-Filho G, Boguszewski MC, Mastronardi CA, et al. Whole exome sequencing of extreme morbid obesity patients: translational implications for obesity and related disorders. *Genes* 2014; 5(3): 709-25.
3. Castiblanco J, Arcos-Burgos M, Anaya JM. What is next after the genes for autoimmunity? *BMC Med* 2013; 11: 197.
4. Yang Y, Muzny DM, Reid JG, et al. Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *N Engl J Med* 2013; 369(16): 1502-11.
5. Schnekenberg RP, Nemeth AH. Next-generation sequencing in childhood disorders. *Arch Dis Child* 2014; 99(3): 284-90.
6. Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, et al. Exome sequencing as a tool for Mendelian disease gene discovery. Nat Rev Genet. 2011 Sep 27;12(11):745-55.
7. MacArthur DG, Manolio TA, Dimmock DP, Rehm HL, Shendure J, Abecasis GR, et al. Guidelines for investigating causality of sequence variants in human disease. *Nature* 2014; 508(7497): 469-76.
8. Liew WK, Ben-Omran T, Darras BT, Prabhu SP, De Vivo DC, Vatta M, et al. Clinical application of whole-exome sequencing: a novel autosomal recessive spastic ataxia of Charlevoix-Saguenay sequence variation in a child with ataxia. *JAMA neurology* 2013; 70(6): 788-91.
9. Need AC, Shashi V, Hitomi Y, Schoch K, Shianna KV, McDonald MT, et al. Clinical application of exome sequencing in undiagnosed genetic conditions. J Med Genet. 2012 Jun;49(6):353-61.

10. Veltman JA, Brunner HG. De novo mutations in human genetic disease. *Nature reviews Genetics* 2012; 13(8): 565-75.
11. Yang Y, Muzny DM, Reid JG, Bainbridge MN, Willis A, Ward PA, et al. Clinical whole-exome sequencing for the diagnosis of mendelian disorders. N Engl J Med. 2013 Oct 17;369(16):1502-11.
12. Consortium EP. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* 2004; 306(5696): 636-40.
13. Anaya JM. The diagnosis and clinical significance of polyautoimmunity. Autoimmun Rev. 2014 Apr-May;13(4-5):423-6.
14. Anaya JM, Castiblanco J, Rojas-Villarraga A, Pineda-Tamayo R, Levy RA, Gómez-Puerta J, Dias C, Mantilla RD, Gallo JE, Cervera R, Shoenfeld Y,Arcos-Burgos M. The multiple autoimmune syndromes. A clue for the autoimmune tautology. Clin Rev Allergy Immunol. 2012 Dec;43(3):256-64.
15. Quail MA, Kozarewa I, Smith F, Scally A, Stephens PJ, Durbin R, Swerdlow H, Turner DJ. A large genome center's improvements to the Illumina sequencing system. Nat Methods. 2008 Dec;5(12):1005-10.
16. Minoche AE, Dohm JC, Himmelbauer H. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. Genome Biol. 2011 Nov 8;12(11):R112.
17. Nakamura S, Nakaya T, Iida T. Metagenomic analysis of bacterial infections by means of high-throughput DNA sequencing. Exp Biol Med (Maywood). 2011 Aug;236(8):968-71.
18. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009 Jul 15;25(14):1754-60.
19. Andrews TD, Whittle B, Field MA, Balakishnan B, Zhang Y, Shao Y, et al. Massively parallel sequencing of the mouse exome to accurately identify rare, induced mutations: an

immediate source for thousands of new mouse models. Open Biol. 2012 May;2(5):120061.

20. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078-9.
21. Fairbrother WG, Holste D, Burge CB, Sharp PA. Single nucleotide polymorphism-based validation of exonic splicing enhancers. PLoS Biol. 2004 Sep;2(9):E268.
22. Wu J, Jiang R. Prediction of deleterious nonsynonymous single-nucleotide polymorphism for human diseases. ScientificWorldJournal. 2013;2013:675851.
23. Sunyaev SR, Eisenhaber F, Rodchenkov IV, Eisenhaber B, Tumanyan VG, Kuznetsov EN. PSIC: profile extraction from sequence alignments with position-specific counts of independent observations. Protein Eng. 1999 May;12(5):387-94.
24. Ramachandran GN, Ramakrishnan C, Sasisekharan V. Stereochemistry of polypeptide chain configurations. J Mol Biol. 1963 Jul;7:95-9.
25. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010 Apr;7(4):248-9.
26. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003 Jul 1;31(13):3812-4.

Legends to the Figures

Figure 1. Main steps of next-generation sequencing for whole genome and exome sequencing. Extracted DNA is broken into <1000 bp fragments and adaptor sequences (in green and red) are ligated to fragments. In whole-genome sequencing, all fragments are sequenced. In whole-exome and targeted sequencing only a subset of the original fragment pool is sequenced. Fragments are separated on a slide and clonal amplification by PCR to generate fragment clusters. Four fluorescently labeled nucleotides are added to the slide and compete to be incorporated to the growing chains. In each cycle, the clusters are excited by laser and the emitted fluorescence (colored circles) is recorded by an image-capturing device.

Figure 2. Bioinformatics algorithm for aligning, curating, and filtering of data obtained from next generation sequencing. The genetic analysis for candidate pathogenic variants can be performed by classical association and/or linkage analysis if variants are common, or by collapsing methods as the Kernel Based Adaptive Cluster (KBAC), in the case of rare ones (those with a minor allele frequency < 1% in the population.