Candidate Sequence Variants for Polyautoimmunity and Multiple Autoimmune Syndrome from a Colombian Genetic Isolate: Implications for Population Genetics

Angad Singh Johar

October 2017

A thesis submitted for the degree of the Doctor of Philosophy of the Australian National University

Declaration

This work was conducted from January 2014 to October 2017 at the Genomics and Predictive Medicine Group, John Curtin School of Medical Research, The Australian National University, Canberra, ACT. During my commencement of this project on 14th January 2014, I was initially. My transfer into the PhD. Program officially took place on the 1st July 2015.

Unless otherwise referenced, this thesis comprises only my original work towards the PhD degree of the Australian National University.

This document has not been submitted for qualifications at any other academic institution.

Angad Singh Johar

Canberra, Australia

October 2017

Acknowledgments

There are many people who undoubtedly deserve thanks and acknowledgment for the success in this project. Firstly, I must thank Dr. Mauricio Arcos-Burgos for his guidance and inspiration that he provided as a supervisor. Since I met you at the end of December 2012, to commence an Honour's program the following year, your kindness as well as the strong leadership and unity that you have created within our lab group during our time together at JCSMR has been amazing. It is these qualities that have enabled me to strive harder and engage in genetic research with pride and passion, both during my Honour's year as well as during my Ph.D. tenure. Another fantastic aspect of you is that whenever I wanted to explore new ideas for our research, you would always give me good advice, but more importantly full confidence and backing, as you truly appreciate the value for innovation. Furthermore, our success and wonderful collaboration together has been enhanced by the fact that our families are now close friends, due to your wonderful nature as a human being. Muchas gracias para todos, hermano!

To Dr. Claudio Mastronardi, words cannot describe the value of your contribution. Along with Mauricio, your conversations with me always provided plenty of belief that I had the skills to be able to conduct this project, and generate quality publications as a result of that. Not only that, your insights and knowledge of immunological mechanisms, Sanger Sequencing and pharmacogenomics along with Maria Fernada Silva Lara had a pivotal role in our publication success in the Journal of Translational Medicine and the Journal of Autoimmunity.

I am equally grateful also to Dr. Juan-Manuel Anaya for all of his efforts at the CREA (Centro de Estudio Enfermedades Autoinmunes or Centre for Autoimmune Disease Research). Without the rigorous efforts of you, your staff members at the CREA and of course Mauricio in recruiting and diagnosing patients and later obtaining their DNA samples, this project would not have gone ahead. Furthermore, words can't describe how much I appreciate the kindness of you, Andrea and Simon for accommodating me during my attendance at the Colombian Symposium for Autoimmunity in Bogota. This was despite the fact you already had enormous responsibilities on your shoulders for organising and hosting this massive event. !Muchas gracias amigo!

To Dr. Nicolas Molano González, even though we only spent a short time together in Bogota, you too deserve big thanks! We had many productive sessions together, in which we exchanged ideas on approaches to genetic research from a statistical and bioinformatics perspective. You are truly a hard-worker and amazing bioinformatician, and it was a great learning experience for me to be with you. The CREA is truly blessed to have you. !Espero que podemos trabajar en el futuro hermano!

For Dr. Jorge Vélez, I can only say this: Tu estas magnifico! Your help and support has been nothing short of amazing. Your insight into statistics, bioinformatics is like no one else that I have worked with, apart from Mauricio. I have to say that I learnt more about the approach required to address research questions more than what I did in any undergraduate course. Also amigo, if there was any way to personify skill, ingenuity, toughness, resilience and determination simultaneously, you would be the first that comes to mind. !Tambien amigo, muchas gracias para mi ayudar en el tarea de Español! I think I can say that my Spanish has improved thanks to you and the rest of our lab group! I hope there will be many more opportunities to work with all of you in the future and continue our friendship!

To the BRF (Biomolecular Resource Facility) staff at JCSMR, thanks for your reliability and efficiency in processing all of our sequencing needs. Your work facilitated our analysis. Within the BRF, one man who deserves special thanks is Dr. Stephen Ohms. I really appreciate your efforts in allowing me access to the Metacore Pathway and Network Suite for conducting part of my analysis for this project, as well as other projects performed by our lab group. Furthermore, we really appreciate you for dealing with the cost of the license for the software!

To the Genome Discovery Unit (GDU), in particular Dr. Hardip Patel, your skills and dedication in generating variant calls from our whole exome capture data was an essential step, without which our downstream analysis would not have been possible. Furthermore, Hardip, your advice was extremely useful in enhancing the success of these studies. I will also miss our discussions of cricket together!

Finally, and most importantly, my beloved parents: Your unconditional support throughout my life, and in particular during the last 3.5 years was the key driving force in getting to this point. Thank you for always backing and believing in me! The strength required to attain whatever success comes from this project was ultimately from you. I can’t thank you enough.

Publications and Presentations

Published Papers

1. A. Johar, J.M. Anaya, D. Andrews, H.R. Patel, M. Field, C. Goodnow, and M. Arcos-Burgos. Candidate gene discovery in autoimmunity using extreme phenotypes, next generation sequencing and whole exome capture. *Autoimmun Rev,* 14(3):204-9, 2014.

2. A.S. Johar, C. Mastronardi, A Rojas-Villararga, H.R. Patel, A. Chuah, K. Peng, A. Higgins, P. Milburn, S. Palmer, M.F. Silva-Lara, J.I. Vélez, Novel and rare functional genomic variants in multiple autoimmune syndrome and Sjögren's syndrome. *J Transl Med,* 13:173, 2015. The content of this manuscript is presented in chapter 3 of the thesis

3. A. Johar, J.C. Sarmiento-Monroy, A. Rojas-Villaraga, MF Silva-Lara, H.R. Patel, R.D. Matilla, J.I. Vélez, K.M. Schulte, C. Mastronardi, M. Arcos-Burgos, J.M. Anaya. Definition of mutations in polyautoimmunity. *J Autoimmun,* 72:65-72, 2016. The content for this manuscript is presented in chapter 4 of the thesis.

4. J.I. Vélez, F. Lopera, D. Supulveda-Falla, H.R. Patel, A.S. Johar, A. Chuah, C. Tobón, D. Rivera, A. Villegas, Y. Cai, K. Peng, R. Arkell, F.X. Castallenos, S.J. Andrews, M.F. Silva Lara, P.K. Creagh, S. Easteal, J. de Leon, M.L. Wong, J. Licinio, C.A. Mastronardi, and M. Arcos-Burgos. APOE*E2 delays age of onset in PSEN1 Alzheimer's disease. *Mol Psychiatry,* 21(7):916-24, 2016.

5. J.I. Vélez, F. Lopera, H.R. Patel, A.S. Johar, Y. Cai, D. Rivera, C. Tobón, A. Villegas, D. Sepulveda-Falla, S.G. Lehmann, S. Easteal, C.A. Mastronardi, and M. Arcos-Burgos. Mutations modifying sporadic Alzheimer's disease age of onset. *Am J Med Genet B Neuropsychiatr Genet,* 171(8):1116-30, 2016.

6. G. Paz-Filho, M.C. Boguszewski, C.A. Mastronardi, H.R. Patel, A.S. Johar, G.A. Huttley, C.A. Boguszewski, M.L. Wong, M. Arcos-Burgos, and J. Licinio. Whole exome sequencing of extreme morbid obesity patients: translational implications for obesity and related disorders. *Genes(Basel)*, 5(3):709-25, 2014.

7. A.C.P. da Fonesca, C. Mastronardi, A. Johar, M. Arcos-Burgos, and G. Paz-Filho. Genetics of non-syndromic obesity and use of high-throughput DNA sequencing technologies. *J. Diabetes Complications*, 31(10):1549-61, 2017.

Manucripts Prepared for Submission

8. A.S. Johar, M.T. Acosta, A.F. Martinez, J.I. Vélez, J. Swanson, A. Stehli, B. Molina and the MTA Team, H.R. Patel, D.A. Pineda, F. Lopera, J.D. Palacio, F.X. Castallenos, C. Mastronardi, M. Muenke, and M. Arcos-Burgos. Rare and common variants in *HYDIN*, a gene implicated in laterality and brain development, confers risk to ADHD.

9. A.S. Johar, J.I. Vélez, C.A. Mastronardi, J.M. Anaya and M. Arcos-Burgos. Genetics of Population Isolates: Comparisons of the Paisa community with other cohorts from Europe and South America. The content of this drafted manuscript is given in Chapter 5.

Table of Contents

List of Tables

List of Figures

Abbreviations

A2M	Alpha-2-Macroglobulin
ADs	Autoimmune Diseases
ANA	Anti-Nuclear Antibodies
AITD	Autoimmune Thyroid Disease
ANU	Australian National University
BAM	Binary Alignment Map
BLOSUM	Block Substitution Matrix
BRF	Biomolecular Resource Facility
CDCV	Common Disease Common Variant
CEU	Central Europeans of Utah
CLM	Colombians in Medellin
CLRT	Composite Likelihood Ratio Test
EP	Extreme Phenotypes
EPS	Extreme Phenotype Sampling
ESP	Exome Sequencing Project
ExAC	Exome Aggregation Consortium
FDR	False Discovery Rate
FIN	Finnish in Finland
GATK	Genome Alignment Toolkit
GBR	British in England and Scotand
GVF	Genome Variant Format

GSMA	Genome Scan Meta Analysis
GWAS	Genome-Wide Association Studies
HGC	Human Gene Connectome
IFN	Interferon
IBS	Iberian Spanish
IGV	Integrative Genomics Viewer
JAK	Janus Kinase
JCSMR	John Curtin School of Medical Research
INDELs	Insertions and Deletions
KBAC	Kernel Based Adaptive Cluster
LD	Linkage Disequilibrium
LPS	Lipopolysaccharide
LOD	Logarithm of Odds
LHS	Left Hand Side
MAF	Minor Allele Frequency
MAS	Multiple Autoimmune Syndrome
MXL	Mexicans in Los Angeles
MNP	Multi-nucleotide Polymorphism
NHLBI	National Heart, Lung and Blood Institute
OMIM	Online Mendelian Inheritance in Man
pbGWAS	Pooling/Bootstrap Genome-Wide Association Studies
PCA	Principal Component Analysis
PEL	Peruvians in Peru

PSIC	Position Specific Independent Counts
PUR	Puerto Ricans in Puerto Rico
pVAAST	Pedigree Variant Annotation and Analysis Search Tool
RA	Rheumatoid Arthritis
RF	Rheumatoid Factor
RVAT	Rare Variant Association Test
VAAST	Variant Annotation and Analysis Search Tool
WEC	Whole Exome Capture
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing
WTCCC	Welcome Trust Case Control Consortium
SAM	Sequence Alignment MAP
SNP	Single Nucleotide Polymorphism
SLE	Systemic Lupus Erythematosus
SR	Sum Rank
SS	Sjogren’s Syndrome
T1D	Type 1 Diabetes
TSI	Toscani in Italy
VAT	Variant Annotation Tool

Abstract

Autoimmunity is an immunological disorder whereby patients have lost immunological tolerance to self-antigen. It has extreme financial and socioeconomic burden with costs of over 100 billion dollars in the USA alone, and an estimated prevalence of 9.4%, and evidence indicates that this estimate has increased at a rate of 5% per year for the past 3 years. These phenotypes can be manifested in more severe forms through polyautoimmunity, whereby patients are carrying 2 or more autoimmune conditions. In addition to that, there is also the most extreme phenotype of autoimmunity known as the Multiple Autoimmune Syndrome (MAS), consisting of cases where patients have 3 or more autoimmune diseases. These extreme phenotypes are extremely important for genetic research as will be elaborated upon in this thesis. For more than 20 years, pedigrees from the world's largest known genetic isolate, from the Paisa region of Colombia have been ascertained and thoroughly followed by Dr. Juan-Manuel Anaya and Dr. Mauricio Arcos-Burgos. This population has maintained its status as a genetic isolate since the 16th century, during the early colonization by the Spanish Conquistadors.

In this thesis, our attempts in identifying potential candidate variants potentially underpinning the genetic etiology of autoimmune phenotypes in this population is facilitated by the fact that families are derived from cohorts where genetic homogeneity is maximized. Candidates are identified in both sporadic as well as familial cases. This is primarily achieved through combination of linkage analysis and association tests for both rare and common variants, derived from variant-calling pipelines and that had undergone quality control, filtering and functional annotation, via bioinformatic anlayses. Genes harbouring variants with significant evidence of linkage and association were primarily involved in negative regulation of apoptosis, phagocytosis, regulation of endopeptidase activity, response to lipopolysaccharides and plasminogen urokinase receptor activity. These findings, that were obtained by utilizing the combinations of statistical as well as network-based analyses have relevant potential implications in autoimmunity, and can be further supported with additional studies.

Chapter 1: Introduction

1.1 Autoimmunity: Impact Upon Populations

Medical research into autoimmunity first began over 100 years ago during its initial discovery [1]. The disease comprises an immune system defect in recognition of self-tissues that leads to them being attacked as foreign antigens [1]. Studies suggest there are 81 autoimmune diseases (ADs) worldwide [2]. Some estimates indicate that autoimmunity prevalence is 4.5-5.3% worldwide [3-4]. However, in recent studies, this figure has increased to 9.4% [4-6]. The financial burden for 7 of the 81 known ADs in the USA alone is over $100 billion [6]. Currently, there is no cure for autoimmunity, and only symptoms can be treated. However, further knowledge of the genetic etiology of these conditions has been increased in recent years, to which our studies contribute.

1.2 Physiological Understanding of Autoimmunity

Patients with diseases such as Rheumatoid Arthritis (RA) often have high levels of IL1 beneath the synovial membrane. These cytokines stimulate the production of proteinases (such as collagenases), which destroy matrix proteins within the cartilage and bone, allowing synovial fluid to enter these regions. This results in irreversible damage and painful swelling around the join structures. Also the presence of high Rheumatoid Factor (RF) titres is associated with the disease. Whilst the actual mechanism is unknown, one theory for the way in which this autoantibody leads to the disease phenotype is that the localization of RF in the synovial membrane leads to excess cytokine and chemokine release creating a positive feedback mechanism that increases the deposition of immune complexes of RF with the Fc portion of IgG [5]. This binding leads to activation of macrophages and neutrophils which seemingly play an important role in the intense inflammatory symptoms and joint destruction observed in RA.

Another example of highly prevalent AD is Systemic Lupus Erythematosus (SLE). The presence of anti nuclear antibodies (ANA) is often associated with the presence of SLE. Studies indicate that defective apoptosis in SLE patients prevents effective clearance of all components of apoptotic cells. This potentially exposes these nucleic acids (antigens) to the immune system, resulting in the stimulation of ANA production [6]. Production of these antibodies can be enhanced by simultaneous occurrence of B lymphocytes that are unable to respond to suppressor signals or defective T regulatory lymphocytes. These ANA complexes can be deposited in different organs such as the heart, kidney and the skin, leading to activation of complement, emigration of neutrophils and release of kinins and prostaglandins ultimately resulting in inflammation [5].

As well as joint and organ damage and excess inflammation, metabolic impacts are also evident in autoimmunity as mentioned earlier. In the case of AITD, there are often elevated levels of anti thyroglobulin antibodies. This prevents efficient homeostatic maintenance of T3, T4 and Thyroid Stimulating Hormones (TSH) [5], which are crucial in metabolism. Thus the disease will affect iodine levels, which lead to Goitre, and also affects processes such as protein and carbohydrate metabolism, neural maturation, Sodium ion transmembrane transport and protein synthesis for stability of bones and joints [5]. Such impacts on metabolism lead to rapid fluctuation in body weight and heart rate, and brittle bones, and these symptoms are potentially serious as those of the diseases described above.

1.3 Current Genetics Knowledge of Autoimmunity

Early functional studies revealed that antigen presentation to T cells has a crucial role in triggering onset of AD. This prompted investigation into the HLA gene locus, a 7.6Mb region, encoding the Major Histocompatibility Complex (MHC) [7]. The polymorphic nature of this locus means that it can produce many HLA subtypes and haplotypes, resulting in a greater chance for finding candidate mutations in immunologically relevant genes. It has now been well established that molecular studies validated early findings from the functional investigations. This is demonstrated from linkage analysis results, which thus far have revealed strong evidence of associations of various HLA subtypes with autoimmunity. For example HLA-DRB1 has been shown to have significant association to RA, particularly in Latin American populations [8], whilst HLA-DQA2 has been linked to Autoimmune Thyroid Disease (AITD) [8]. There are many proposed mechanisms for illustrating the role of these HLA subtypes in autoimmunity. One theory is that the genetic variants will lead to variation antigen binding grooves of the MHC protein molecule, such that it binds to self-antigen, which is then subsequently destroyed by the immune system. Many other plausible theories exist, but the exact mechanism will depend on the deleterious base change that has occurred within the gene’s DNA sequence [7].

Whilst many autoimmune diseases such as SLE, RA and AITD have been shown to have a common association with the HLA locus, this genomic region only accounts for a minor fraction of genetic etiology for AD phenotypes. Thus it is logical to assume that the disease inheritance in patients (including those from whom these DNA samples were obtained) will have a multifactorial inheritance [9]. In order to address this issue, multiple candidates must be found to account for the presence of AD and MAS phenotypes observed in populations where these diseases are prevalent. Our Colombian population is no exception in this case.

Since the introduction of GWAS (Genome Wide Association Studies) substantial progress has been made in accounting for the multifactorial genetic origins of autoimmunity. For example the *BLK* gene encodes B cell kinases that are important for B cell differentiation, whilst *IRGM* encode a GTPase used for intracellular pathogen defence [10]. Often many of these variants have been associated with more than one disease. For example, *PTPN22* contains a single coding SNP (Single Nucleotide Polymorphism) linked to T1D, SLE and Rheumatoid Arthritis phenotypes. However according to the data by Arcos-Burgos and Anaya [8], individuals show a high concordance of autoimmune diseases between families in comparison to other members of the region’s population, suggesting that variants not necessarily common to the wider population, contribute to the observed phenotypes. Thus, individuals with a single autoimmune disease are susceptible to MAS. It is hypothesized that both common and rare variants contribute to the disease manifestations of this complex disorder. Arguments in favour of common variants follow the common disease common variant (CDCV) hypothesis. I.e., due to the fact that rare variants have low allele frequencies in any given population, it is very unlikely that (individually) they will have sufficient overall effect size to show statistically significant evidence of contributing to a particular disease phenotype (in this case MAS, autoimmunity and polyautoimmunity). This may be the case even with very large sample sizes [11]. Common variants on the other hand are theoretically more likely to be identified amongst affected individuals. For this reason, it is argued that many common variants of small individual effects, simultaneously contribute to observed disease phenotypes [11].

In total than 2000 disease associated common variants have been identified through GWAS [12], so this approach has had considerable success in many complex diseases. Many of these variants have high population frequencies, in accordance with the CDCV hypothesis [13-15].

Although GWAS findings have made considerable inroads to this problem in the past, the biggest challenge for finding candidate genetic variants is the issue of missing heritability. In fact, indications are that over 50% of autoimmune loci aren’t elucidated [9]. For example, even though in a study with more than 150,000 individuals, more than 70 loci exhibited genome-wide significance in association with Type 2 Diabetes, only 11% of heritability has been explained [16]. Likewise in Crohn’s disease, with 210,000 individuals, only 23% of inheritance has been accounted [16].

Due to reliance on common variants, GWAS are often underpowered to detect rare variants due to insufficient effect size. I.e. the presence of rare variants amongst patients is not frequent enough to give a strong statistical association between the presence of these variants and the occurrence of disease [11]. Furthermore, the genetic basis of these complex diseases cannot be explained by the rare variant hypothesis or the CDCV hypothesis alone. Instead results from many previous studies into MAS indicate that the disease phenotypes are due to an interaction of many common variants which have small but additive impacts, as well as major effect variants of lower frequency [11].

To further support the rare-variant hypothesis, evolutionary and selection theory must also be considered. That is, disease-causing variants are likely to have reduced allele frequency due to their deleterious nature, as a result of purifying selection [11]. In particular, loss-of-function mutations, which disrupt the production of functional proteins, should be very rare [11]. Therefore, one can argue that pathogenic rare variants will potentially have greater disease effects and penetrance compared to common variants.

Therefore there has been a greater shift towards analysis of rare variants in order to underpin the genetic etiology of complex diseases, including autoimmunity [11, 16]. Due to the low allelic frequency of rare variants as mentioned earlier, such objectives are more likely to be achieved via linkage analysis and Rare Variant Association Testing (RVAT) algorithms, as explained later [16-18].

Thus, in order to address the issue of missing heritability, methods to analyse rare variants are essential for studying complex diseases such as autoimmunity [16-18]. In particular, the power of such studies can be enhanced from genetic isolates [19].

1.4 The Value of Population Isolates in Genetic Studies

As stated before, that the benefit of population isolates in genetic studies arises from the maximization of genetic and environmental homogeneity. Also, there is also minimization of population stratification [20]. Thus, this increases the likelihood of finding variants that are overrepresented and shared amongst affected individuals (be it familial or sporadic cases) with autoimmunity, polyautoimmunity and MAS [20], compared to more admixed populations.

In this case, our studies were conducted on the Paisa genetic isolate from Medellin, within the province of Antioquia in Colombia. Even though the initial founder population had substantial Amerindian and Iberian Spanish admixture, the descendants of these founders have been largely geographically and culturally isolated for over 400 years [19-20]. This is largely due to the mountainous environment, the resource rich landscape as well as the cultural restrictions/domination imposed by the conquistadors and representatives of the Spanish crown during the initial founding stages of this population, and for a long time thereafter. I.e., during colonial periods, the Spanish rulers would enforce strict segregation laws, whereby individuals of mixed Iberian-Amerindian ancestry were only permitted to marry with those from invading Hispanic heritage and not within their own communities [19-24]. Meanwhile, unmixed Native Amerindians were isolated over time. Continuation of this practice for many generations and exclusion of individuals with full Amerindian ethnicity, eventually led to a genetic structure whereby Paisas had greater than 85% Caucasian ancestry, deduced by identity coefficient analyses [19-24]. This was followed by wars for independence, whose outcome resulted in change of territorial distribution (for loyalists and separatists respectively) and hence further separation between the Paisa population and other regions [19-24]. Thus, from all of these historical events, for a large proportion of the time since the initial colonization, this population remained genetically isolated due to limited additional admixture and gene flow, from neighbouring subpopulations within the remainder of Colombia [19-24].

Thus owing to genetic evolutionary forces such as founder effects and genetic drift, large DNA sequence regions and haplotypes are shared between many individuals. This can lead to not only enrichment in common, but also rare variants, due to allelic fixation. In the case of the Paisa community, allelic architecture within this genetic isolate may have also been potentially influenced population bottlenecks, which may have occurred due to the large impact of historical, geographical and political influences. All of these factors will contribute to the unique allelic architecture of this population, thereby influencing the outcome of our studies, with regard to the likelihood of success in identifying disease-causing variants, during genetic studies.

A unique aspect about this population is the presence of a multi-ethnic founder effect [19-20]. As mentioned before, the founder population of the Paisas consisted of both Spanish invaders and Amerindian locals [19-24]. In addition to the colonizers of Extramadura, Andalusia and to a lesser extent the Catalan province of Spain, the Paisa region also received Basque and Sephardic Jewish immigrants. These groups migrated to this area in order to escape cultural and religious persecution, particularly as a consequence of the Spanish Inquisition, where tribunals were set in many colonies of the empire. Around the same time, the Spanish empire also sought to maintain a cultural monopoly on a linguistic level as well, by only allowing the use of the Castilian language. For this reason, the Basque and Sephardi Jewish immigrants sought geographical isolation from the urban bases of the Spanish Conquistadors [19-24]. Thus, spatial and temporal admixture amongst the Basque, Sephardic Jewish and majority ruling Castilian ethnicities over time shaped the Paisa population, as well as the initial Amerindian founders in the colonization stage. This was followed by a long period (> 400 years or 20 generations) of geographic and cultural isolation from other regions. These phenomena were influential in the population genetics and multi-founder effect of the Paisa community. Such phenomena contribute to increased levels of genomic linkage disequilibrium (LD) [19]. Over time, the combination of these evolutionary forces also increases the overall frequency and proportion of variants shared across a given genomic region, especially with the loss of heterozygosity, which can occur across multiple generations via genetic drift [7], as previously mentioned.

1.5 Utility of Extreme Phenotypes from Genetic Isolates

In addition to the genetic homogeneity, the extreme phenotypes of autoimmunity carried by this cohort further enhance the power for these genetic studies. This is due to the fact that extreme phenotypes can facilitate the identification of rare as well as common variants. The reason is that the likelihood of identifying causative, major effect variants is higher in individuals carrying extreme phenotypes, compared to those with less severe manifestations of the phenotype in question [25]. This is illustrated using proofs based upon the principles of mathematical induction [25]. The proof elaborates that major effect variants are more likely to be enriched in individuals with more extreme/severe disease phenotypes. This was achieved by performing the mathematical proof, conditional on the effect size as well as the probability that a particular variant is causative, given the extremity of the observed disease phenotype, as detailed in chapter 2 [25]. Individuals can be selected, after adjusting for covariates, based on trait values or risk factors. For example, in case-control analysis, one may select individuals with early onset and family history, comparing those to late onset, and no family history or known lifestyle factors that confer disease risk. For quantitative and binary traits, EPS approaches are more powerful than random sampling. Empirical studies suggest that selecting from the upper and lower tails of phenotypic distribution can reduce the required sample size by up to 50% or in some cases 7 fold [25-27]. Therefore, analyses utilizing such extreme phenotypes are advantageous, as they can remain substantially powerful, even with small sample sizes, particularly when cohorts are derived from genetic isolates, as is the case with the Paisa community.

1.6 Large Output Next Generation Sequencing Data

Inspite of the difficulties described above in uncovering the genetic basis of MAS (which also apply to other complex diseases), the availability of Next Generation Sequencing Technologies (NGS) data generated from sequencing projects has enabled substantial advancements to be made in this area. The use of automated Illumina HiSeq Sequencing by Synthesis pipelines (relying on cluster generation, bridge amplification and light wavelength detection from fluorescently labelled nucleotides) for exome and whole genome sequencing (WGS) data allows massive parallel sequencing of reads, making it a high throughput technology [28-31]. In fact up to 1Gb of DNA is generated for each run with average error rates of less than 0.3% per base [28-31].

Also, during every sequencing cycle all 4 nucleotides are present as single separately fluorophore-labelled molecules that are competing for hybridisation to the template DNA, thereby minimising chances of incorporation bias. In addition, the combination of paired end technology and mate pair libraries allows flexibility in insert sizes (200bp-5kb), sufficient to maximise coverage, minimise gaps and enable accurate sequencing over small repetitive regions within the exome [28-31]. Otherwise the short reads can align to vastly different regions in the exome and create gaps in the assembly, particularly in repetitive sequences. At the same time, excessively large insert sizes may have the disadvantage of low coverage. Hence Illumina sequencing is a powerful tool, as it can implement a combination of large insert mate pair reads with contiguous sequences and high coverage reads with smaller insert sizes to address both of these issues. This property will be useful for de novo assembly of new versions of the human genome, which may be required to uncover any hidden variation that is present in poorly annotated genomic regions [28-31].

Subsequently, more projects have emerged that generate a higher quantity of variant data, which can be used to search for candidate disease genes in ADs. The emergence of NGS has substantially increased the volume of variant data available, as it can develop billions of short sequence reads in a cost-effective manner [28-31]. One such project is the 1000 genomes project. This project involved the sequencing of 1092 human genomes worldwide (in phase 1), including 60 Colombians from Medellin, as well as other regions from the Americas, Africa and Europe [32]. The aim of this project is to use the genetic variation obtained from these different populations as a representative catalogue of known human genetic variation. In total the project consortium has genotyped and generated haplotype maps for a large number of variants comprising of 38 million SNPs, 1.4 million indels and 14000 large deletions [32]. Also a large fraction of the reference genome (94%) used in this project for variant annotation is accessible due to long sequence read lengths. Therefore, this database provides is useful in assisting experimental design and analyses aimed at enhancing our understanding of the genetic basis of autoimmunity.

As well as the advances in next generation sequencing (technologies), sequencing procedures can also be enhanced in efficiency by the use of exome capture. Firstly, it is a cost effective technique that employs efficient capture by hybridisation techniques, using exon specific oligo nucleotides to enrich only protein coding sequences that can be later used for sequencing. Prior estimates in various studies indicate that the proportion of disease variants located in protein coding regions and canonical splice sites is as low as 25% and a maximum of 85% [33-34]. This is a considerably high estimate, given that the exome sequence comprises of only 1.2% of the entire genome. This makes exome capture more cost effective, because fewer sequence reads are required than Whole Genome Sequencing in order to maximise coverage [35]. The comparatively low cost of this technology means that exome databases are rapidly expanding. Therefore there is a large abundance of chromosomes that can be used as controls for identifying candidate disease genes [36]. The most extensive of these is that of Exome Aggrehation Consortium (ExAC) [37-38]. The ExAC project has combined many samples including but not limited to; exome sequencing data from the 1000 Genomes project, National Heart Lung and Blood Institute (NHLBI) Exome Sequencing Project (ESP) cohort [39]. In total there are over 10 million catalogued variants across 60,706 individuals [37-38], making it the largest single exome sequence database, compared to its predecessors, thereby offering several analytical advantages. The size of ExAC database allows for detection of mutational occurrence, whereby the same mutations repeatedly occur throughout the population history of a given cohort [37-38]. Large proportions of variants found in external unrelated trio samples are also shared in the ExAC database, suggesting independent origins of mutations within given cohorts at different points in time.

The ExAC database is also valuable for the diversity of sampled populations as reflected in the geographic ancestry axes of variation from the Principal Component Analysis [37-38]. The large overall sample size and cataloguing of variants from continentally diverse population groups, ensures minimal identification of spurious results from stratification and population structure. Empirical data suggests that the total number of candidate variants by 7 fold compared to the NHLBI ESP project [37-38]. To compliment its variant filtering ability, the database also consists of many MNPs (multi-nucleotide polymorphisms) which are often missed by many sequencing, variant calling and quality control pipelines. This is because the variant calling pipeline, Genome Alignment Tool Kit (GATK) HaplotypeCaller [37-38, 40-42], is capable of assembling reads into haplotypes and later realigning them, when variants falling within the same codon are encountered, for multi-nucleotide polymorphism (MNP) identification [37-38, 41-43]. These realignment algorithms are adapted from those, used for mapping insertions and deletions (INDELs) [37-38, 40-42]. Thus, the utility ExAC as a filtering tool for identifying rare and de novo candidate disease alleles, based on minor allele frequencies is increased.

Despite its advantages, one must be aware of the fact that exome sequencing has limitations. This is because hybridisation probes are not available for all annotated exons within the ENSEMBL database, particularly those harbouring repeat sequences on chromosomal ends. Also exome sequencing will have little success to detect mutations in non-coding DNA that alter gene function by various regulatory mechanisms and enhancer effects [43]. Such variants (in recent times) are emerging as important contributors to genetic disease and they occur in >98% of the human genome, which is undetectable by exome capture, and can only be obtained by WGS [43-44]. Nevertheless, when performed in well-selected and phenotyped cohorts and powerful statistical analyses, this approach is still capable of substantial success, especially considering the small fraction of exome DNA across the genome.

1.7 Developments in Linkage and Association Analyses

The increase in data output of NGS means more rare-variants are seen, which previously were undetected by Sanger or SNP-chip technologies. To utilize the higher data output from NGS, enhancement of statistical analyses algorithms was essential, especially to identify undiscovered rare, pathogenic variants. Therefore, linkage and RVATs have been further enhanced, as alluded to earlier [16, 18]. When there is limited availability of families, the best alternative is RVAT [16, 18]. Unlike traditional GWAS (Genome Wide Association Studies) analyses, this method relies on variant aggregation/clustering. That is, rare variants within a gene can be collapsed together into single multi-site genotypes [16, 18]. Therefore, the association test statistic and the corresponding P-value to evaluate the degree of enrichment/overrepresentation of cases over controls is in fact reported as a gene-based, rather than a variant-based quantity. I.e., instead of utilizing information from a single variant in association analyses, multiple variants are simultaneously evaluated as a single genotype [16, 18]. This overcomes the issue of low effect size for single rare variants and reduces the need for collecting extremely large sets of samples.

Previous results have shown that obtaining odds ratios in excess of 1.4, with 80% power will require more than 500,000 samples, with disease prevalence levels of 5% assumed and nominal significance of 5×10^{-8}. Naturally, as sample sizes are increased, the number of rare causal variants will also rise along with the proportion of explained phenotypic variance and heritability. Using the exome sequencing project (ESP) by NHLBI, it was observed that explanation of phenotypic variance of rare variants were as high as 84% and as low as 18% for varying effect size simulations [16]. In fact, by these calculations it can be inferred that the estimates of phenotypic variance will still be underestimated, even with ~ 1000,000 individuals, reiterating the point of rare variant effect sizes [16].

Also, given that human genomes predominantly consist of higher proportions of low frequency, rare and/or de novo mutations over common ones, more stringency on variant filtering may be required [11, 16]. This further reduces analytical power, especially by increasing false negatives [11, 16]. Therefore, when applying multiple testing corrections to gene-based variant collapsing tests, one only has to adjust for ~ 20-25000 genes in the genome, instead of millions of SNPs, thereby reducing the number of hypotheses to be tested.

To supplement rare variant association tests, many bioinformatics tools to predict protein functionality of SNPs affecting amino acid sequence changes are prevalent [45-49], based upon sequence conservation and molecular structures. Nevertheless, these tools are prone to misclassification, with some studies obtaining false prediction rates as high as > 40% [50]. To overcome this, rare-variant tests can conduct adaptive weighting for combinations of multi-site genotypes within given genes according to their sample risk (number of individuals with genotype and disease divided by total number of affected people), similar to algorithms such as the Kernel Based Adaptive Cluster (KBAC) [18]. Therefore, the likelihood of obtaining false positive or false negative results, due to bias introduced by premeditated filtering strategies (based on protein deleteriousness predictors) is minimized [18]. However, even though variant effect predictors are susceptible to misclassification, information regarding protein functionality still needs to be utilized. This has been recently achieved with the rare-variant based composite likelihood ratio test in the variant annotation and analysis search tool (VAAST), by incorporating information of amino acid sequence conservation in various protein databases.

Whilst this strategy is successful, large sample sizes may still be required to overcome genome-wide significance levels for genes harbouring sets of extremely rare causative variants [16-18]. Also, one has to overcome the issue of ascertainment bias causing potential population structure. Therefore, familial-based linkage studies have been helpful in this regard [16-18]. In this case, the power of linkage analysis lies in the homogeneity within families [17]. This is because, even though the disease causative variants may be rare in a particular population, individuals within families are expected to share large numbers of variants and haplotypes, compared to non-related individuals [17]. Therefore, the segregation patterns within families are likely to show significant evidence of linkage, as long as the candidate disease variant has greater representation in the affected (compared to unaffected) pedigree members, even if the sample size is not large [17]. For this reason, linkage analysis will be more powerful than single variant based association studies of rare variants.

The fact that familial samples have greater homogeneity than population-based cohorts is advantageous for detection of highly penetrant rare and/or low frequency disease variants. Here linkage analysis algorithms can identify the likelihood of disease cosegregation, depending on the transmission patterns of variants from parental to offspring generations. Nevertheless, obtaining familial samples as large as typical case-control cohorts is challenging. Thus, the power of genetic studies, especially rare-variant analyses is enhanced with the integration of linkage and association information [51]. The reason is, pathogenic rare variants are more likely to be shared in families by cosegregation, than case vs. control cohorts.

Moreover, linkage analysis has been enhanced with developments that facilitate detection of de novo variants and compound heterozygotes, as implemented in tools such as the Pedigree Variant Analysis and Annotation Search Tool (pVAAST) [51], explained in chapter 2. These features explain why this algorithm substantially more powerful than other linkage approaches for disease analyses. Also, previously developed linkage algorithms are based upon the notion that selected markers for testing are not causative, but only cosegregating/linked with the disease variant. Conversely, the Elston-Stewart based pVAAST algorithm (considered as a modern form of linkage analysis) is based upon the assumption that causative variants can be directly assayed, under a gene-based model [51-52]. This allows the algorithm to be successful in sequence-based analyses with large variant numbers. Also, the gene-based nature of the analysis enables information from multiple variants to be combined (as is the case with RVATs), during the LOD score calculation [51]. Thus, once again, detection of potentially underrepresented variants is potentially empowered via this strategy.

Despite the benefits of combining linkage and association, genetic heterogeneity must also be addressed, for complex diseases. Genetic heterogeneity is accounted for in pVAAST, by the structure of the null and alternate model. The alternate model allows for current loci (genetic loci under investigation) and latent loci (those unlinked to the current loci) to both simultaneously contribute to the disease, thereby allowing for locus heterogeneity [51]. The null model, conversely only attributes disease phenotypes to latent loci. Hence, pVAAST-based linkage analysis is informative and powerful for single-gene Mendelian and polygenic complex disorders. Therefore, given all of these factors, it is no surprise that under conditions of reduced penetrance, missing phenotypes and heterogeneity, that this approach outperforms other previously developed linkage analysis algorithms [51]. Specifically, when these situations were simulated on previously tested datasets with known causative disease genes, the unified linkage and association algorithms, evidently present more consistent solutions than other algorithms [51]. Hence, this provides the platform for implementation of a powerful analytical framework for analyzing large outputs of NGS variant data in determining the genetic etiology of complex diseases including autoimmunity and MAS.

1.8 Scope of Study

The main topic of this thesis describes the results from implementation of a framework used to identify candidate genes in autoimmunity and patterns of genomic variation in selected genetic isolates that may influence disease susceptibility. Both sporadic and familial autoimmunity are analysed. The framework combines linkage analysis, gene-based association tests, pathway and network analyses as a means of data mining to achieve these goals.

In chapter 2, the algorithmic procedures for each of the aforementioned analyses required to attain our overarching aims is described. Their mathematical concepts are fully explained. Key points emphasize the importance of genetic isolates, control for population stratification as well as the mathematical and statistical concepts detailing the linkage, rare-variant association, pathway/network analyses as well as calculations of physiological relatedness by biological distance.

In chapter 3, we performed genetic analysis on whole exome capture (WEC) data from patients with sporadic forms of MAS and autoimmunity, compared against a set of controls. Sequencing was performed via the Illumina HiSeq 2000 NGS facilities. Variants were filtered based on minor allele frequency and functionality (determined by predicted pathogenicity of encoded proteins or appearance in regulatory regions). The collected sample has phenotypic heterogeneity (4 out of the 12 disease carriers have Sjogren's syndrome only, whilst others carry MAS) and they are unrelated cases, along with the small sample size. Although the autoimmune tautology argues for similar genetic origins of these diseases, this cannot apply for all phenotypes. Thus it is unlikely that any single variant will be enriched in a large proportion of affected individuals. Hence, in this scenario, a hard filtering approach is considered as the most successful approach to identify possible disease genes [51]. Genes harboring variants that were predicted to have damaging amino acid sequence changes or important in regulatory regions, absent from the controls and the 1000 Genomes Project were retained after filtering. Additional evidence of pathogenicity of these genes, in particular *LRP1* was supported by pathway and network enrichment analysis, and metrics of biological distance.

Having identified disease genes in sporadic cases, we aimed to build upon these findings via familial analysis. From the Paisa genetic isolate, we analyzed families ascertained by Dr. Arcos-Burgos and Dr. Juan-Manuel Anaya. In total, 18 families were sampled from this particular cohort, out of which 10 were selected for whole exome sequencing (WES) and linkage studies. Variants identified from the GATK Best Practices pipeline, were analysed under the pVAAST unified framework of linkage and association tests, integrated with functionality information. LOD scores and permutation-based p-values were used as a means of gene prioritization. Those meeting criteria for LOD score, P-value thresholds and sequence read quality were also extracted for studies of functionality. We identified that the *SRA1, DHX34, ABCB8, MLL4* and *PLAUR* genes all harbor mutations exhibiting statistical evidence of familial linkage, and are also involved in apoptotic-related processes as suggested by network analyses. It is granted that affected patients carrying these mutations have differences in phenotypic manifestations. However, the effectiveness of the study is not only due to the relatedness of the sampled individuals, but also the fact that they share phenotypes conforming to the autoimmune tautology (in particular RA and SLE). Hence it is hypothesized that some disease phenotypes in these patients share similar immunogenetic etiology and pathophysiological mechanisms. This increases the likelihood of overrepresented variants, identified within these families, in having clinically relevance.

The success of the studies involving sporadic and familial autoimmune patients can be attributed to the population genetics of the Paisa isolate. This is explored in chapter 5. Here, we describe the comparison of genome-wide LD between 1000 Genomes individuals from the Paisa population and other cohorts obtained from Europe and South America. We found that the Paisa had considerably higher overall ranked levels of LD and greater levels of rare variation than all of the European and most of the South American populations. As discussed, these findings have string correlations with population history.

In chapter 6, we place our statistical findings of candidate genes from linkage and case-control analysis in a biological context. We also discuss the utility of our LD and rare variation comparisons of our population genetics analysis, and how this can be complimented with future studies that can aid knowledge of complex disease.

References

[1] J.J. Condemi. The Autoimmune Diseases. *JAMA.* 268(20):2882-92, 1992.

[2] S.M. Hayter and M.C. Cook. Updated assessment of prevalence, spectrum and case definition of autoimmune disease. *Autoimmun Rev,* 11(10):754-65, 2012.

[3] G.S. Cooper, M.L.K. Bynum, and E.C. Somers. Recent insights in the epidemiology of autoimmune diseases. Improved Prevalence Estimates and Understanding of Clustering of Diseases. *J Autoimmun,* 33(3-4):197-207, 2009.

[4] R. Tozzoli, M.C. Torrentino, and N. Bizzaro. Detecting multiple autoantibodies to diagnose autoimmune co-morbidity (multiple autoimmune syndromes and overlap syndromes): a challenge for the autoimmunologist. *Immunol Res,* 56(2-3):425-31, 2013.

[5] P.S. Ramos, A.M. Shedlock and C.D. Langefeld. Genetics of autoimmune diseases: insights from population genetics. *J. Hum Genet,* 60(11):657-64, 2015.

[6] American Autoimmune Related Diseases Association (AARDA) and National Coalition of Autoimmune Patient Groups (NCAPG). The Cost and Burden of Autoimmune Disease: The Latest Front in the War on Healthcare Spending. http://www.diabetesed.net/page/_files/autoimmune-diseases.pdf , 2011.

[7] S.C.L. Gough and M.J. Simmons. The HLA Region and Autoimmune Disease: Associations and Mechanisms of Action. *Curr Genom,* 8(7):453-65, 2007.

[8] J.M. Anaya, J. Castiblanco, A. Rojas-Villarraga A, et al. The multiple autoimmune syndromes. A clue for the autoimmune tautology. *Clin Rev Allergy Immunol,* 43(3):256-64, 2012.

[9] M.F. Seldin. The Genetics of Human Autoimmune Disease: A Perspective of Progress in the Field and Future Decisions. *J Autoimmun*, 64(11):1-12, 2015.

[10] G. Lettre and D. Rioux. Autoimmune Diseases: Insights from genome-wide association studies. *Hum Mol GenGenet,* 17(2):116-21, 2008.

[11] G. Gibson. Rare and common variants: twenty arguments. *Nat Rev Genet,* 13(2):135-45, 2012.

[12] P.A. Vischer, M.A. Brown, M.I. McCarthy, and J. Yang. Five Years of GWAS Discovery. *Am J Hum Genet,* 90(1):7-24, 2012.

[13] A.B. Begovich, V.E.H. Carlton, L.A. Honigberg, et al. A Missense Single Nucleotide Polymorphism in a Gene Encoding a Protein Tyrosine Phosphotase (*PTPN22)* is Associated with Rheumatoid Arthritis. *Am J Hum Genet,* 75(2):330-37, 2004.

[14] The International Consortium for Systemic Lupus Erythematosus Genetics (SLEGEN), J.B. Harley, M.E. Alarcón-Riquelme et al. Genome-wide association scan in women with systemic lupus erythematosus identifies susceptibility variants in *ITGAM, PXK, KIAA1542* and other loci. *Nat Genet,* 40(2):204-10, 2008.

[15] C.J. Lessard, J.A. Ice, I. Adrianto et al. The Genomics of Autoimmune Disease in the Era of Genome-wide Association Studies and Beyond. *Autoimmun Rev*, 11(4): 267-75.

[16] S. Lee, G.R., Abecassis, M. Boehnke, and X. Lin. Rare-Variant Association Analysis: Study Designs and Statistical Tests. *Am. J. Hum. Genet*, 95(1):5-23, 2014.

[17] J. Ott, J. Wang and S.M. Leal. Genetic linkage analysis in the age of whole-genome sequencing. *Nat Rev Genet*, 16(5):275-84, 2015.

[18] D.J. Liu, and S.M. Leal. A Novel Adaptive Method for the Analysis of Next-Generation Sequencing Data to Detect Complex Trait Associations with Rare Variants Due to Gene Main Effects and Interactions. *PLoS Genet*, 6:e1001156. https://doi.org/1.1371/journal.pgen.1001156, 2010.

[19] M. Arcos-Burgos, and M. Muenke. Genetics of Population Isolates. *Clin Genet*, 61(4):233-247, 2002.

[20] M.L. Bravo, C.Y. Valenzuela, and O.M. Arcos-Burgos. Polymorphisms and phyletic relationships of the Paisa community from Antioquia (Colombia). *Gene Geogr*, 10(1):11-7, 1996.

[21] L.G. Carvajal-Carmona, I.D. Soto, N. Pineda, et al. Strong Amerind/White Sex Bias and a Possible Sephardic Contribution among the Founders of a Population in Northwest Colombia. *Am. J. Hum. Genet*, 67(5):1287-1295, 2000.

[22] N.R. Mesa, M.C. Mondragón, I.D. Soto, et al. Autosomal, mtDNA, and Y-Chromosome Diversity in Amerinds: Pre- and Post-Columbian Patterns of Gene Flow in South America. *Am. J. Hum. Genet*, 67(5):1277-1286, 2000.

[23] G. Bedoya, P. Montoya, J. Garcia, et al. Admixture dynamics in Hispanics: A shift in the nuclear genetic ancestry of a South American population isolate. *Proc Natl Acad Sci USA*, 103(19):7234-39, 2006.

[24] L.G. Carvajal-Carmona, R. Ophoff and S. Service. Genetic demography of Antioquia and the Central Valley of Costa Rica. *Am J Hum Genet*, 112(5-6):534-41, 2003.

[25] I.J. Barnett, S. Lee and X. Lin. Detecting Rare Variant Effects Using Extreme Phenotype Sampling in Sequencing Association Studies. *Genet Epidemiol*, 37(2):142-51, 2013.

[26] G.M. Peloso, D.J. Rader, S. Gabriel, et al. Phenotypic extremes in rare variant study designs. *Eur J Hum Genet*, 24(6):924-930, 2016.

[27] C. Gu, A. Todorov and D.C. Rao. Combining Extremely Concordant Sibpairs with Extremely Discordant Sibpairs Provides a Cost Effective Way to Linkage Analysis of Quantitative Trait Loci.

[28] R. Shen, J.B. Fan, D. Campbell, et al. High Throughput SNP genotyping on universal bead arrays. *Mutat Res*, 573(1-2):70-82, 2005.

[29] S. Bennett. Solexa Ltd. *Pharmacogenomics*, 5(4):433-8, 2004.

[30] S.T. Bennett, C. Barnes, A. Cox, et al. Towards the $1000 Genome. *Pharmacogenomics,* 6(4): 373-82, 2005.

[31] M. Baker. *De Novo* genome assembly: what every biologist should know. *Nat Methods,* 9(4):333-337, 2012.

[32] The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature,* 526(7571):68-74, 2015.

[33] M. Choi, U.I. Scholl, and R.P. Lifton. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc Natl Acad Sci USA,* 106(45):19096-101, 2009.

[34] Y. Yang, D.M. Muzny, J.G. Reid, et al. Clinical Whole-Exome Sequencing for the Diagnosis of Mendelian Disorders. *NEJM,* 369(16):1502-11, 2013.

[35] C.S. Ku, N. Naidoo, and Y. Pawitan. Revisiting Mendelian disorders through exome sequencing. *Hum Genet*, 129(4):351-70, 2011.

[36] M.J. Bamshad, S.B. Ng, A.W. Bigham, et al. Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Rev Genet,* 12(11):745-55, 2011.

[37] M. Lek, K.J. Karczewski, E.V. Minikel, et al. Analysis of protein-coding variation in 60,706 humans. *Nature*, 536(8):285-91, 2016.

[38] K.J. Karczewski, B. Weisberd, B. Thomas, et al. The ExAC browser: displaying reference data information from over 60 000 exomes. *Nucleic Acids Res,* 45(1):840-45, 2017.

[39] W. Fu, T.D. O'Connor, G. Jun, et al. Analysis of 6515 exomes reveals recent origin of most human protein-coding variants. *Nature*, 493(1):216-20, 2013.

[40] H. Li. Towards better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics,* 30(20):2843-51, 2014.

[41] L.E. Mose, M.D. Wilkerson, D.N. Hayes, et al. ABRA: improved coding indel detection via assembly-based realignment. *Bioinformatics*, 30(19):2813-15, 2014.

[42] A.L. Hoffman, J. Behr, J. Singer, et al. Detailed simulation of cancer exome sequencing data reveals differences and common limitations of variant callers. *BMC Bioinform,* 18:8, https://doi.org/10.1186/s12859-016-1417-7, 2017.

[43] M.H. Li, J.H. Abrudan, M.C. Dulik, et al. Utility and limitations of exome sequencing as a genetic diagnostic tool for conditions with pediatric sudden cardiac arrest/sudden cardiac death. *Hum Genomics,* 9(1):15 doi:10.1186/s40246-015-0038-y, 2015.

[44] A. Hamington, M. Tétreault, D.A. Dyment, et al. Concordance between whole-exome sequencing and clinical Sanger sequencing: implications for patient care. *Mol Genet Genomic Med*, 4(5):504-12, 2016.

[45] J.M. Schwarz, C. Rodelsperger, M. Schuelke, and D. Seelow. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat. Methods,* 7(8):575-6, 2010.

[46] J.M. Schwarz, D.N. Cooper, M. Schuelke, and D. Seelow. MutationTaster2: mutation prediction for the deep sequencing age. *Nat. Methods,* 11(4):361-2, 2014.

[47] I.A. Adzhubhei, S. Schmidt, L. Peshkin, et al. A method and server for predicting damaging missense mutations. *Nat Methods*, 7(4):248-9, 2010.

[48] G.N. Ramachadran, C. Ramakrishnan, and V. Sasisekharan. Stereochemistry of polypeptide chain configurations. *J. Mol Biol*, 7(1):95-9, 1963.

[49] S. Henikoff and P.C. Ng. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res,* 31(13):3812-4, 2003.

[50] Y.M. Di, E. Chan, M.W. Wei, et al. Pediction of Deleterious Non-synonymous Single-Nucleotide Polymorphisms of Human Uridine Diphosphate Glucouronosyltransferase Genes. *AAPS J,* 11(3):469, 2009.

[51] H. Hu, J.C. Roach, H. Coon, et al. A Unified Test of Linkage Analysis and Rare-Variant Association for Pedigree Sequencing Data. *Nat Biotechnol*, 32(7):663-9, 2014.

[52] M. Fishelson and D. Geiger. Exact genetic linkage computations for general pedigrees. *Bioinformatics,* 18(suppl 1):S189-S198, 2002.

Chapter 2: General Framework to Detect Causative Autoimmune Genes and Comparing Genomic Variations Within and Between Population Isolates

Abstract

Identifying causative mutations using rigorous analysis is key for understanding the genetic aetiology of all complex diseases including autoimmunity. In this chapter, a combined framework from genetic, statistical and bioinformatics resources is elaborated. Essential components and algorithms part of the framework include extreme phenotypes (EPs), genetic isolates, linkage and association analysis and biological networks. The key properties and mathematical theory behind each of these elements that enhances their utility in complex disease studies is explained in detail. This comprehensive strategy ensures that thorough analysis is conducted genetic, statistical and physiological perspective.

2.1 Extreme Phenotypes from a Genetic Isolate

The crucial element for genetic analysis in searching for disease-causing variants is the quality of phenotyping for patients carrying autoimmunity. The advantages of the sampling efforts by Dr. Arcos-Burgos and Dr. Juan-Manuel Anaya are that many patients, part of the Paisa isolate are carrying extreme phenotypes of autoimmunity, manifesting as MAS Proofs based on mathematical inductions plus probabilistic and statistical axioms, support the notion that the likelihood of identifying causative rare variants is greater via EP approaches than otherwise obtained from random sampling of cases and controls. Hence, this means that minor allele frequencies (MAFs) of rare causal variants are more likely to be enriched via extreme phenotype sampling (EPS), compared to affected members of the given population [1]. Frequencies of given rare causative variants is conditioned on the effect size of given genotypes, regression coefficients for the analysed variants and covariates, genotypic effect size, threshold for phenotypic cut-off as well as the mutation's frequency in other members (carrying non-EPs) of studied populations. Using the phenotypic threshold cut-off (for severe and non-severe phenotypes respectively) and the conditioning on the premise that the studied genotype has a positive (causative) effect on phenotypes, the proof of rare-variant enrichment within EPs is obtained. The first step in this procedure is to model the phenotype as follows:

$$y_i = a_0 + X'\alpha + G'_i\beta + \varepsilon_i \text{ (2.1)}$$

In the above equation, Yi is the coded phenotype, whilst α and β are the vectors of regression coefficients for the covariates represented by X and genetic variants denoted by G respectively [1]. The term a0 represents the intercept value, a quantity that contributes to the phenotypic variance when all other terms are equal to 0. The final variable is the error term, representing the quantity of phenotypic variance not accounted for by the regression coefficients and the intercept value. It is a number, calculated as a function of the variance of the phenotypic term, y_i. Thus, the higher the value of the beta regression coefficient, the greater is the strength of the phenotype-genotype association for a given variant. The genotype and phenotype terms derived from this equation are then used to prove the notion hypothesised for EPS vs. non-EPS sampling, for the single causal variant/no covariate model [1]. Thus the covariates term and y-intercept values are equal to 0, giving:

$$Yi = G'_i\beta + \varepsilon_i \text{ (2.2)}$$

The likelihood of present variants under EPS can now have a rearranged mathematical expression [1]. The probability that the genotype (G) contains the causative variant (either in homozygous or heterozygous form), given the presence of the EPs (i.e. y > c, where c is the phenotypic cut-off), can be equivalently expressed by the equation on the right hand side in (2.3. In this case, P(y>c) is conditioned on G [1]. This is derived with the assumption, beta > 0 (i.e. variant has causative effects).

$$\Pr(G > 0)\,|(y > c) = \Pr(G > 0)\frac{\Pr(y > c)\,|(G > 0)}{\Pr(y > c)} \text{(2.3)}$$

To achieve the proof of concept, we must find Pr(y>c)|(G>0) – Pr(y>c) which equals:

$$= \mathrm{Pr}(y>c)|(G>0) - \mathrm{Pr}(y>c|G=0)\mathrm{Pr}(G=0) - \mathrm{Pr}(y>c|G>0)\mathrm{Pr}(G>0) \quad (2.4.1)$$

$$= (\mathrm{Pr}(y>c)|(G>0) - \mathrm{Pr}(y>c)|(G=0))\mathrm{Pr}(G=0) > (\mathrm{Pr}(y>c)|(G=1) - \mathrm{Pr}(y>c|G=0))\mathrm{Pr}(G=0)$$

$$= (\mathrm{Pr}(\beta + \varepsilon > c) - \mathrm{Pr}(\varepsilon > c))\mathrm{Pr}(G = 0) \quad (2.4.3)$$

Once this is obtained, it is necessary to determine the probability of the disease trait exceeding the cut-off for EPs (y>c), conditioned on the inference that the genotype is carrying the alternate, causative allele, in a heterozygous or homozygous state [1]. Later, the equation can be reorganized, enabling us to subtract from this value, the likelihood that the potentially causative allele is carried in a heterozygous state only. Thus the overall difference between the 2 conditional probabilities is given. As this mathematical proof is conducted via an additive model, this explains the outcome of the steps b and c in the equation sequence. Therefore, it is intuitive that Pr(y>c)|(G>0) > Pr(y>c)|(G=1), as the conditioned probability term on the LHS of the inequality accounts for the presence of homozygous genotypes, which in additive modes of inheritance will theoretically generate a greater likelihood of EPs [1]. However, this same expression also clarifies that this theorem can also be proven via the dominant and recessive models (as G=1, only considers heterozygotes). Furthermore, it is apparent that if the genotype isn’t causative, Pr(G>0) will be the same when conditioned under EPs or otherwise. Thus the MAF of variants under EPS is the same as the general population.

When genotypes are causative, this should lower the value of Pr(G=0) and Pr(epsilon > c) in EPs. Simultaneously, the regression coefficient should have a significantly high value, raising the probability that the beta coefficient and error term will contribute a large proportion to the value of y, such that y > c, compared to the likelihood when beta = 0 (i.e. non-causative genotypes) [1]. Subsequently, as the covariates = 0, it is clear that Pr(y>c)|(G>0) > Pr(y>c). Hence, it is proven that the MAF of causal variants (or likelihood to identify causal variants) is greater in EPs, relative to the remaining sampled population.

The proof of concept can be extended to multiple causative disease variants models.

The key variables for determining the MAF of pathogenic variants in EPs have been previously mentioned [1]. One of these variants, the genotype effect size, can be derived from the strength of the phenotype-genotype association (a), and the logarithm of the MAF as follows:

$$\beta j = \text{-alogMAF} \quad (2.5)$$

Therefore, based upon the outcome of single variant case, as well as the effect size and phenotype-genotype association parametric quantity (a), the proof is also replicated in the case of polygenic traits (i.e. Pr(y>c)|(Gj=g) > Pr(y>c). When obtaining Pr(y>c)|(Gj=g) and Pr(y>c) for in the multilocus model, the cross-product for genotype probabilities [1], plus likelihood of all variables contributing to y > c, summed across all tested loci is required, as expressed below: This avoids inflation of the error term and other variables, including those from other genotypes not corresponding to the jth variant.

$$\Pr(y > c)\,|\big(G_j = g\big) - \Pr(y > c) =$$

$$\Sigma_{g1=0}^{2} \ldots \Sigma_{gp=0}^{2} \phi\big(\alpha\beta_j + \Sigma_{l=j}\beta_l g_l - c\big)\Pi \Pr(G_l = g_l) - \Sigma_{g1=0}^{2}\Sigma_{gp=0}^{2}\phi(\Sigma_{l=1}^{p}\beta_l g_l - c)$$

(2.6)

Hence, similar to the previous example, the parameter incorporating alpha and beta must be sufficiently large for y (phenotype) to exceed c, under the assumption beta > 0 [1]. Hence, presence of EPs is likely to yield greater enrichment of potentially deleterious genotypes, compared to minor other sampled individuals, as supported by mathematic evaluation. Extending equation 3 and 4 to the multiple variants assumption, once the probability of genotypes is obtained, expected MAF can be derived, based upon the following formula:

$$\Pr(G > 1)\,|(y > c) = \frac{E\big(G_j\big)(y > c)}{2}$$

$$= 0^{*}\Pr(G = 0)|(y > c) + 0.5^{*}\Pr(G = 1)|(y > c) + 1^{*}\Pr(G = 2)|(y > c)$$

(2.7)

The formula takes accounts for the presence of compound heterozygotes, thereby explaining the inclusion of 0.5*Pr(G=1)|(y>c) in the equation, as well as homozygotes [1]. In this particular case, a recessive model is being tested, from which the estimated frequency of 2 alleles present in given loci (i.e. compound heterozygote or homozygote state) can be calculated. Thus, the mathematical evidence shows the benefits of EPS from the Paisa community, in addition to their genetic homogeneity.

2.2 Principal Component Analysis

Nevertheless, in case-control designs, small levels of microdifferentiation can still have confounding effects in case-control studies. One way to control for this is the application of Principal Component Analysis (PCA) [2-4]. This requires an orthogonal transformation of linearly correlated variables into uncorrelated variables/factors. The purpose of this is to identify potential evidence of population structure across all variants and individuals used during analysis. In brief, this firstly involves representation of a database as a matrix X containing m markers and n individuals [2-4]. This matrix is later decomposed into USV^T whereby S is considered as the n by n matrix for which the eigenvalues (values denoting the quantity of variation explained by a single component or dimension in the dataset) of the original matrix entries are present, V is a matrix of the principal components, or in this case 'ancestral vectors' relating genotypes to samples (i.e. eigenvectors or axes of variation, corresponding to each individual sample) [2-4]. U represents the coordinate values for a given eigenvector. Given that the purpose is to find a potential mathematical relation between these elements of the matrix (i.e. individuals and markers) these eigenvectors can be thought of as genotypic ancestry components in PCA [2-4]. The matrix decomposition is expressed as:

$$X^T X = VS^T U^T USV^T = VS^T SV^T, \text{(2.8)}$$

Where ^{T} is the matrix transposition. It is known that V is an eigenvector matrix, which forms the principal components [2-4]. Given that the eigenvalues and eigenvectors of the m by n matrix in X are represented by lambda$_k$ and v$_k$ respectively, it is known that:

$$X^T X v_k = \gamma_k v_k \text{ (2.9)}$$

This step is followed by correction of individual data elements by markers. Assume that g is a genotype vector value (response variable), corresponding to other independent variables [2-4]. The average value for all g vectors is subtracted from each element. This gives the value (written as g_0), for the first element of g (g_1), and g_{k-1} for the kth g vector [2-4]. Therefore, for all g, we can now calculate the corrected value of g_k ($g_{k\ \text{corrected}}$):

$$g_{k\ \text{corrected}} = g_{k-1} - \gamma_k v_k \text{ (2.10)}$$

$_k$ is derived from the eigenvectors of the original marker and sample matrix, as well as individual genotype vectors. Thus, further derivation gives:

$$\gamma_k = v_k . g_{k-1} \text{ (2.11)}$$

The value of γ_k in turn is also proportional to quantities of the variance-covariance matrix, quantifying the mathematical relation between every genotype and sample, summed across all individuals [2-4]. This equation is expresnnnned as:

$$\Sigma_j a_{jk} g_{j(k-1)} / \Sigma_j a^2_{jk} \text{ (2.12)}$$

Thus, after decomposition of the original matrix, mathematical relations between each dimension or principal component, upon which the data points are projected can be quantified, using a variance-covariance matrix [2-4]. In this case, a is the value showing these between-dimensional relations, specifically between genotypes and samples in the variance-covariance matrix, for j individuals and k markers [2-4]. This enables derivation of the Euclidean distances of each of the coordinates (markers) within the multi dimensional space, constructed by multiple dimensions from the eigenvectors.

Finally, from all of the above formulas, principal components accounting for the largest proportion of variance in a given dataset can be determined, by identifying the eigenvectors with the largest eigenvalues [2-4]. This facilitates identification and subsequent adjustments for potential population structure. Hence, PCA was applied to the analyses of the sporadic autoimmune cases against the sampled unaffected controls, whereby diagrammatic plots can ensure minimisation of differentiation between disease carriers and non-carriers [2-4]. The PCA quality control algorithm is implemented in EIGENSTRAT software, from the SVS Golden Helix Suite [2].

2.3 Whole Exome Capture, Illumina Next Generation Sequencing and Sequence Read Quality Evaluation:

Whole exome capture was performed via the Axlegen and Nimblegen kits, and sequencing was conducted via Illumina NextGen Hiseq facilities at the BRF (Biomolecular Resource Facility) for all samples [5-8]. The FASTX Toolkit was then implemented in trimming of adaptors and removal of low quality bases. Once the FASTQ files are generated, GATK (Genome alignment ToolKit) Best practices pipeline is applied for sequence read alignment (using BOWTIE). This generates a sequence alignment map (SAM) file, which are later converted into a binary (BAM) aligned format. Duplicates that emerge through PCR amplification in the sequencing by synthesis process are removed, after this step and base scores are recalibrated, along with the variant calls derived from the alignment maps [9-11]. Accuracy of variant calls can be obtained by the mapping quality and Phred score values, using the Integrative Genomics Viewer (IGV) software. The mapping quality determines the likelihood of the read being aligned to the correct portion of the reference sequence. The Phred score is the probability of the correct base call from the GATK Pipeline [12-13]. Both of these measures were used for quality control purposes in our sporadic and familial autoimmune studies.

2.4 Unified Combination of Linkage and CLRT (Composite Likelihood Ratio Test)

Firstly, the VCF files containing variants for each sequenced individual were converted into GVF (Genome Variant Format) files using the vaast_converter script, part of the pVAAST package [14]. After that, the GVF files are subsequently annotated with the Variant Annotation Tool (VAT) Perl script in order to determine the variants in each individual that are non-synonymous, found in splice sites, intronic etc. This annotation process is based through comparison of variants against the reference human genome sequence (hg19) [14].

After implementation of VAT, the annotated variants for each family member are unified into a single condensed variant file (cdr file). There is one CDR file generated for each family, and once these files are generated, we then proceed to construction of the pedigree file. Once these input files are generated we then perform the pVAAST analysis, whilst simultaneously integrating individuals from the 1000 genomes database as our control population. The linkage parameters for mode of inheritance are set in a parametric file [14]. Within the same file, we also activate the detection of de novo mutations, using the default mutation rate per site per generation in the human genome. This final step is the component of analysis for generation of Logarithm of Odds (LOD) and CLRTp scores (using the VAAST script), which are produced through integration of linkage information and composite likelihood ratio (CLRT data) test data. Note also that the CLRT in turn includes information about the severity of amino acid substitutions [14].

2.4.1 Mathematical Algorithm of Linkage Analysis

Linkage analysis was conducted for 47 individuals across 10 families subjected to WEC and NGS from the Paisa community. In total, there were 32 affected and 15 unaffected individuals (full study described in chapter 4). In brief, the algorithm assumes that the observed disease phenotype is caused by either a latent (unlinked) locus or loci under investigation (current loci). The null hypothesis is assuming that only the latent locus is causal, of the observed disease phenotype. The alternate hypothesis is that either one of the loci could be causal [15]. The calculation is derived, based upon adaptations of the Elston Stewart algorithm [15-16]. This is different to the Lander-Green algorithm, whereby Markov-Chain formulas are used for determination of genotype and Identity by Descent (IBD) probabilities for inheritance vectors at candidate loci (v_{i+1}), derived by conditioning genotypic IBD states and observed phenotypic data, at a known locus (v_i) and transition matrix values (between v_i and v_{i+1} for loci M_i and M_{i+1}). The probability of observed phenotypes across all possible values of inheritance vectors are also required [17]. This is evidently a Markov-Chain algorithm as the probability calculations of inheritance vectors and genotypic states in v_{i+1} are dependent on those in $v_{i,}$ to obtain the eventual recombination numbers and fraction [17]. In contrast, the Elston-Stewart algorithm relies on exact likelihood derivations, via recursive genotype probability calculations from observed data across each generation of given pedigrees. Firstly, the likelihood of the obtained genotype is determined conditional on the parental and founder genotypes, as well as the probability of an individual's phenotype given their respective genotype (thereby deriving the genotypic disease probability or penetrance) and phenotypes of the descendants [15-16].

$$\mathrm{L} = \Sigma \mathrm{G}_1 \Sigma_\mathrm{n} \mathrm{Pr}(\mathrm{X}_\mathrm{i}|\mathrm{G}_\mathrm{i}) \Pi \mathrm{Pr}(\mathrm{G}_\mathrm{founder})\, \mathrm{Pr}(\mathrm{G}_0|\mathrm{G}_\mathrm{f}, \mathrm{G}_\mathrm{m}) \ (2.13)$$

After that, the conditional probability of the genotypes and the observed phenotypes for the pedigree data (as per the Elston-Stewart algorithm), given the present allele frequencies in the analysed cohort as well as the genotypic disease probability is then obtained [15-16]. These likelihood values (under the null and alternate models) are calculated for the current and latent loci respectively by:

$$\mathrm{L} = \mathrm{Pr}(\mathrm{g}_\mathrm{c}, \mathrm{g}_\mathrm{l}, \mathrm{p})|(\mathrm{P}_\mathrm{c}, \mathrm{P}_\mathrm{l}, \mathrm{f}_\mathrm{c}, \mathrm{f}_\mathrm{l})(2.14)$$

The variables g_c, g_l, f_c, f_l and P_c, P_l are the genotypes, frequencies and genotype disease probabilities (or penetrance) of the current and latent loci respectively, and p is the phenotypes [15]. This calculation is performed under the null and alternative hypothesis, for derivation of the LOD score.

$$\Sigma_{\mathrm{i}=1}^{\mathrm{n}} \log_{10} \Pr \left(\frac{\mathrm{Obs|Linkage}}{\mathrm{Obs|No\ Linkage}} \right) (2.15)$$

This will identify variants within the gene that maximise the gene-based LOD score, across all pedigrees.

2.4.2 Extension of Analysis to Compound Heterozygotes

The advantage of pVAAST implementation is that under a recessive model, it is also possible to evaluate compound heterozygotes during the joint linkage and association algorithm [15]. Compound heterozygotes are evaluated in such a way that one doesn't falsely assume independence in the genotype disease probabilities of heterozygous variant sites. Instead, one constructs Boolean risk vectors for multi site genotypes for heterozygous variant sites, denoting them as either D (conferring disease susceptibility) or N (neutral) [15]. Using these risk vectors, it is then possible to implement a joint likelihood function, conditional on the probability that individual with a risk and neutral associated genotypes respectively, are affected, as follows:

$$\mathrm{L_i} = \mathrm{Pr}^{\mathrm{na}}(1 - \mathrm{Pr})^{\mathrm{nb}}\mathrm{Pn}^{\mathrm{nc}}(1 - \mathrm{Pn})^{\mathrm{nd}} \ (2.16)$$

Pr and Pr^{n} represent the probabilities that individuals with risk and neutral genotypes are affected, where as na and nb are total number of affected/unaffected individuals with risk genotypes and nc, nd are corresponding, with neutral genotypes. Multiple iterations are later conducted, randomly switching Boolean risk vectors of variant sites, until likelihood calculations are converged [15]. Markov chains are subsequently constructed from these risk vector likelihoods. Furthermore, using the summed indicator functions ($I_D(k)$) to denote the causality and predicted deleteriousness of all selected variants, the functional score (F) can be used to generate the modified likelihood for compound heterozygotes:

$$\mathrm{F} = \frac{\Sigma_{\mathrm{k}} \mathrm{CLRT}_{\mathrm{v(k)}} . \mathrm{I}_{\mathrm{D}}(\mathrm{k})}{2\Sigma_{\mathrm{k}} \mathrm{I}_{\mathrm{D}}(\mathrm{k})}$$

Therefore, the most likely risk vector can be selected and used for the aforementioned LOD score formula under the recessive model [15].

2.4.3 Inclusion of De Novo Mutations

The predominant drawback of many linkage analysis algorithms is that de novo mutations occurring in the offspring but absent within the parents are often disregarded as Mendelian Errors [15, 18]. However, in this case these Mendelian errors are given a transmission probability of 1.2 x 10^{-8}, which from previous studies is established as the mutation rate per site per generation in the human genome. These de novo mutations are then incorporated during permutation and gene-drop simulation [15].

2.4.4 CLRT: Integration of Case-Control Algorithms

The composite likelihood ratio test (CLRT) statistic is derived from 2 components: the severity of amino acid substitution (in the case of non synonymous variants) as well as indel and or splice site severity weights (which depend on the levels of sequence conservation between residues) and the allele frequency differences between cases vs controls [19]. The controls in this case are derived not only from the unaffected members of the pedigrees, but from exome sequencing data within the 1000 genomes population (from now onwards referred to as the background population) [20].

In the case of non-synonymous variants, the potential deleterious nature of amino acid substitution is calculated using an algorithm, similar to the derivation of the block substitution matrix (BLOSUM) [21]. This involves deriving the observed and expected values of amino acid frequencies when performing multiple sequence alignments of the query and target protein sequences as well as target sequence alignments with other proteins within the same conserved family [19, 21]. This calculation is performed for the background population as well as the affected individuals. This value, obtained when sampling the affected individuals is then expressed as a ratio compared to the proportion of cases in the OMIM (Online Mendelian Inheritance in Man) database database, specifying the given amino acid substitution as a disease causing variant [19, 22-23]. Both these values (denoted as hi and ai) are then included in the calculation of the CLRTv test statistic.

When evaluating non-protein coding variants, such as annotated splice sites, their importance in functionality is evaluated by transition matrix scores, based upon the conservation of sequence data between various species [19]. Once again, this is calculated for the background population and affected pedigree members. The latter of these two values is then compared to the variant proportions in the OMIM database. Both values (ai and hi) are included in the calculation of the CLRTv test statistic for the alternate and null hypothesis respectively [19, 22-23].

Once these components are obtained, the following variables are required to calculate the test statistic:

m - the number of variant collapsing categories or groups (i.e. a set of multiple grouped variants) [19].

$\mathbf{l_{k+m}}$ – Number of individual collapsed variants from various sites in each of the mth collapsing category [19].

$\mathbf{ni^A}$ and $\mathbf{ni^U}$ – Total number of affected and unaffected individuals respectively [19].

k - the number of variant sites that are not grouped or collapsed with other variants [19]

$\mathbf{Pi^A}$ and $\mathbf{Pi^U}$ - minor allele frequencies of affected and unaffected individuals for collapsed and uncollapsed variant sites respectively [19].

Xi^{A} and Xi^{U} - The number of affected and unaffected individuals containing given variant allele(s) (at uncollapsed sites) or multi-locus allele(s) (for collapsed variant sites). [19]

These components are incorporated into the derivation of the test statistic as below:

$$\lambda = \ln(\frac{\mathrm{L_{Null}}}{\mathrm{L_{Alt}}}) \ (2.18)$$

$$= \left[\frac{\mathrm{hi(pi)^{Xi}(1-pi)^{2lini-Xi}}}{\mathrm{ai(pi^{U})XiU(1-pi^{U})^{2lini^{U}-Xi^{U}}\left(pi^{A}\right)^{Xi^{A}}\left(1-pi^{A}\right)^{2lini^{A}-Xi^{A}}}}\right] (2.19)$$

The calculation simplifies to the likelihood under the null (numerator) vs. alternate (denominator) hypothesis [19]. Equation shows that the CLRTv score is conditional on the observed minor allele frequencies across all, affected and unaffected individuals at a given variant or variant collapsing site. Hence statistic based on allele frequencies plus collapsed variant sites under null and alternate model is summed across all collapsed and uncollapsed variant sites [19]. As well as minor allele frequencies, the calculation also takes into account the functionality of non-synonymous variants. The variable *hi* is the proportion of expected amino acid substitutions in the background 1000 genomes dataset for a given pair when aligned with proteins in the BLOCK database [21], where as *ai* is the observed/expected frequency of aa pairs in affected individuals relative to the known disease causing mutations involving these specific aa substitutions in the OMIM database [22-23].

2.4.5 Derivation of the Combined Linkage/Association Test Statistic

The combined test statistic is obtained through combination of the summatory LOD score (calculation explained above) along with the CLRTv value. The natural logarithm of the CLRTv score multiplied by 2 is subtracted from the summatory LOD score [15].

$$\mathrm{CLRT_p} = 2^{*}\ln 10(\Sigma_{\mathrm{i=1}}^{\mathrm{n}}\mathrm{LOD_i}) - 2\lambda \quad (2.20)$$

This calculation is performed across all families from i =1 to i = n.

2.4.6 Evaluation of Statistical Significance

After generation of the test statistic, statistical significance is determined by randomization and gene drop simulation [15]. Firstly, affection status is shuffled and the expected genotype frequencies are derived conditional on parental alleles [15]. The score is calculated after each permutation. In total 10000 permutations were conducted, based on Monte Carlo simulations.

2.5: Evaluating Pathogenicity of Candidate Variants and Variant Quality Scores

Following the Linkage and Rare-Variant Association Analysis, variants were further evaluated based on their potential functional impact. This was achieved using SIFT [24], Polyphen2 [25-26], MutationTaster [27], and Provean [28-29].

2.5.1 Variant Effect Prediction Algorithms

Classification algorithms across all of these variant effect predictors are performed using any of the following criteria:

UniProt- When a query amino acid sequence of a protein is submitted to Uniprot, 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 [25-26, 28-29]. For example, changes between 2 amino acids in a sequence that are hydrophilic are tolerated, but changes from hydrophilic to hydrophobic and vice-versa are considered as damaging.

PSIC (Position Specific Independent Counts) profiles- Present in the Polyphen variant classifier, PSIC profiles compute the probability that any given variant amino acid is likely to be found at a particular position in the protein, after Polyphen2 identifies homologous sequences from the BLAST search of the UniRef 100 database [25]. Sequence alignments are maintained if the alignment length is greater than 75 amino acids and if the sequence identity is between [30] and 94% [25-26]. Thus regions of the protein that are prone to amino acid changes in regions that have a high level of conservation are likely to have an important function, and therefore are more likely to be damaging [25-26]. This and the Uniprot-based approach are implemented in the PolyPhen predictor.

Sequence Homology and Conservation. This approach uses a normalized transition probability matrix, i.e. based on the amino acid position what is the probability that the amino acid (which in this case is the state), will change states to another particular amino acid. Those with low transition probabilities will be deleterious [24]. These algorithms are predominantly seen in SIFT [24]. Further extensions of this are evident in MutationTaster, whereby sequence conservation at splice sites as well as amino acid substitutions are evaluated [27]. Also, using various biomedical databases, sequence changes that influence mRNA levels are identified [27].

Sequence Cluster Alignment Score: The score is derived in the PROVEAN algorithm. It quantifies the difference between a reference and variant protein sequence, against a comparator homologue when a variation (insertion, deletion or amino acid substitution) is introduced [28-29]. The score is calculated across all sequence clusters. Firstly, all homologous and distantly related proteins are clustered. The delta alignment scores for individual sequences are dependent upon BLOSUM (in turn derived from the observed and expected frequencies of amino acid substitutions of the Block database, containing multiple alignments of conserved regions in protein families) [21, 28-29]. The calculation for the delta alignment values is:

$$\Delta(Q, v, S) = A(Q', S) - A(Q, S) (2.21)$$

A(Q'S) and A(Q,S) are the alignments scores for the reference and alternate variant sequence against the comparator respectively, derived from the BLOSUM62 substitution matrix values [28-29]. Low delta scores indicate minimal differences between the 2 protein sequences on the semi-global alignments [28]. Hence, any introduced variation amongst this sequenced region is predicted to be deleterious. Scores are summed across all sequence clusters to give the unbiased average delta score.

$$\frac{1}{N}\Sigma_{c=1}^{N}\left(\frac{1}{N_c}\Sigma_{i=1}^{Nc}\Delta_{c,i}\right) (2.22)$$

Calculations can be extended to neighbouring points of a given amino acid sequence position, and therefore applied to INDEL variations [28-29].

2.6 MetaCore Pathway and Network Analysis Workflow

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 in the Metacore software suite, licensed by GeneGo [30]. The network and pathway analysis algorithms are available through a web interface, and the software suite also includes a manually curated gene ontology database. These networks can be constructed via the metacore web interface [30]. The algorithm to build these networks is known as incorporates a subset of 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.

The local networks are prioritised based on the extent of interaction between genes from the input list and other network nodes from the Metacore GeneGo database [30]. The probability that a given number of genes from the input list are present in one of these subnetworks is represented by a hypergeometric distribution, as illustrated in the equation below:

$$\mathrm{P(r,n,R,N)} = \frac{\mathrm{C_R^r . C_{N-R}^{n-r}}}{\mathrm{C_N^n}} = \frac{\mathrm{R!\,.\,(N-R)!}}{\mathrm{N!}} . \frac{\mathrm{n!.\,(N-n)!}}{\mathrm{r!.\,(R-r)!}} . \frac{1}{\mathrm{(n-r)!.\,(N-R-n+r)!}}$$

(2.23)

In this equation, the variables are as follows:

N – The total number of nodes in the Metacore database, collectively known as the global network [30].

R – Nodes in the global network containing genes that match the input list [30].

n – The total number of nodes corresponding to a **single** network under investigation, constructed from the input list. It can be thought of as a subnetwork to the global network [30].

r – A subset of n, corresponding to nodes matching the input genes, within the single analysed network being studied [30]. The sum of the probabilities for values equal to or larger than r, gives the likelihood that genes from the input list are present within the constructed network by chance, under the null hypothesis [30].

$$\mathrm{P - value} = \Sigma_{\mathrm{r}}^{\mathrm{n}} \mathrm{P(r,n,R,N)} (2.24)$$

The probability values generated from equation 3 that form this hypergeometric distribution are also used to calculate the Z score [30]. This is a measure to determine the extent to which each network differs from the mean of the hypergeometric distribution, regarding enrichment of candidate input genes [30]. The Z-score requires mean and variance for these distributions of given networks is obtained as follows:

$$\mu = \frac{\mathrm{n.R}}{\mathrm{n}} (2.25)$$

$$\sigma^2 = \Sigma_{\mathrm{r=0}}^{\mathrm{n}} \mathrm{r}^2 . \mathrm{P(r,n,R,N)} - \mu^2 = \frac{\mathrm{n.R.(N-n).(N-R)}}{\mathrm{N^2.(N-1)}} (2.26)$$

Using these values, the Z-Statistic is calculated by:

$$\mathrm{Z} = \frac{\mathrm{r} - \mathrm{n}\frac{\mathrm{R}}{\mathrm{N}}}{\sqrt{(\mathrm{n}(\frac{\mathrm{R}}{\mathrm{N}})(1-\frac{\mathrm{R}}{\mathrm{N}})(1-\frac{\mathrm{n}-1}{\mathrm{N}-1}))}} = \frac{\mathrm{r} - \mu}{\sigma} (2.27)$$

Thus, in essence, the Z-score is a representation of the magnitude of differentiation of r (number of nodes in a particular network or subnetwork, overlapping with the gene list) from its expected mean value corresponding to the hypergeometric distribution [30].

Within these networks, each node (connected by 2 or more genes) can also match corresponding subsets of gene ontology processes [30]. 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 p values which are calculated based on the size of the intersection between the network process in question and nodes in the Metacore database of interactions (a subset of which contain genes marked by intersection with the input gene list [30]. This p value is a statistical probability genes from the input gene list, within the network would randomly overlap with a particular GeneGo ontology process.
This is calculated by considering the following variables:
N – The total number of nodes within the ‘global network’, with direct physical interactions with the genes from the input dataset [30].
R – Number of nodes that are associated with the GO process category of interest (e.g. Calcium signalling, T cell costimulation etc) [30].
n- The number of nodes under evaluation in the subnetwork (also known as the local network) [30].
r - The number of nodes containing genes in the subnetwork associated with the GeneGo ontology process of interest (subset of n) [30].

$$\frac{R!\,n!\,(N-R)!\,(N-n)!}{N!}\Sigma_{i=\max(r,R_n-N)}^{\min(n,R)}\frac{1}{i!\,(R-i)!\,(n-i)!\,(N-R-n+i)!}\,(2.28)$$

This indicates whether or not there is a random association between the genes matching the input list and the GO category [30]. The p-value to ascertain this is given below (as obtained from the Metacore Sotware Suite, Licensed by Thomson Reuters):

The p-value in this case is defined as the probability that within the nodes of a particular subnetwork (which are subsets of N), there would be a random intersection of genes of the user’s list associated with a particular GeneGO ontology process of size r or larger [30].

The only nodes used for statistical evaluations are those with direct physical interactions with network elements connected with our input genes (i.e. our gene list generated from the filtration strategies described above) [30]. 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 our filtered list.

2.7 Human Gene Connectome (HGC)

The Human Gene Connectome is applied in order to prioritize candidate genes associated to or in linkage with autoimmunity based on their potential functional roles in disease pathogenesis. However, unlike the Metacore database, the HGC is not designed to construct biological pathways or networks for this purpose. Instead, genes are assessed for mechanistic relevance to the disease of interest by generating measurements of biological distance [31-32]. These measures provide an estimate of quantifying the functional relatedness between the candidate genes and core genes

One of the challenges of identifying specific candidate genes for diseases that exhibit a monogenic, Mendelian mode of inheritance is that current bioinformatic methods in combination with high throughput NGS data output 1000s of variants per individual exome sample [31-32]. Hence it is often difficult to identify a single genetic variant that is overrepresented in affected compared to unaffected individuals in monogenic Mendelian traits. This difficulty is further enhanced by the fact that most bioinformatic analyses methods for variant filtration and prioritization analysis don’t contain well developed metrics for estimating the relatedness of genes that are not part of the same pathways or networks [31-32]. Also, by using only genetic means (i.e. population genetic association, variant effect predictors, allele frequency databases), it is difficult to identify morbid variants [31-32]. It is for this purpose the HGC is implemented (to evaluate the degree of physiological homogeneity between a given set of genes which may be implicated in a disease pathogenesis pathway, by estimation of biological distance) [31-32]. The workflow of the algorithms used by the programming scripts in the HGC database is as follows:

Lists of binding and pathway interactions from the Genes of interest in the STRING database are collected. These interactions in STRING are obtained from multiple sources including: Protein Databank, MINT (molecular interactions) and GO (gene ontology). The interactions are then applied to derive a raw quality score which determines the ratio of annotations showing an interaction between a given pair of genes vs. annotations which show no interaction [31-32]. This is illustrated in the following formula:

$$Q = \log \frac{(N_{together} \cdot N_{total})}{(N_{alone1}+1) \cdot (N_{alone2}+1)} (2.29)$$

N together represents the number of times a given set of genes were found to have a binding and/or pathway interaction, N alone denotes otherwise and N total is the full number of interactions for any given gene pair of interest [33-34]. After that, empirical evidence from the quality scores (Q) is used to derive the confidence scores of protein interactions [33-34] (see figure 2.1).

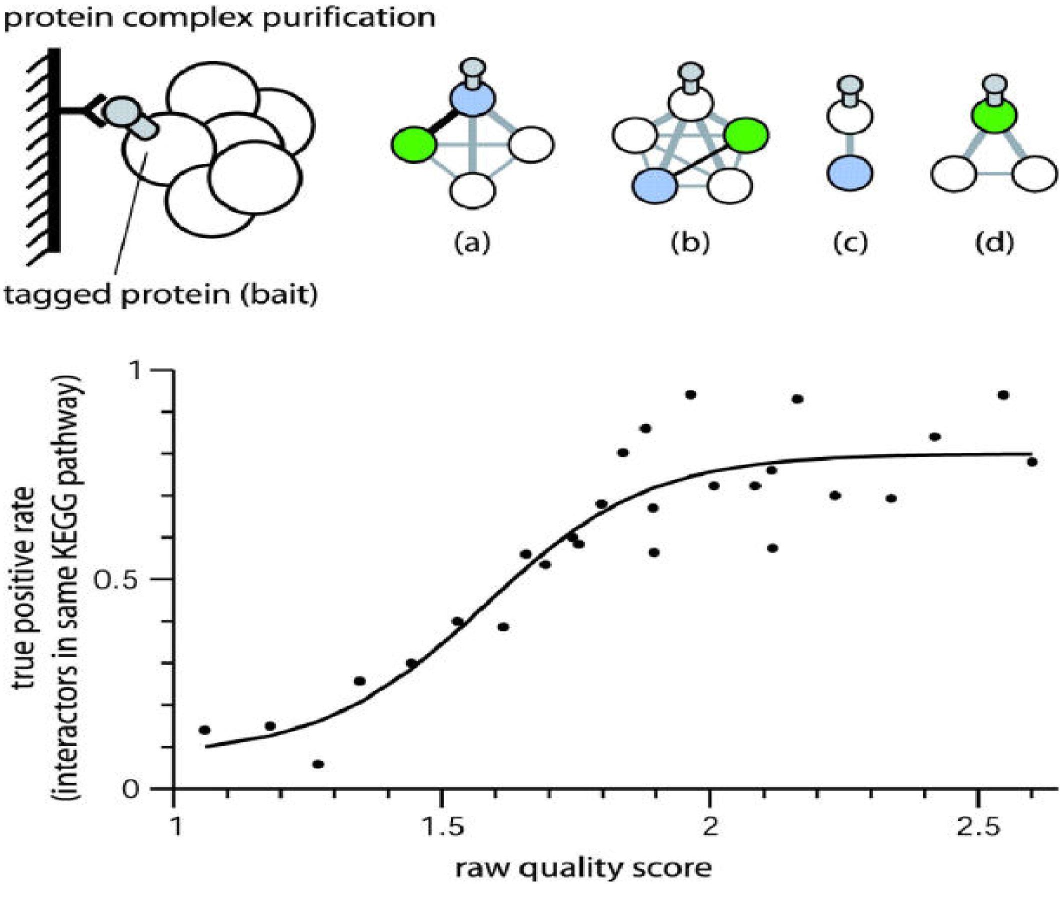

Figure 2.1: Empirical estimates of confidence scores derived from quality scores for HGC algorithm. Generated from equation 2.29 above**.** These results are determined based on the number of interactions (either through direct protein binding as shown in this figure or co occurrence in the same pathway) annotated in databases such as Gene Ontology, STRING and Protein Databank for a particular set of proteins.

This is achieved by using the plot depicting proportion of annotations that are shown to have a pathway interaction in the manually curated KEGG database against the raw quality score, for each database source in the STRING repository (Gene Ontology, Protein Databank and MINT) [33-34]. The pathway interactions in the KEGG database correspond to the probability or confidence score (from 0 to 1) that the binding interactions in the STRING repository contain functional relevance. For example, empirical estimate of the graph above shows that a given set of genes with a Q score (from equation 1) of 2.5, corresponds to a confidence level of approximately 0.8 (or 80% chance) that the protein interactions have mechanistic importance in biological pathways that underpin disease pathogenesis [33-34]. This value is denoted by the variable, Si.

For each value of S (derived from benchmarking GO, Protein Databank and MINT protein interactions against the KEGG database), a combined score of confidence is then calculated (equation 2):

$$\mathrm{S} = 1 - \Pi_{\mathrm{i}}(1 - \mathrm{S}_{\mathrm{i}}) \text{ (2.30)}$$

S_i is the confidence score of interaction for protein annotations for each separate database source (mentioned above) in the STRING repository, for the ith set of proteins. Substituting this value in equation 2 gives the combined or weighted score of confidence for protein interactions [33-34]. This calculation fulfils the assumption that if protein interactions (pathway or binding) are reported in multiple database sources that are contained within the STRING repository, they will have a greater weighted confidence score.

Finally the biological distance between the genes (i and j respectively) of interest is obtained by taking the inverse value of $S_{i,j}$.

$$\mathrm{D}_{\mathrm{i,j}} = \frac{1}{\mathrm{S}_{\mathrm{i,j}}} \text{ (2.31)}$$

Once this process is complete for direct interactions (via a custom Python Script by Itan et al 2013) [31-32], the nodes and edges can be subsequently connected for genes that do not have direct interactions, by combining nodes that interact directly [31-32]. This generates a complete connectome network. If there is more than one node separating the genes of interest, then the summed distance (between the 2 genes) is multiplied by the number of nodes separating them. Hence, genes with non-direct interactions will have a larger, weighted biological distance.

It should be noted that the Human Gene Connectome provides a comparatively more powerful analysis than HumanNet and Funcoup (alternate database repositories) respectively [31-32]. HumanNet is able to identify new genes in a biological pathway, whilst FunCoup is sufficient for identifying only closely related genes (predominantly those with direct interactions). Also they are powerful for analysis of polygenic disease, as in such cases there is no core gene required, when the candidate gene list is submitted for analysis, and networks can be provided that are inferred to be associated with a disease of interest [31-32]. However approaches, using these sources are not gene centric, and therefore provide no information of the biological proximity to the core gene, and the routes between the central and candidate genes. Also Funcoup and HumanNet are only able to provide a qualitative (yes/no) answer to gene interaction and functional relatedness, and therefore it is difficult to distinguish between multiple genes with close functional relatedness to core gene(s) [31-32]. Thus it is clear that the gene centric approach of the HGC as well as its ability to rank genes by biological distance and route between genes of interest makes it adaptable for analysis of monogenic as well as polygenic disease, as it can differentiate between genes with close relatedness to core genes [31-32]. This enables more efficient detection of new candidate disease genes, especially in cases where pathways with a high likelihood of involvement in disease pathogenesis are available.

2.8 Genetic Analysis of Population Isolates:

As further outline in chapter 6, population isolates often exhibit unique patterns of genomic variation that may be influence disease susceptibility. These include linkage disequilibrium and enrichment of rare variation. The mathematical algorithms to calculate their significance and to compare them between the Paisa and selected populations from the 1000 Genomes Project are outlined below.

2.8.1 Linkage Disequilibrium

Linkage disequilibrium (LD) is the occurrence of non-random association of alleles and genotypes at different loci, more than what is expected by chance from the population allelic frequencies [35]. It is a powerful quantitative indicator of factors such as genetic drift, admixture and genomic structural variation within and between given populations. There are 2 main measures of LD. These are D' and r^2. The D' statistic depends upon the difference (D) between the observed frequencies at which alleles of 2 separate loci are segregated together in the analysed population (i.e. $D = P_{ab} - P_aP_b$) [35-36]. Afterwards, the theoretical maximum of D is required, derived from multiplication of the observed allele frequencies, at the respective loci [36-39]. I.e. suppose that P_A and $P_{a,}$ represent the allele frequencies at locus x, whilst q_A and q_b denote those at locus y [36-39]. Therefore, the theoretical maximum of D is determined by the alleles giving the smallest product of multiplied allele frequencies, as per the following formula:

$$D' = \frac{D}{D_{max}} = \frac{D}{\min(P_A q_b \text{or } P_a q_B)} \text{ when } D'_{AB} > 0 \ (2.32)$$

$$D' = \frac{D}{D_{max}} = \frac{D}{\min(P_A q_B \text{ or } P_a q_b)} \text{ when } D'_{AB} < 0 \ (2.33)$$

Thus it is clear from the formula, that the D' statistic is able to detect LD in a non-variant frequency dependent manner, as it is dependent upon the value of D and its theoretical maximum, which can include the product of frequencies from the reference and/or alternate allele [36-39]. This is also the case for rare variants. However, this method is prone to yielding false positives, as estimates can be inflated in small samples or when one variant allele frequency is low. Whilst the D' may still maintain its accuracy with one common and one rare allele at respective loci, the statistic may still be inflated in this instance [36-39]. However, if both alleles are rare, it is difficult to reliably establish LD, unless the sample size is considerably large.

The alternate statistic that is preferred by most geneticists is the r^2 value. The r^2 is obtained by:

$$\mathrm{r}^2 = \frac{\mathrm{D}_{\mathrm{AB}}^2}{\mathrm{PA}(1-\mathrm{PA})\mathrm{PB}(1-\mathrm{PB})} \ (2.34)$$

Unlike the D’, the r^2 value is totally dependent upon the frequency of the variant allele. Thus, as the variant allele frequency decreases, so does the r^2 value [36-39]. Therefore, r^2 is more powerful for detecting LD with common variants. Nevertheless, the likelihood of false positives (which can arise through D’) amongst rare variants can also be reduced [36-39]. Thus, for the purposes of comparing LD between the Paisa and other selected 1000 Genomes populations, the r^2 statistic was used. The next step was to evaluate the significance of LD comparisons. The LD calculations are derived from PLINK software [40].

2.8.2 P-values of Pairwise LD comparisons

When differentiating the relative magnitude in LD between the Paisa and other European plus Latin American cohorts, pairwise analyses of the Wilcoxon Rank Sum Test were conducted [41]. The advantage of this test is that it is not conditional on the assumption of data or observations following a normal distribution. Let m be the number of observations in Cohort 1 and n be the number of observations in cohort 2. The test procedure first involves combining all observations from both cohorts into a single groups and ranking them from smallest to largest. Afterwards, the rank sum (denoted by W1 and W2) of both treatments are obtained [41-44]. The smaller of the two rank sums (assume W1) is selected. This is followed by the permutation status for the values corresponding to given observations. The total number of possible permutations is given by X = $^{m+n}C_{n.}$ However, for very large values, an approximate p-value estimate is derived, as it can be computationally intractable to perform all permutations [41-44]. In our case, the number of permutations conducted in R is such that the most significant estimate of the P-value is < 2.2×10^{-16}. The number of occasions that the permuted rank sum (#W_{perm}) is smaller than the observed value of W1 divided by the number of permutations performed (N) gives the one-tailed p-value [41-44]. Thus:

$$\frac{\#\mathrm{Wperm} \leq \mathrm{W1obs}}{\mathrm{N}} \ (2.35)$$

The p-value therefore gives the probability of obtaining a greater rank sum in one cohort over another by chance. This statistical analysis can be performed through the R programming language, using the ‘coin’ package [45-46].

2.8.3 Quantifying Differences in Rare Variation Proportions: Z-Proportion Test

To contrast the proportion of rare variation across all loci between populations, we then determine the respective percentages of variants in each cohort with <= 5% MAF in the 1000 Genomes database, via custom python scripts [47]. The significance of these differences is quantified by pairwise Z-proportion tests, via the ‘MASS’ package in R [48].

$$\text{Z proportion score} = \frac{p_1 - p_2}{((\sqrt{p_{\text{pooled}}(1 - p_{\text{pooled}})\left(\frac{1}{N_1} + \frac{1}{N_2}\right)})} \quad (2.36)$$

The variables p1 – p2 are the differences between the percentages. N_1 and N_2 are the sample sizes for each cohort, whilst p_{pooled} is the combined proportion of rare variants, added across both cohorts [49]. p_{pooled} is used in the formula, because the calculation is done under the assumption and hypothesis of no difference between population groups 1 and 2.

2.8.4 P-value of Z proportion test

The statistical significance depends upon where the score fits within a normal distribution, as per the null hypothesis [49]. To do this, the area between the tail of the distribution and the critical value on the distribution, best corresponding to the test statistic is obtained [49]. Thus the p-value can be subsequently derived. The greater the difference between proportions, the more likely that a score at least as extreme as the observed statistic is not obtained by chance alone.

References

[1] I.J. Barnett, S. Lee and X. Lin. Detecting Rare Variant Effects Using Extreme Phenotype Sampling in Sequencing Association Studies. *Genet Epidemiol,* 37(2):142-51, 2013.

[2] Bozeman MT: Variant Classification. In: SNP and Variation Suite Manual version 8.7.2, Copyright 2017, accessed Feb 2017.

[3] K. Pearson. On lines and Planes of Closest Fit to Systems of Points in Space. *Philos Mag,* 2(11):559-72, 1901.

[4] A.L. Price, N.J. Patterson, R.M. Plenge, et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet,* 38(8):904-9, 2006.

[5] N. Patterson, A.L. Price, and D. Reich. Population Structure and Eigenanalysis. *PLoS Genet,* 2(12):e190, doi:10.1371/journal.pgen.0020190, 2006.

[6] R. Shen, J.B. Fan, D. Campbell, et al. High Throughput SNP genotyping on universal bead arrays. *Mutat Res,* 573(1-2):70-82, 2005.

[7] S. Bennett. Solexa Ltd. *Pharmacogenomics,* 5(4):433-8, 2004.

[8] S.T. Bennett, C. Barnes, A. Cox, et al. Towards the $1000 Genome. *Pharmacogenomics,* 6(4): 373-82, 2005.

[9] A. McKenna, M. Hanna, E. Banks et al. The Genome Analysis Toolkit: a Map-Reduce framework for analysing next-generation DNA sequencing. *Genome Res,* 20(9):1297-1[30]3, 2010.

[10] M. DePristo, E. Banks, R. Poplin, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet,* 43(5):491-8, 2011.

[11] G.A. Van der Auwera, M. Carneiro, C. Hartl, et al. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices. *Curr Prot Bioinformatics,* 43:11.10.1-33. Doi: 10.1002/0471250953.bi1110s43, 2013.

[12] J.T. Robinson, H. Thorvaldsdóttir, W. Winckler, et al. Integrative Genomics Viewer. *Nat Biotech,* 29(1):24-6, 2011.

[13] H. Thorvaldsdóttir, J.T. Robinson, J.P. Mesirov, et al. Integrative Genomics Viewer: high performance genomics data visualisation and exploration. *Brief Bioinform,* 14(2):178-92, 2013.

[14] B. Kennedy, Z. Kronenberg, H. Hu, et al. Using VAAST to Identify Disease-Associated Variants in Next-Generation Sequencing Data. *Curr Prot Hum Genet*, 81: 6.14.1-25. doi: 10.1002/0471142905.hg0614s81, 2014.

[15] H. Hu, J.C. Roach, H. Coon, et al. A Unified Test of Linkage Analysis and Rare-Variant Association for Pedigree Sequencing Data. *Nat Biotechnol*, 32(7):663-9, 2014.

[16] M. Fishelson and D. Geiger. Exact genetic linkage computations for general pedigrees. *Bioinformatics*, 18(suppl 1):S189-S198, 2002.

[17] E.S. Lander and P. Green. Construction of multilocus genetic linkage maps in humans. *Proc Natl Acad Sci,* 84(8):2363-67, 1987.

[18] T.M. Darlington, R. Pimentel, K. Smith, et al. Identifying rare variants for genetic risk through a combined pedigree and phenotype approach: application to suicide and asthma. *Transl. Psychiatry,* 4:e471, doi: 10.1038/tp.2014.111, 2014.

[19] M. Yandell, C. Huff, H. Hu, et al. A probabilistic disease-gene finder for personal genomes. *Genome Res,* 21(9):1529-42, 2011.

[20] The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature,* 526(7571):68-74, 2015.

[21] S. Henikoff and J. Henikoff. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci,* 89(22):10915-19, 1992.

[22] A. Hamosh, A.F. Scott, J.S. Amberger, et al. Online Mendelian Inheritance in Man, a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res,* 33:D514-D517, 2005.

[23] M. Yandell, B. Moore, F. Salas, et al. Genome-wide analysis of human disease alleles reveals that their locations are correlated in paralogous proteins. *PLoS Comput Biol,* 4:e1000218, doi:journal.pcbi.1000218, 2008.

[25] I.A. Adzhubhei, S. Schmidt, L. Peshkin, et al. A method and server for predicting damaging missense mutations. *Nat Methods,* 7(4):248-9, 2010.

[26] I. Adzhubhei, D.M. Jordan, and S.R. Sunyaev. Predicting functional effect of human missense mutations using PolyPhen 2. *Curr Prot Hum Genet,* Chapter 7:Unit7.20, Jan 2013.

[27] J.M. Schwarz, D.N. Cooper, M. Schuelke, and D. Seelow. MutationTaster2: mutation prediction for the deep sequencing age. *Nat. Methods,* 11(4):361-2, 2014.

[28] Y. Choi, and A.P. Chan. Provean web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics,* 31(16):2745-7, 2015.

[29] Y. Choi, G.E. Sims, S. Murphy, et al. Predicting the Functional Effect of Amino Acid Substitutions and Indels. 7(10):e46688, doi:10.1371/journal.pone.0046688, 2012.

[30] Metacore Version 6.24, Build 67895, Thomson Reuters, New York, USA, 2015.

[31] Y. Itan, S.Y. Zhang, G. Vogt, et al. The human gene connectome as a map of short cuts for morbid allele discovery. *Proc Natl Acad Sci USA,* 110(14):5558-63, 2013.

[32] Y. Itan, M. Mazel, B. Mazel, et al. HGCS: an online tool for prioritizing disease-causing gene variants by biological distance. *BMC Genomics,* 15:256, doi: 10.1186/1471-2164-15-256, 2014.

[33] C. Von Mering, L.J. Jensen, B. Snel, et al. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res,* 33(1):D433-37, 2005.

[34] D. Szklarkzyk, A. Francheschini, M. Kuhn, et al. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res,* 39(Database Issue):D561-8, 2011.

[35] M. Slatkin. Linkage disequilibrium – understanding the evolutionary past and mapping the medical future. *Nat Rev Genet,* 9(6):477-85, 2008.

[36] J.M. VanLiere and N.A. Rosenberg. Mathematical properties of the r^2 measure of linkage disequilibrium. *Theor Popul Biol,* 74(1):130-37, 2008.

[37] R.B. Robbins. Some applications of mathematics to breeding problems. *Genetics,* 3(4):375-89, 1918.

[38] R.C. Lewontin, and K. Kojima. The evolutionary dynamics of complex polymorphisms. *Evolution*, 14(4):458-72, 1960.

[39] R.C. Lewontin. Some applications of mathematics to breeding problems. The interaction of selection and linkage. 1. General considerations; Heterotic Models.

[40] S. Purcell, B. Neale, K. Todd-Brown et al. PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am. J. Hum. Genet,* 81(3):559-75, 2007.

[41] H.B. Mann, and D.R. Whitney. On a Test of Whether one of Two Random Variables is Stochastically Larger. *Ann. Math. Stat*, 18(1):50-60, 1947.

[42] W.J. Conover and R.L. Iman. Rank Transformations as a Bridge Between Parametric and Nonparametric Statistics. *Am. Stat*, 35(3):124-29, 1981.

[43] D.F. Bauer. Constructing confidence sets using rank statistics. *JASA,* 67(339):687-90, 1972.

[44] M.A. Hollander, and D.A. Wolfe. *Nonparametric Statistical Methods*: 2nd Ed. New York: John Wiley & Sons, 1999.

[45] T. Hothorn, K. Hornik, M.A. van de Wiel and A. Zeileis. A Lego System for Conditional Inference. *Am Stat*, 60(3):257-63, 2006.

[46] T. Hothorn, K. Hornik, M.A. van de Wiel and A. Zeileis. Implementing a Class of Permutation Tests: The coin Package. *J Stat Soft*, 28(8):1-23, 2008.

[47] Python Software Foundation Python Language Reference, Version 2.7. Available at http://www.python.org, Accessed: Jun 2017.

[48] W.N. Venables, and B.D. Ripley. *Modern Applied Statistics with S*. Fourth Edition. New York: Springer, ISBN 0-387-95457-0, 2002.

[49] R.C. Sprinthall. *Basic Statistical Analysis*: 9th Ed. Pearson Education, 2011

Chapter 3: Novel and Rare Functional Genomic Variants in Multiple Autoimmune Syndrome and Sjögren's Syndrome

Abstract:

Multiple autoimmune syndrome (MAS), an extreme phenotype of autoimmune disorders, is a very well suited trait to tackle genomic variants of these conditions. Whole exome sequencing (WES) is a widely used strategy for detection of protein coding and splicing variants associated with inherited diseases. The DNA of eight patients affected by MAS [all of whom presenting with Sjögren's syndrome (SS)], four patients affected by SS alone and 38 unaffected individuals, were subject to WES. Filters to identify novel and rare functional (pathogenic-deleterious) homozygous and/or compound heterozygous variants in these MAS patients were applied. Bioinformatics tools such as the Human Gene Connectome as well as pathway and network analysis were applied to test overrepresentation of genes harbouring these variants in critical pathways and networks involved in autoimmunity. Eleven novel and rare functional variants were identified in cases but not in controls, harboured in: *MACF1*, *KIAA0754*, *DUSP12*, *ICA1*, *CELA1*, *LRP1/STAT6*, *GRIN3B*, *ANKLE1*, *TMEM161A*, and *FKRP*. Notably, the *LRP1/STAT6* novel mutation was homozygous in one MAS affected patient and heterozygous in another. *LRP1/STAT6* disclosed the strongest plausibility for autoimmunity. Network analysis indicates *LRP1/STAT6* are involved in extracellular and intracellular anti-inflammatory pathways that play key roles in maintaining the homeostasis of the immune system. Further; networks, pathways, and interaction analyses showed that *LRP1* is functionally related to the *HLA-B* and *IL10* genes and it has a substantial impact within immunological pathways and/or reaction to bacterial and other foreign proteins (phagocytosis, regulation of phospholipase A2 activity, negative regulation of apoptosis and response to lipopolysaccharides). Amongst all the novel and rare variants identified, the *LRP1/STAT6* novel mutation has the strongest case for being categorised as potentially causative of MAS given the presence of intriguing patterns of functional interaction with other major genes shaping autoimmunity.

3.1 Introduction

Recent evidence supports the involvement of rare variants (population allele frequency < 1%) in the aetiology of common diseases [1],[2]. 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. The detection of these rare genomic variants harboured in coding regions has shown to be achievable using extreme phenotypes (those exhibiting an unexpected and extreme accumulation of signs and/or symptoms than those expected by the disease's natural history) and pedigrees segregating exceptional phenotypes [1],[2].

Polyautoimmunity is defined as the presence of more than one autoimmune disease (AD) in a single patient [3]. When three or more ADs coexist, the condition is called multiple autoimmune syndrome (MAS), which characterises the best example of polyautoimmunity, and probably the most conspicuous extreme autoimmune phenotype [4] i.e., ***(i)*** MAS amalgamates signs and symptoms that are present in several ADs, ***(ii)*** the MAS signs and symptoms clustering is not random but it outlines the presence of subtypes, ***(iii)*** in many occurrences, it clusters in families, and ***(iv)*** major gene effects and the potential location of these MAS major loci have been established [4],[5]. Consequently, it is fair to consider that MAS, as an extreme phenotype of autoimmunity, would be critical for dissecting genes of major effect conferring susceptibility to autoimmunity [5],[6]. Sjogren´s syndrome (SS), an autoimmune exocrinopathy, is frequently observed in MAS patients [7].

Whole exome sequencing (WES) is a cost effective technique, becoming the first-line approach for monogenic disorders, and an alternative one for dissecting extreme phenotypes of complex inherited conditions [5]. WES is a highly effective approach for identifying homozygous, compound heterozygous, novel, germinal, and *de novo* rare coding variants [5]. Its ultimate rationale remains in that genetic variants located in exons are more likely to be pathogenic, with major effect than many of those located in introns or between genes. The power of this strategy has increased with available access to large numbers of publicly available exome sequence databases that allow the controlled comparison of frequencies, as well as the identification of *de novo* variants and stratification by ethnicity. In this manuscript we report the identification of rare and novel variants observed in sporadic MAS and SS patients.

3.2 Methods

3.2.1 Patients and controls

Eight patients with MAS and 4 patients with SS alone, fulfilling validated classification criteria as previously reported [3],[4],[8], were included. Patients were assessed at the Center for Autoimmune Diseases Research (CREA), at the University of Rosario, in Bogotá and Medellin, Colombia (Table 3.1). Patients and controls did not present other phenotypes such as cardiovascular disease (i.e., ischemic heart disease or stroke), or diabetes.

Sample	Phenotype	Gender
1	MAS (SS, AITD, VIT)	F
2	MAS (SS, SSc, AIH, AITD)	F
3	MAS (RA, SS, AITD,)	F
4	MAS (PSO, RA, SS)	F
5	MAS (AITD, RA, SS)	F
6	MAS (AITD, RA, SS)	F
7	MAS (SLE, SS, AITD)	F
8	MAS (RA, SLE, SS)	F
9	SS	F
10	SS	F
11	SS	F
12	SS	F

Table 3.1: Phenotypic Information for individuals carrying a MAS or Sjögren's phenotype, amongst sporadic cases of autoimmunity. Abbreviations: AITD: Autoimmune Thyroid Disease; SLE: Systemic Lupus Erythematous; SS: Sjögren's syndrome; SSc: systemic sclerosis; RA: Rheumatoid Arthritis; VIT: Vitiligo; PSO: Psoriasis; AIH: Autoimmune Hepatitis.

3.2.2 DNA library preparation, exome capture and sequencing protocol

Libraries were constructed from 1 μg of genomic DNA using an Illumina TruSeq genomic DNA library kit at the Biomolecular Resource Facility, John Curtin School of Medical Research. Libraries were multiplexed with 6 samples pooled together (500 ng of each library). Exons were enriched from the pooled 3 μg of library DNA using the Nimblegen Exome enrichment kit. Each exome-enriched pool was run on a 100-base-pair paired and run on an Illumina HiSeq 2000 sequencer.

3.2.3 Sequence read processing, alignment, variant calling and analyses by filtering

The sequencing image data was converted to FASTQ files containing DNA base calls (A, C, G and T) and quality scores using the Illumina CASAVA pipeline in order to convert raw image data into sequences. The resulting FASTQ files were further processed for variant analysis.

The workflow for data curation and analysis for variant calling was developed by the Genome Discovery Unit (GDU), at the Australian National University (ANU). Key components of the workflow include: i) Quality assessment; ii) Read alignment; iii) Local realignment around the known and novel insertions/deletions (indel) regions to refine indel boundaries; iv) Recalibration of base qualities; v) Variant calling using the Genome Alignment Tool Kit (GATK) algorithm; and vi) Assigning quality scores to variants [9-13].

Subsequently, we included a filtering phase (using information from dbSNP and the 1K Exome Project), with the following sequential steps: 1) identification of *novel* variants i.e. those variants absent from the 1000 genomes and dbSNP databases (the 1000 genomes - phase3 - has a set of 95 individuals recruited from Colombia; the same area of ascertainment of these sporadic cases); 2) filtering of variants to include either pathogenic or specific variants associated to disease with numerous tools i.e., SIFT, PolyPhen2, Mutation Taster, Mutation Assessor, and FATHMM (more detailed information in the Additional file 1) as implemented in the DNA-seq Analysis Package SVS7.7.6, Golden Helix, Bozeman, USA [9]. Variants were not excluded if classified as potentially damaging by at least one of these filtering tools. These variants are not necessarily non-synonymous, but can also include those found in splice sites or that are a part of splicing regulatory elements, as identified by the Variant Classification and Human Splice Finder algorithms respectively [14]-[18] and 3) Filtering of damaging variants based on genes known to be associated with human disease. The identification of likely compound heterozygous polymorphisms, and rare recessive homozygous polymorphisms, was performed with different modules of the DNA-seq Analysis Package in SVS7.7.6, Golden Helix, Bozeman, USA [14], in combination with custom Python scripts. For any homozygous intronic variants identified (in cases only) during the initial filtration process, further analysis was conducted using algorithms of the Human Splice Finder [17], in order to identify possible motifs harbouring mutations that might have an effect on splicing regulation (Splicing Regulatory Elements or SREs). In brief, Position Weight Matrices are constructed for the predicted sequence motifs, in order to measure the level of nucleotide sequence conservation, as well as their enrichment in introns vs. exons [17]. Sequences that have more enriched matrix scores in a given intronic region compared to other locations in the gene's exons and introns are considered as candidate splicing regulators [17]. Thus our approach is attempting to extract as much information as possible from non-synonymous and splicing variants as well as other non-coding variants proximal to exon boundaries, in order to reduce the risk of excluding genes that may have substantial importance in the phenotypes of these autoimmune patients.

3.2.4 Principal Component Analysis

Population stratification and substructure can generate spurious association and consequently inaccurate conclusions about the enrichment of candidate variants in cases over controls. Although our dataset contains exome-sequencing variants from individuals who are from a homogeneous region, small levels of microdifferentiation may be present. We control this potential confounder by applying genotype based Principal Component Analysis (PCA), as implemented in SVS 7.7.6, Golden Helix, Bozeman, USA [14],[23], to identify outliers. PCA is applied to both filtered and unfiltered datasets.

3.2.5 Network analysis

To identify potential enriched MAS related physiological pathways, network analyses were performed. For constructing networks and pathways, variants with potential functional changes, detected as novel and in homozygote state, were examined with the 'Analyse Network', 'Process

Networks', 'Shortest Paths' and 'Direct Interactions' algorithms implemented within the MetaCore software suite (Version 6.2, Build 66481, Thomson Reuters, New York, USA) (Details regarding some the differences between the algorithms can be found in the MetaCore Manual). These procedures allowed us to use a heuristic integration of maps and networks and rich ontologies for diseases based on the biological role of candidate genes.

3.2.6 The Human Gene Connectome (HGC) Analysis

Similar to MetaCore, the rationale of implementing the HGC is also for prioritizing candidate genes on the basis of their functional relevance to autoimmune phenotypes. In this case however, candidate genes were chosen on the basis of their quantitative relatedness or biological distance to genes already established as having functional importance in ADs. This was used to calculate biological distances between candidate genes identified from the aforementioned filtering strategies and previously identified genes with potential functional relevance in autoimmunity, including but not limited to rheumatoid arthritis (RA), SS, systemic lupus erythematous (SLE) and autoimmune thyroid disease (AITD) [19]-[21]. The genes with known functional/physiological relevance and/or association to autoimmunity were obtained from the Gene Prospector database [21]. Genes within the top 10% listed for each disease and shared amongst multiple ADs of interest (present in the MAS patients) were selected for the HGC analysis of candidate genes [21]. To evaluate the significance of these distances, P-values were estimated via random permutation of pairwise gene interactions in the HGC database. These values were subsequently corrected using the Benjamini and Hochsberg false discovery rate (FDR) method [22].

3.3 Results

3.3.1 Evaluating for potential population structure

After applying PCA, there was no evidence of stratification effect between cases and controls. No outliers from both groups were identifiable, across the eigenvectors and eigenvalues.

3.3.2 Candidate genetic variants identified from filtering strategies

The filtering strategies were essential tools in order to successfully obtain a refined, prioritized list of candidate genes, with potential importance in the MAS patients. Using the aforesaid approach, we successfully identified 11 variants within the following genes: *DUSP12, GRIN3B, MACF1, KIAA0754, LRP1, STAT6, ANKLE1, BABAM1, TMEM161A, MICAL1, ICA1, FKRP* and *CELA1*. As shown in table 3.2 Ten out of the 12 affected individuals had at least 1 homozygous or a pair of compound heterozygous variants within genes, which were not observed in the controls.

By definition, mutations in the *CELA1* and *TMEM161A* genes were considered as splice mutations. This is because these variants are located within a *GT-AG* nucleotide portion of the intron along the DNA sequence that encodes the messenger RNA, which is evident after implementation of the Integrative Genomics Viewer (IGV) [16],[24],[25]. It has been previously observed, that variants in these regions outside the exon boundary are well conserved in splice sites [16]. In addition, two heterozygous variants within the *MACF1* gene were present in one of the four patients with SS. With the exception of the rare variants harboured in *KIAA0754* and *CELA1,* each of these variants was absent from both: the dbSNP and the 1K databases. All of them had potentially deleterious effects according to at least one of the following variant effect predictors: Polyphen, SIFT, FATHHM, MutationTaster and Mutation Assessor (Table 3.2).

Of particular importance was the *DUSP12* gene; harbouring one homozygous novel mutation in a MAS patient affected by AITD, RA and SS, as well as a second heterozygous novel mutation in another MAS individual (diagnosed with Psoriasis, RA and SS). The *LRP1* gene has one novel, non-synonymous mutation variants in 2 individuals. Both were affected with MAS. The heterozygote individual had AITD, SS and vitiligo, whilst the homozygote individual was diagnosed with AITD, RA and SS (Table 3.2 and 3.3). Interestingly, apart from non-synonymous and predicted splicing variants, 2 intronic single nucleotide polymorphisms (SNPs) within *MICAL1* and *ICA1* respectively were also identified as part of our list of candidate autoimmune causing mutations. Both of these variants are considered as 'Disease Causing' by the MutationTaster algorithm [26],[27]. After implementing IGV, it was found that the *ICA1* homozygous variant was located in an adenosine rich region, proximal to the 3'UTR of the mRNA sequence in one of the gene's exons [24],[25]. In the case of *MICAL1*, the homozygous variant is 22bp from the intron-exon boundary. In addition, it was also found that this variant is harboured within the vicinity of (i.e. <10bp from) a sequence region containing 2 hexamer non-coding elements (also named as intron identity elements or IIE), which may act as splicing motifs, according to the algorithms implemented in Human Splicing Finder [17]. These motifs contain the following sequences: *ATGGTG* and *TGGTGG* [17-18].

Chr	Position	Type of Mutation	Gene	Exon	HGVS Protein
1	39,854,131	Nonsyn	MACF1	52	p.Asn3144Thr
1	39,879,412	Nonsyn	KIAA0754/MACF1	1	p.Ala1159Thr
6	109,767,639	Intronic/potential regulatory	MICAL1		
1	161,719,833	Nonsyn	DUSP12	1	p.Pro81Arg
7	8,196,577	Intronic/potentially regulatory	ICA1		
12	51,740,405	Splicing	CELA1	1	?
12	57,522,754	Nonsyn	LRP1/STAT6	1	p.Thr3Pro
19	1,009,552	Nonsyn	GRIN3B	9	p.Ala1028Gly
19	17,392,775	Nonsyn	BABAM1/ANKLE1	1	p.Arg70Trp
19	19,245,591	Splicing	TMEM161A	2	?
19	47,259,734	Nonsyn	FKRP	4	p.Glu343Gln

Table 3.2: Candidate genetic variants identified amongst individuals carrying autoimmunity, which are absent from controls. The gene name and nucleotide position, amino acid change (HGVS Protein: for non-synonymous variants) along with transcript ID at a particular variant is listed. The reference and alternate alleles are also given, along with identifiers to distinguish novel variants (i.e. absent from 1000 genomes and dbSNP databases).

Variant	Individual ID and Phenotype	Genotype
1:39,854,131	9 (SS)	(AC)
1:39,879,412	9 (SS)	(AG)
1:161,719,833	5 (AITD, RA, SS)	(GG)
	4 (PSO, RA, SS)	(CG)
6:109,767,639	6 (AITD, RA, SS)	(CC)
7:8,196,577	12 (SS)	(TT)
12:51,740,405	6 (AITD, RA, SS)	(GG)
12:57,522,754	1 (AITD, SS, VIT)	(AC)
	3 (AITD, RA, SS)	(CC)
19:1,009,552	4 (PSO, RA, SS)	(GG)
19:17,392,775	11 (SS)	(TT)
19:19,245,591	11 (SS)	(CC)
	2 (SS, SSc, AIH)	(AC)
19:47,259,734	11 (SS)	(CC)
	7 (SLE, SS, AITD)	(CC)

Table 3.3: Phenotypes and genotypes of individuals carrying potentially genetic deleterious variants in autoimmunity, absent from controls. The chromosome and nucleotide position of the variant harboured within the candidate gene is given with the corresponding individuals, their phenotypes and the genotypes.

3.3.3 Quality evaluation of sequence reads

The information about mapping quality, which measures the confidence that a sequence read, corresponds to its aligned position in the genome, based on the strength of the alignment, and the Base Phred Score (a quantitative estimate of the probability of an incorrect base call), reported a high quality of reads. In our case homozygous variants harboured in the *DUSP12*, *ICA1* and *LRP1/STAT6* genes had a mapping quality of 42, greater than any other variant. In addition, the Phred Quality Scores for each of these genes was 34, 29 and 27 respectively (table 3.4). This shows that for these variants, the probability of correct mapping during the alignment of these reads harbouring the variants is greater than 99.99%. Also, the likelihood of accurate base calls at each of these nucleotide positions is more than 99.8%.

Gene	Variant	Individual (Phenotype)	Mapping Quality	Phred Score	Variant Effect Predictors that considered each variant as potentially deleterious
DUSP12	1:161,719,833	5 (AITD, RA, SS) 4 (PSO, RA, SS)	42	34	FATHMM (damaging)
ICA1	7:8,196,577	12 (SS)	42	28	Mutation Taster (Disease Causing)
MACF1	1:39,854,131	9 (SS)	42	32	PolyPhen2 (possibly damaging), Mutation Taster (Disease Causing)
MACF1/KIAA0754	1:39,879,412	9 (SS)	42	33	Mutation Taster (Disease Causing)
CELA	12:51,740,405	6 (AITD, RA, SS)	23	27	Mutation Taster (Disease Causing)
LRP1/STAT6	12:57,522,754	1 (SS, AITD, VIT) 3 (AITD, RA, SS)	40	27	FATHMM (Damaging)
GRIN3B	19:1,009,552	5 (AITD, RA, SS)	31	35	Mutation Taster (Disease Causing)
BABAM1/ANKLE1	19:17,392,775	11 (SS)	42	30	SIFT (damaging)
MICAL1	6:109,767,639	6 (AITD, RA, SS)	40	26	Mutation Taster (Disease Causing)
TMEM161A	19:19,245,591	4 (PSO, RA, SS)	42	33	Mutation Taster (Disease Causing)
FKRP	19:47,259,734	11 (SS) 7 (SLE, SS, AITD)	40 42	34 34	PolyPhen, FATHMM, Mutation Taster (Probably Damaging, Disease Causing, Damaging)

Table 3.4 Measures of quality scores for sequence reads containing potentially deleterious variants in individuals with autoimmunity. Apart from the position of the genetic variants of interest and phenotype information for each individual, the mapping quality and Phred scores for the sequence reads containing the variants are also present. Finally, we also have the list of specific variant effect predictors which calculated each variant as being potentially deleterious.

3.3.4 Pathway and Network Analysis

Significant results from the ‘Analyze Network’ algorithm show that the alpha 2-macroglobulin receptor/low density lipoprotein receptor-related protein (alpha 2 MR/LRP/A2M receptor, a large cell-surface glycoprotein (encoded by the *LRP1* gene) is phosphorylated by Protein kinase C- *alpha* (PKC alpha) during the following processes: phagocytosis, negative regulation of apoptosis and phospholipase A2 activity. According to MetaCore, these processes are also seemingly activated when Plasminogen Activator Urokinase Receptor (PLAUR) binds to the A2M receptor (Figure 3.1 and Table 3.5). In addition, the application of the ‘shortest paths’ network algorithm, found that interferon (IFN) gamma interacts with the A2M receptor by regulating its transcription, which is important in response to lipopolysaccharides (Figure 3.2 and Table 3.5). *MICAL1* is another intriguing gene identified through network analyses (Figure 3.3). Like *LRP1, MICAL1* is also involved in apoptosis regulation, actin filament depolymerisation, and negative regulation of cysteine type endopeptidase activity .

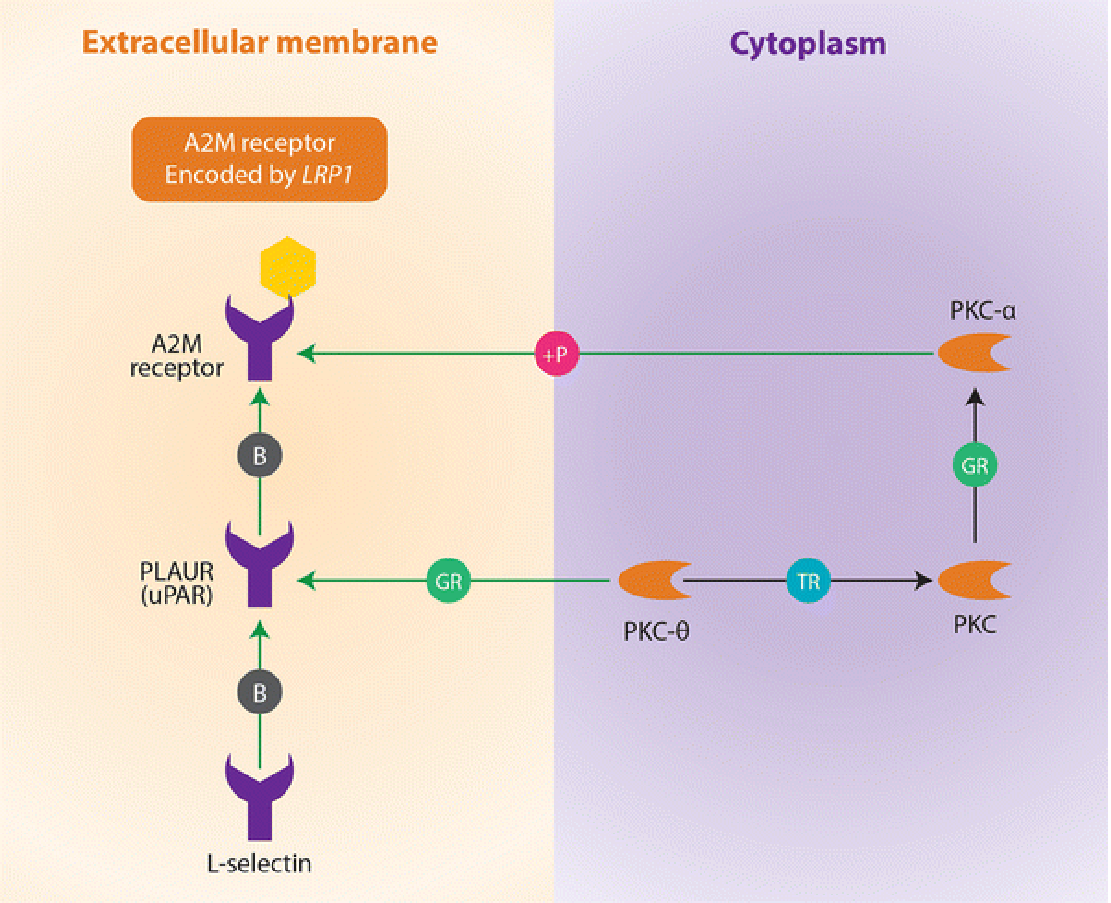

Figure 3.1 Network analysis of candidate genes involving *LRP1* and its potential role in autoimmunity. The network is showing the mechanisms by which protein kinase molecules activate the A2M receptor encoded by the *LRP1* gene. The protein highlighted with a *hexagonal yellow dot* is formed from one of the genes that were identified from the preliminary filtration strategies and used as an input list for the network-building algorithm (in this case gene was *LRP1*). The cellular locations (i.e., cytoplasm and extracellular membrane) of the interacting molecules, which in this case include protein kinases and the A2M receptor is given. Also included are the mechanisms by which one molecule interacts with another. *P* phosphorylation, *B* binding, *GR* group relation, *TR* transcriptional regulation. The effect of these mechanisms is denoted in the colour of the symbols corresponding to the respective nodes is as follows: *pink* activation (by phosphorylation), *grey* activation (by binding), *blue* activation (by transcriptional regulation), *green* unspecified effect due to group relation.

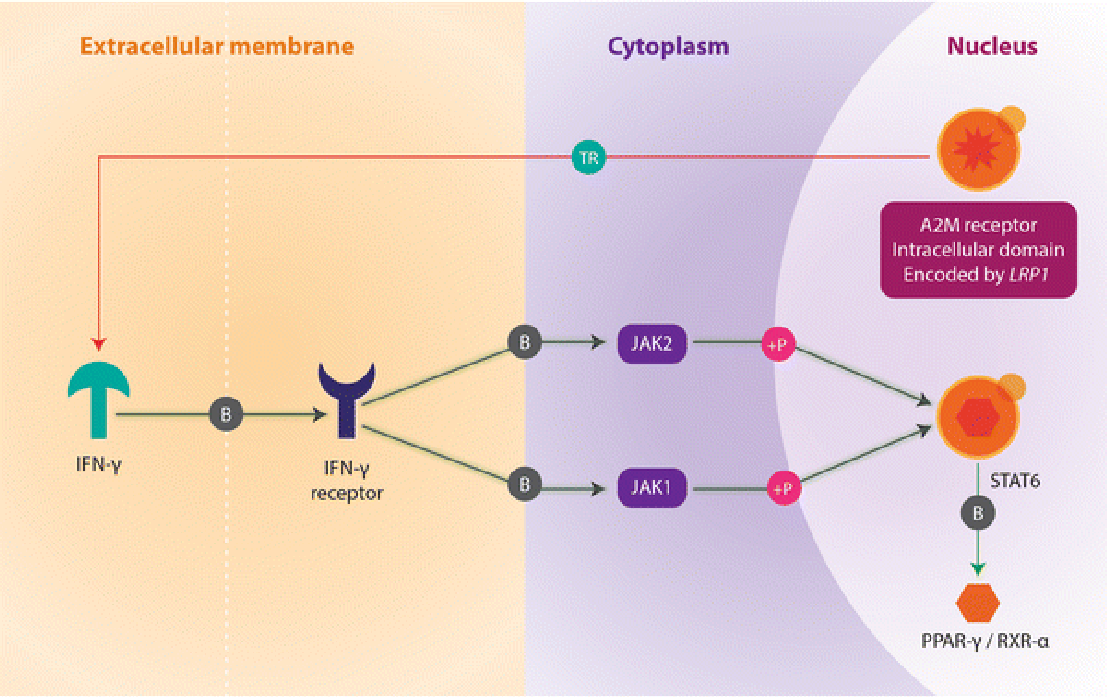

Figure 3.2 Network analysis of candidate genes involving the A2M receptor intracellular domain. In this network, the effect of the A2M receptor (encoded by *LRP1*) intracellular domain upon IFN-gamma is illustrated. Locations of relevant proteins in this network are shown in the nucleus, cytoplasm and extracellular membrane respectively. The mechanistic nature of the protein interactions in the network are as follows: *TR* transcriptional regulation, *B* binding, *P* phosphorylation. The downstream effects exhibited by the protein–protein interactions between a given set of nodes are represented by the following colours on each of the mechanism symbols: *green* inhibition (by transcriptional regulation), *grey* activation (by binding), *pink* activation (by phosphorylation). The A2M receptor and the STAT6 transcription factor are highlighted with a *yellow dot*, showing that they are part of the candidate gene list used as an input source for the network-building algorithm implemented to generate this biological network.

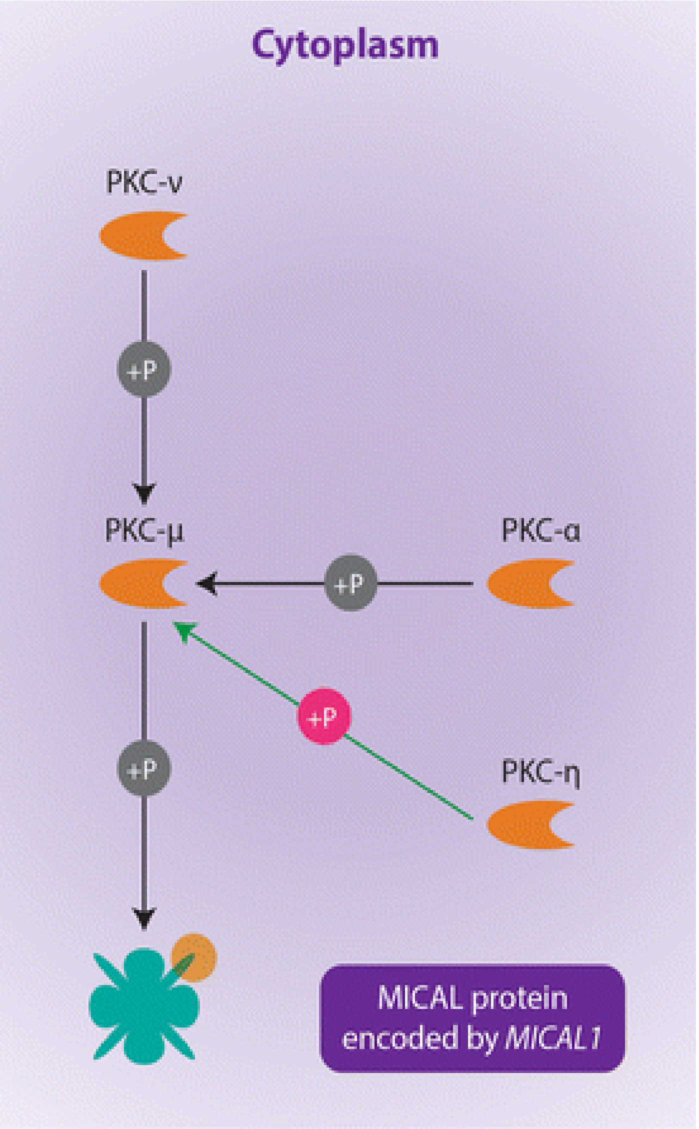

Figure 3.3 Network analysis illustrating the function of *MICAL1* in autoimmune related processes. Of particular importance in the network is the interaction between protein kinase C mu and MICAL. The MICAL protein is highlighted with a *circular yellow dot* (as was the case for the A2M receptor in Figures 1 and 2) because it is encoded by the *MICAL1*that was part of the user generated input list for the MetaCore network-building algorithm. The mechanistic nature of the protein interactions in the network are as follows: *P* phosphorylation. The downstream effects exhibited by protein–protein interactions between a given set of nodes are represented by the following: *pink* activation by phosphorylation, *grey* phosphorylation with unspecified effect.

Gene	**Network Algorithm** **(P-value)**	**Network Node**	**GeneGO ontology Process**	**Processes P-Value**
LRP1	Analyse Network (3.03e-7)	PKC alpha (Phosphorylation of the A2M Receptor encoded by LRP1) **(Figure 1.)**	Phagocytosis	7.596e-8
LRP1	Analyse Network (3.03e-7)	PKC-alpha (Phosphorylation of the A2M Receptor encoded by LRP1) **(Figure1.)**	Regulation of Phospholipase A2 Activity	3.597e-13
LRP1	Analyse Network (3.03e-7)	PKC-alpha (Phosphorylation of the A2M Receptor encoded by LRP1) **(Figure 1.)**	Negative Regulation of Apoptosis	6.703e-21
LRP1	Shortest Paths (N/A)	LRP1 (Transcription Regulation) IFN-gamma **(Figure 2.)**	Response to Lipopolysaccharide	7.616e-21
MICAL1	Analyse Network (7.32e-10)	PKC-mu MICAL1 **(Figure 3.)**	Negative Regulation of apoptotic process	7.901e-15
MICAL1	Analyse Network (7.32e-10)	PKC-mu MICAL1 **(Figure 3.)**	Actin Filament Depolymerisation	2.34e-2
MICAL1	Analyse Network (7.32e-10)	PKC-mu MICAL1 **(Figure 3)**	Negative regulation of cysteine type endopeptidase activity	5.403e-3

Table 3.5: Network and pathway analysis showing the most likely candidate genes with functional relevance in autoimmunity. The first P-Value is of the constructed network. This gives the probability of obtaining a certain number of genes obtained from a given network algorithm from the input list by random chance. Also given are the network nodes and their corresponding biological processes that may have functional importance in ADs.

3.3.5 Human Gene Connectome output

The *LRP1* gene has very short biological distances from the *HLA-B*, *MBL2* and *IL10* genes, as is the case with the distance between *STAT6* and *IRF5* (table 3.6). The functional relatedness of *LRP1* with *HLA-B* and *IL10* is closer than most pairwise comparisons of core and candidate genes, used in this analysis. After FDR correction, the probability of obtaining shorter distances after random permutation and sampling of pairwise distance measurements for *STAT6* and *LRP1* against the remaining genes in the HGC database was less than 0.05 in all cases. The FDR corrected P-values for these distances involving *LRP1* were 0.02486, 0.04428 and 0.04938 respectively. The distance measurement for *STAT6* and *IRF5* yielded an adjusted P value of 0.0388 (see table 3.6). It must also be noted that *STAT6* and *ICA1* have already been identified as genes with established functional importance in SLE and SS respectively within the GeneProspector database [16]. *MICAL1* is another gene, which had close relatedness to important immune system genes such as *PTPN22* and *TLR9*, whilst *DUSP12* had a significantly short distance to *TSHR*. However, these distances only had nominal significance. Even though the variant in *MICAL1* is not within a coding region or splice site, it is still considered functionally relevant, according to the variant effect predictor and biological network analyses.

Core Gene	Candidate Gene	Distance	P-Value	FDR
MBL2	LRP1	4.329	0.00113	0.02486
HLA-B	LRP1	4.20532	0.00552	0.04928
IL10	LRP1	5.38512	0.00672	0.04928
IRF5	STAT6	4.72812	0.01408	0.03888
AIRE	ICA1	1.25	0.00014	0.028

Table 3.6: Biological distances and between core and candidate autoimmune genes. Core genes represent those that are associated or found to have strong functional and mechanistic evidence with ADs in previous studies. The candidate genes listed are part of those that were identified from our studies of the 12 sporadic cases from all of the analyses we conducted before applying the HGC algorithm. For each of the distance values, the significance levels before (P-value) and after (FDR) multiple testing are given.

3.4 Discussion

As a whole, our strategy has been successful in identifying candidate genetic variants that may account for the MAS phenotype present in a subset of affected patients as well as in patients with SS. One factor that must be acknowledged in this approach is the identification of compound heterozygotes for the *MACF1* gene. Given that this gene spans 92 exons and more than 402Kb [28], this increases the likelihood of identifying more than 1 heterozygote in a particular individual by chance (compared to smaller genes), regardless of whether these variants are causative or not. On this basis, one can argue that such genes should be excluded, but at the same time, size alone cannot rule out the fact that these variants may be potentially causative. In this case, these variants were included, as part of our analysis. However their inclusion or exclusion does not change our conclusions about which genes are the best candidates for observed MAS phenotypes.

Other studies involving correlated meta-analyses and factor analysis for inflammatory markers and metabolic traits have suggested that *MACF1* and *KIAA0754* contained significant pleiotropic association with high density lipoprotein cholesterol and C-reactive protein levels. Consequently, this renders these genes as risk factors for metabolic syndrome, which may result in a genetic predisposition for cardiovascular disease and diabetes [29]. Although no patients from our study were diagnosed with either condition those carrying the *KIAA0754* and *MACF1* may have an increased susceptibility to these disorders [29].

Based on these comprehensive analyses we also found intriguing evidence that *LRP1* and *STAT6* have the strongest case for being categorised as potentially causative genes of MAS. This observation came along with: ***(i)*** the ascertainment of patients with extreme autoimmune phenotypes; ***(ii)*** the recruitment from a population exhibiting features of a well-established homogeneous population; ***(iii)*** the identification by whole exome capture and sequencing of novel (*i.e.* not present in dbSNP or 1000 genomes projects) and rare functional coding variants (some of them in at least two patients); and ***(iv)*** the presence of intriguing patterns of functional correlations among them, or with other major genes shaping autoimmunity.

Undoubtedly, the association of *LRP1* with the phenotypes of two MAS patients constitutes an interesting finding that validates the results of the present study. Indeed, several lines of evidence suggest that LRP1 product is involved in crucial extracellular and intracellular anti-inflammatory pathways that play key roles in maintaining the homeostasis of the immune system [30]-[34]. Therefore, a damaging mutation in this gene might largely contribute to the occurrence of MAS.

LRP1 is largely expressed in phagocytic cells such as peripheral macrophages and brain microglia that play crucial roles in engulfing cellular debris such as apoptotic cell bodies, amyloid ß peptide and chromatin [32],[33],[35]. Remarkably, it has been previously described that reduced clearance of dying cells by macrophages causes accumulation of cellular fragments in several tissues [36-37]. This process appears to induce dendritic cells (DC), professional antigen presenting cells that activate naïve T-cells, to uptake apoptotic debris [36]. After that, DCs might present self-antigens to naïve T cells and activate autoreactive T cells [36]. Thus, impaired *LRP1* action could ultimately cause autoimmunity.

The crucial anti-inflammatory role of *LRP1* in counteracting deleterious effects of neurodegenerative diseases has been previously reported [30-33]. For instance, decreased expression of *LRP1* has been hypothesized to be crucial in the extracellular accumulation of beta amyloid protein occurring during Alzheimer's disease [30]. Furthermore, LRP1 has also been hypothesized to play a crucial role in clearing apoptotic cells during multiple sclerosis [33]. In summary, the involvement of LRP1 in the removal of cellular debris might constitute a key step in preventing autoimmunity.

There are other lines of evidence suggesting that LRP1 has anti-inflammatory roles, which indirectly could also aid in the prevention of autoimmunity. First, one of the key LRP1 ligands, alpha-2-macroglobulin (A2MG), enhances survival during sepsis through a novel mode of interaction between cells that involve plasma membrane-shed vesicles containing large proteins and lipid mediators [34]. These vesicles are termed microparticles and one of their key components to prevent sepsis is A2MG, which acts through LRP1 [35]. Secondly, increased levels of glucocorticoids occur during inflammatory challenges aimed at self-containing the inflammatory cascade also increase the expression of *LRP1* in phagocytic cells such as macrophages, which contribute to the removal of apoptotic cells as described above [32]. Thirdly, there is an intracellular self-limiting anti-inflammatory process that involves *LRP1* [31]. Recent *in-vitro* studies described that proteolytic processing of the intracellular domain of the protein encoded by *LRP1* triggered nuclear signalling to dampen the expression of key inflammatory lipopolysaccharide (LPS)-induced genes such as IFN regulatory factor-3 (*IRF-3)* [34]. More specifically, it was shown that the soluble intracellular domain encoded by *LRP1* translocates to the nucleus to repress the LPS-induced increase of *IRF-3*, a crucial transcription factor that regulates the expression of other inflammatory genes.

The HGC and the MetaCore analyses provided additional evidence for *LRP1* and *STAT6* as potentially causal genes within these particular individuals from a functional perspective. As mentioned earlier, *IL10* is related to *LPR1* and has an important role in immunological function acting as a negative regulator of the inflammation response [38]. Therefore, a mutation that disrupts this gene's function would lead to a hyper inflammatory response, which might account for the elevated IL-10 levels in RA [38] and SS [39].

Based on the significance of the distance measurements, the functional proximity between core and candidate genes on the cluster plot and the assumptions of the connectome analysis, *LRP1* may have an important role in ADs such as RA, SS and AITD via similar mechanisms, networks and/or pathways as *IL10*. Evidence for this interpretation is further enhanced by the fact that the individual homozygous for the *LRP1* variant contains these precise phenotypes (i.e., RA, SS, AITD).

Although *LRP1* has the strongest evidence as a candidate gene, *MICAL1* also may have physiopathological relevance in ADs, as it has a close biological proximity to *PTPN22,* an autoimmune gene. A functional SNP C1858T in *PTPN22* which alters the responsiveness of T and B cells is associated with some ADs in our population including SS [40],[41].

In addition to the genes above, it is also clear that this approach identified well known genes associated with autoimmunity within the exome variant data of these patients, which in this case are *ICA1* and *STAT6* (encoded by the same variant as *LRP1*). This suggests that the filtration strategies we applied have good validity and reliability in identifying potential MAS causing genes. The significantly short functional proximities of these genes is to be expected, given that *ICA1* interacts with *AIRE* in the production of self-antigen [42] and therefore has been functionally linked with SS [43]. The signal transducers and activators of transcription (STATs) including STAT6 are latent cytoplasmic proteins that undergo tyrosine phosphorylation by Janus kinases (JAKs) in response to cytokine exposure (mainly IL-4 and IL-13) in the extracellular milieu [44]. This involves phosphorylation of JAKs, which allows dimerization of STAT molecules, enabling transcriptional regulation of target genes. Transcriptional regulation by *STAT6* occurs as a result of its capability to transform chromatin between open and closed states at target loci [44]. It should be stressed that the variants identified within *ICA1* and *MICAL1* are not categorised as coding or splice-site SNPs. However, this does not mean that these variants are not functionally relevant, because the *ICA1* homozygous variant is seemingly part of a poly A tail, which is suggested through its sequence analysis via the use of the IGV [24-25]. Another possible explanation is that the sequenced region has high levels of sequence conservation. Both of these explanations may account for the assignment of this variant as ‘Disease Causing’ by the MutationTaster algorithm (table 3.3). Conversely, the variant harboured in *MICAL1* is located in a region that could be important in intron splicing regulatory element activity as mentioned earlier [17-18]. This inference is not only based on the results from the Human Splice Finder (motif predictor) [17-18]. Instead, empirical observations from previous studies have illustrated that intronic splicing regulatory elements up to 150 bp from alternatively spliced exons are highly conserved compared to constitutive exons [45]. Therefore, given that this sequence is 22bp from the intron exon boundary of *MICAL1*, it may have high levels of conservation. Thus if this SNP is located in a highly conserved region, it makes sense as to why it is predicted as a functional variant, as it may have important regulatory mechanisms in splicing, based on the evidence obtained thus far [24]-[27].

It must also be noted that *ICA1*, coding for ICA69 autoantigen, has been previously associated with Diabetes mellitus type 1 (T1DM), based on cDNA expression analysis in islet cells, as well as being implicated in SS [46],[48]. It has been observed in past investigations that mice heterozygous for *ICA1* as well as those carrying mutations for *ICA1* and *AIRE* (thereby hindering thymic ICA69 expression), exhibited suboptimal negative selection of ICA69 reactive T cells in the thymus. This can drive autoreactive T cell mediated destruction, as is the case with T1DM, and also cause impaired function of other organs expressing ICA69 (i.e., the thyroid, the salivary glands, the brain, the stomach), meaning that it can contribute to a potential mechanism in the pathogenesis of ADs. This will occur especially if the autoreactive T cells affect the target organ. Further verification of this proposed mechanism is evident through the fact that no islet destruction was observed in cells carrying the ICA69 wildtype [46-47]. Hence this mutation may be important in the SS phenotype observed in the individual homozygous for *ICA1*. In addition, ICA69 autoantibodies have been reported in SS and may reflect the broad spectrum of autoimmune abnormalities in this condition [48].

Although SS and T1DM share several genetics factors, the coexistence of both diseases is uncommon [41]. On the other hand, patients with SS may be prone to develop early subclinical atherosclerosis and have an altered lipid profile with potential atherosclerotic risk [49]. Nevertheless, the role of dyslipidaemia in favouring organic arterial wall damage in these patients appears to be marginal [49]. Thus, other mechanisms including genetics may play a key role in determining the acceleration of atherosclerosis in SS. Therefore the identified variants in our research may not only be relevant for the observed phenotypes in the MAS and SS patients, but also other subphenotypes that could develop in these individuals later on.

3.5: Conclusion

The application of different databases and quality control for filtering purposes has ensured that identified and filtered variants have been corrected for batch effects as well as analysed by any relevant bioinformatics tools, not just in terms of population frequency, but also from a physiological perspective. Thus, based on our results, these genes (in particular *LRP1*) should be considered as strong candidates for conferring risk to autoimmunity. Furthermore, additional variants found in *MACF1, ICA1* and *KIAA0754* may confer susceptibility not only to autoimmunity but also to other diseases such cardiovascular disease [50]. Hopefully our findings can be supported by future analysis of multigenerational families segregating autoimmunity [51], and will help to decipher the common mechanisms of autoimmunity (i.e., the autoimmune tautology) [6].

References

[1] G. Paz-Filho, M.C.S. Boguszewski, C.A. Mastronardi, et al. Whole exome sequencing of extreme morbid obesity patients: translational implications for obesity and related disorders. *Genes,* 5(3):709-25, 2014.

[2] E.T. Cirulli, and D.B. Goldstein. Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat Rev Genet, 11(6):415-25, 2010.

[3] J.M. Anaya. The diagnosis and clinical significance of polyautoimmunity. *Autoimmun Rev*, 13(4-5):423-26, 2014.

[4] J.M. Anaya, J. Castiblanco, A. Rojas-Villarraga A, et al. The multiple autoimmune syndromes. A clue for the autoimmune tautology. *Clin Rev Allergy Immunol,* 43(3):256-64, 2012.

[5] A.S. Johar, J.M. Anaya, D. Andrews, et al. Candidate gene discovery in autoimmunity by using extreme phenotypes, next generation sequencing and whole exome capture. *Autoimmun Rev,* 14(3):204-209, 2015.

[6] J.M. Anaya. Common mechanisms of autoimmune diseases (the autoimmune tautology). *Autoimmun Rev,* 11(11):781-84, 2012.

[7] M.J. Amador-Patarroyo, J.G. Arbelaez, R.D. Mantilla, et al. Sjögren's syndrome at the crossroad of polyautoimmunity. *J Autoimmun,* 39(3):199-205, 2012.

[8] J.M. Anaya, G.J. Tobon, P. Vega, and J. Castiblanco. Autoimmune disease aggregation in families with primary Sjogren's syndrome. *J. Rheumatol,* 33(11):2227-34, 2006.

[9] A. McKenna, M. Hanna, E. Banks et al. The Genome Analysis Toolkit: a Map-Reduce framework for analysing next-generation DNA sequencing. *Genome Res*, 20(9):1297-1303, 2010.

[10] M. DePristo, E. Banks, R. Poplin, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet,* 43(5):491-8, 2011.

[11] G.A. Van der Auwera, M. Carneiro, C. Hartl, et al. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices. *Curr Prot Bioinformatics*, 43:11.10.1-33. Doi: 10.1002/0471250953.bi1110s43, 2013.

[12] H. Thorvaldsdottir, J.T. Robinson, and J.P. Mesirov. Integrative Genomics Viewer (IGV): high performance genomics data visualisation and exploration. *Brief Bioinform,* 14(2):178-92, 2013.

[13] J.T. Robinson, H. Thorvaldsdottir, W. Winckler, et al. Integrative Genomics Viewer. *Nat Biotechnol*, 29(1):24-26, 2011.

[14] Bozeman, MT: SNP & Variation Suite (Version 7.7.6). Golden Helix, Inc, Available from http://www.goldenhelix.com 2014.

[15] Bozeman MT: Variant Classification. In: SNP and Variation Suite Manual version 8.7.2, Copyright 2014, accessed Feb 2015.

[16] N. Sheth, X. Roca, M.L. Hastings, et al. Comprehensive splice-site analysis using comparative genomics. *Nucleic Acids Res,* 34(14):3955-67, 2006.

[17] F.O. Desmet, D. Hamroun, M. Lalande, et al. Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Res,* 37(9):e67 doi: 10.1093/nar/gkp215, 2009.

[18] C. Zhang, W.H. Li, A.R. Krainer, et al. RNA landscape of evolution for optimal exon and intron discrimination. *Proc Natl Acad Sci USA,* 105(15):5797-802.

[19] Y. Itan, S.Y. Zhang, G. Vogt, et al. The human gene connectome as a map of short cuts for morbid allele discovery. *Proc Natl Acad Sci USA,* 110(14):5558-63, 2013.

[20] Y. Itan, M. Mazel, B. Mazel, et al. HGCS: an online tool for prioritizing disease-causing gene variants by biological distance. *BMC Genomics,* 15:256, doi: 10.1186/1471-2164-15-256, 2014.

[21] Yu W, Wulf A, Liu T, Khoury MJ, Gwinn M: Gene Prospector: an evidence gateway for evaluating potential susceptibility genes and interacting risk factors for human diseases. *BMC Bioinformatics,* 9:528, doi: 10.1186/1471-2105-9-528, 2008.

[22] Y. Benjamini and Y. Hochberg. Controlling the False Discivery Rate: A Powerful and Practical Approach to Multiple Testing. *J.R. Stat Soc Series B Stat Methodol,* 57(1):289-300, 1995.

[23] A.L. Price, N.J. Patterson, R.M. Plenge, et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet,* 38(8):904-9, 2006.

[24] H. Thorvaldsdottir, J.T. Robinson, and J.P. Mesirov. Integrative Genomics Viewer (IGV): high performance genomics data visualisation and exploration. *Brief Bioinform,* 14(2):178-92, 2013.

[25] J.T. Robinson, H. Thorvaldsdottir, W. Winckler, et al. Integrative Genomics Viewer. *Nat Biotechnol,* 29(1):24-26, 2011.

[26] J.M. Schwarz, C. Rodelsperger, M. Schuelke, and D. Seelow. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat. Methods,* 7(8):575-6, 2010.

[27] J.M. Schwarz, D.N. Cooper, M. Schuelke, and D. Seelow. MutationTaster2: mutation prediction for the deep sequencing age. *Nat. Methods,* 11(4):361-2, 2014.

[28] W.J. Kent, C.W. Sugnet, T.S. Furey, et al: The genome browser at UCSC. *Genome Res,* 12(6):996-1006, 2002.

[29] A.T. Kraja, D.I. Chasman, K.P. North, et al. Pleiotropic genes for metabolic syndrome and inflammation. *Mol Genet Metab,* 112(4):317-38, 2014.

[30] M. Shibata, S. Yamada, S.R. Kumar et al. Clearance of Alzheimer's amyloid-ss(1-40) peptide from brain by LDL receptor-related protein-1 at the blood-brain barrier. *J. Clin Invest,* 106(12):1489-99, 2000.

[31] K. Zurhove, C. Nakajima, J. Herz, et al. Gamma-secretase limits the inflammatory response through the processing of LRP1. *Sci Signal,* 1(47):ra15, doi: 10.1126/scisignal.1164623, 2008.

[32] A. Nilsson, L. Vesterlund, P.A. Oldenborg, et al. Macrophage expression of LRP1, a receptor for apoptotic cells and unopsonized erythrocytes, can be regulated by glucocorticoids. *Biochem Biophys Res Commun,* 417(4):1304-9, 2012

[33] A. Fernandez-Castaneda, S. Arandjelovic, T.L. Stiles, et al: Identification of the low density lipoprotein (LDL) receptor-related protein-1 interactome in central nervous system myelin suggests a role in the clearance of necrotic cell debris. *J. Biol Chem,* 288(7):4538-48, 2013.

[34] J. Dalli, L.V. Norling, T. Montero-Melendez, et al. Microparticle alpha-2-macroglobulin enhances pro-resolving responses and promotes survival in sepsis. *EMBO Mol Med,* 6(1):27-42, 2014.

[35] S. Mandrekar, Q. Jiang, C.Y. Lee, : Microglia mediate the clearance of soluble Abeta through fluid phase macropinocytosis. J Neurosci 2009; 29:4252-62.

[36] M.H. Biermann, S. Veissi, C. Maueroeder, et al. The role of dead cell clearance in etiology and pathogenesis of Systemic Lupus Erythematosus: dendritic cells as potential targets. *Expert Rev Clin Immunol,* 10(9):1151-64.

[37] I.K. Poon, C.D. Lucas, A.G. Rossi, and K.S. Ravichandran. Apoptotic cell clearance: basic biology and therapeutic potential. *Nat Rev Immunol,* 14(3):166-80.

[38] J.J. Cush, J.B. Splawski, R. Thomas, et al. Elevated interleukin-10 levels in patients with rheumatoid arthritis. *Arthritis Rheum,* 38(1):96-104, 1995.

[39] J.M. Anaya, P.A. Correa, M. Herrera, et al. Interleukin 10 (IL-10) influences autoimmune response in primary Sjögren’s syndrome and is linked to IL-10 gene polymorphism. *J Rheumatol,* 29(9):1874-6, 2002.

[40] L.M. Gomez, J.M. Anaya, C.I. Gonzalez, et al. PTPN22 C1858T polymorphism in Colombian patients with autoimmune diseases. *Genes Immun,* 6(7):628-31, 2005.

[41] J.M. Anaya, L. Gómez, and J. Castiblanco. Is there a common genetic basis for autoimmune diseases? *Clin Dev Immunol,* 13(2-4):185-95, 2006.

[42] S.M. Bonner, S.L. Pietropaolo, Y. Fan, et al. Sequence variation in promoter of Ica1 gene, which encodes protein implicated in type 1 diabetes, causes transcription factor autoimmune regulator (AIRE) to increase its binding and down-regulate expression. *J. Biol Chem,* 287(21):17882-93, 2012.

[43] T.R. Reksten, S.J. Johnsen, M.V. Jonsson, et al. Genetic associations to germinal centre formation in primary Sjögren's syndrome. *Ann Rheum Dis,* 73(6):1253-8, 2014.

[44] A.V. Villarino, Y. Kanno, J.R. Ferdinand, J.J. O'Shea. Mechanisms of Jak/STAT Signaling in Immunity and Disease. *J. Immunol,* 194(1): 21-7, 2015.

[45] Z. Wang, and C.B. Burge Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. RNA 14(5):802-13, 2008.

[46] M. Pietropaolo, L. Castaño, S. Babu, et al. Islet Cell Autoantigen 69 kD (ICA69) Molecular Cloning and Characterization of a Novel Diabetes-associated Autoantigen. *J Clin. Invest,* 92:359-71, 1993.

[47] Y. Fan, G. Gaultierroti, A. Tajima et al. Compromised central tolerance of ICA69 induces multiple organ autoimmunity. *J Autoimmun,* 53:10-25, 2014

[48] P. Pérez, J.M. Anaya, S. Aguilera, et al. Gene expression and chromosomal location for suceptibility to Sjögren's syndrome. *J Autoimmun,* 33(2):99-108.

[49] T.P. Gordon, D. Cavill, P. Neufing, et al. ICA69 autoantibodies in primary Sjögren's syndrome. *Lupus,* 13(6):483-84, 2004.

[50] R. Gerli, E. Bartoloni Bocci, G. Vaudo, et al. Traditional cardiovascular risk factors in primary Sjögren's syndrome. *Rheumatology (Oxford),* 45(12):1580-1, 2006.

[51] J. Cárdenas-Roldán, A. Rojas-Villararga, and J.M. Anaya. How do autoimmune diseases cluster in families? A systematic review and meta-analysis. *BMC Med,* 73, https://doi.org/10.1186/1741-7015-11,73 2013.

Chapter 4: Identification of Candidate Disease Mutations in Familial Autoimmunity

Abstract

Familial autoimmunity and polyautoimmunity represent extreme phenotypes ideal for identifying major genomic variants contributing to autoimmunity. Whole exome sequencing (WES) and linkage analysis are well suited for this purpose due to its strong resolution upon familial segregation patterns of functional protein coding and splice variants. The primary objective of this study was to identify potentially autoimmune causative variants using WES data from extreme pedigrees segregating polyautoimmunity phenotypes. DNA of 47 individuals across 10 extreme pedigrees, ascertained from probands affected with polyautoimmunity and familial autoimmunity, were selected for WES. Variant calls were obtained through Genome Analysis Toolkit. Filtration and prioritization framework to identify mutation(s) were applied, and later implemented for genetic linkage analysis. Sanger sequencing corroborated variants with significant linkage. Novel and mostly rare variants harbored in *SRA1, MLL4, ABCB8, DHX34* and *PLAUR* showed significant linkage (LOD scores are >3.0). The strongest signal was in SRA1, with a LOD score of 5.48. Network analyses indicated that *SRA1, PLAUR* and *ABCB8* contribute to regulation of apoptotic processes. Novel and rare variants in genetic linkage with polyautoimmunity were identified throughout WES. Genes harboring these variants might be major players of autoimmunity. Our findings here were published in the Journal of Autoimmunity, 2016.

4.1 Introduction

Recent evidence supports the involvement of rare genetic variants with a major effect underpinning the etiology of complex disorders and that a large proportion (1/3) of deleterious alleles have a frequency of <5%. This distribution of rare variants is likely due to the fact that deleterious variants will be unfavorable due to selection, thereby reducing their minor allele frequency in any given population [1]. The detection of these rare variants has shown to be achievable using severe phenotypes segregating in exceptional pedigrees [1]. Polyautoimmunity is defined as the presence of more than one autoimmune disease (AD) in a single patient [2]. If three or more ADs coexist, the condition is called multiple autoimmune syndrome (MAS), which characterizes one conspicuous and extreme example of polyautoimmunity [2,3] i.e.: (i) MAS amalgamates signs and symptoms that are present in several ADs, (ii) the MAS signs and symptoms clustering is not random but linked to subtypes, (iii) MAS frequently clusters in families, and (iv) Mendelian segregation and linkage to major loci have been established for MAS [2-3].

During the last years we have ascertained, through probands affected by polyautoimmunity, several multigenerational and extended pedigrees clustering additional relative members affected by either single or polyautoimmunity syndromes (i.e., familial autoimmunity) [4-6]. Several of these pedigrees were ascertained from the “Paisa” genetic isolate, where several major loci to complex disorders, including MAS have previously been mapped [7-10]. Given the phenotypic characteristics of the polyautoimmunity pedigree members it is fair to hypothesize that autoimmunity segregates in those pedigrees in a Mendelian fashion with different patterns of penetrance (Fig. 4.1). In here we report the analysis of whole exome sequencing (WES) of affected and unaffected members of ten pedigrees clustering autoimmune phenotypes (Fig. 4.1). Using classical and modern techniques of genetic pedigree analyses, we found strong and significant signals of genetic linkage of exonic and regulatory mutations with polyautoimmunity. Given the functional nature of these mutations and the strong linkage to polyautoimmunity it is very likely that genes harboring these mutations are major contributors in causing autoimmunity.

4.2 Methods

4.2.1 Families segregating polyautoimmunity.

Phenotypes of the individuals that were subject of exome capture from the ascertained pedigrees are presented in Table 4.1, and the corresponding pedigrees are in figure 4.1. All the individuals were enrolled in the Center for Autoimmune Diseases Research (CREA) at the Universidad del Rosario, in Bogota and Medellin, Colombia. Written informed consent was obtained from all individuals before enrolment in this study, which was approved by the ethics committee of the Universidad del Rosario. Most of the families (47 individuals) were recruited from the Paisa community living in Antioquia, Colombia. Historical evidence has documented that individuals from the Paisa community are endogenous, homogenous and have very little population stratification [11]. All patients were diagnosed and followed-up by a single team (JMA, JCS, ARV, RDM) according to international classification criteria (table 4.2) [12-29].

Family ID	Individual ID	Gender (1 = male, 2 = female)	Individual Phenotype (1 = Unaffected, 2 = affected)
540	FE44	1	1
540	FE43	1	1
540-A	HUM1377	2	2 (AITD, APS, SSc)
540-B	HUM1378	2	2 (SLE, ITP, MG)
3528-A	HUM1381	2	2 (SS, APS)
3528-B	HUM1382	2	2 (SS, AITD, RA, ITP)
3528-C	FE72	2	2 (RA)
3528	FE73	1	1
3543-A	FE53	2	2 (SS)
3543-B	FE54	2	2 (RA, SLE, VIT, AAV)
3686-B	FE75	2	2 (RA)
3686-A	FE74	2	2 (SS, AITD)
5564-A	FE6	2	2 (RA)
5564	FE7	2	1
5564-C	FE8	2	2 (RA, SS, AITD, SLE, DM)
5564-D	FE9	1	2 (RA, SS, SLE, AITD)
5564	FE10	2	1
5564	FE11	1	1
5564	FE12	2	2 (RA, AITD)
5564-H	FE13	2	2 (RA, AITD)
5564-I	FE14	2	2 (SS)
5564	FE15	1	1
5564-M	FE16	2	2 (T1D, AITD)
5627-A	FE79	2	2 (RA, SS)
5627-B	FE80	2	2 (RA)
5627	FE81	2	1
5627-D	FE82	2	2 (AITD, APS)
5627-E	FE83	2	2 (RA, APS)
5653	FE35	1	1
5653	FE36	2	1
5653-A	HUM1374	2	2 (RA, AITD)
5653-B	HUM1375	2	2 (AITD, SLE, APS)
5653-C	HUM1376	2	2 (SLE, AITD, APS)
5653-D	FE30	2	2 (AITD, APS)
5653	FE31	1	1
5653-H	FE33	1	2 (AITD)
5675-A	HUM1379	2	2 (RA, SS)
5675-B	HUM1380	2	2 (AAV)
5675	FE47	2	1
5675	FE48	2	1
5675	FE49	2	1
5744-A	FE51	2	2 (RA, SS, AITD, SLE)
5744-B	FE52	2	2 (VIT, AITD)
10532-D	HUM1383	2	2 (AITD, MG)
10532-B	HUM1384	2	2 (AITD, RA)
10532-A	HUM1385	2	2 (T1D, AITD)
10532	FE97	2	1

Table 4.1: Phenotypes of Individuals in the 10 families segregating autoimmunity.

Autoimmune Disease (abbreviation)	[Ref]
Antiphospholipid syndrome (APS)	[12]
Autoimmune hepatitis (AIH)	[13]
Autoimmune thyroid disease (AITD)	
Hashimoto’s thyroiditis	[14]
Graves’ disease	[15]
Dermato-polymiositis (DM)	[16]
Diabetes mellitus type 1 (T1D)	[17]
Anti-neutrophil cytoplasmic antibodies - associated vasculitis (AAV)	[18]
Myasthenia Gravis	[19]
Idiopathic thrombocytopenic purpura (ITP)	[20]
Megaloblastic anemia (MA)	[21]
Multiple sclerosis (MS)	[22]
Primary biliary cirrhosis (PBC)	[23]
Psoriasis (Pso)	[24]
Rheumatoid arthritis (RA)	[25]
Scleroderma (SSc)	[26]
Sjögren's syndrome (SS)	[27]
Systemic lupus erythematosus (SLE)	[28]
Vitiligo (VIT)	[29]

Table 4.2: Autoimmune diseases investigated in probands and relatives of extreme pedigrees

4.2.2 DNA library preparation, exome capture and sequencing protocol

Libraries were constructed from 1 mg of genomic DNA using an Illumina TruSeq genomic DNA library kit at the Biomolecular Resource Facility (BRF), John Curtin School of Medical Research (JCSMR). Libraries were multiplexed with 6 samples pooled together (500 ng of each library). Exons were enriched from the pooled 3 mg of library DNA using the Nimblegen Exome enrichment kit (BRF). Each exome-enriched pool was run on a 100-base-pair paired end run on an Illumina HiSeq 2000 sequencer (BRF, JCSMR) [30-32].

4.2.3 Sequence read processing, alignment and variant calling

The sequencing image data was converted to FASTQ files containing DNA base calls (A, C, G and T) and quality scores using the Illumina CASAVA pipeline (a software programs that converts raw image data into sequences). The resulting FASTQ files were further processed for variant analysis. The workflow for data curating and analysis for variant calling was developed by the Genome Discovery Unit (GDU), at the Australian National University. Key components of the workflow include: i) Quality assessment; ii) Read alignment using the BOWTIE aligner; iii) Local realignment around the known and novel insertions/deletions (indel) regions to refine indel boundaries; iv) Recalibration of base qualities; v) Variant calling using the GATK (Genome Alignment Tool Kit) algorithm; and vi) Assigning quality scores to variants (see more detailed workflow information published elsewhere) [30-35].

4.2.4 Linkage analysis and CLRT unified framework

We used the Pedigree Variant Annotation Analysis and Search Tool (pVAAST) [36-38] to identify candidate genetic variants enriched in affected family members co-segregating (in genetic linkage) with polyautoimmunity. In brief, the pVAAST algorithm allows the identification of potential disease causing variants through a method that combines classical linkage analysis (Elston-Stewart algorithm), with casecontrol association, and functional impact of the tested variant using a Composite Likelihood Ratio Test (CLRTv). This produces a gene-based LOD score and CLRTp (pVAAST test statistic) value. Dominant, codominant, and recessive models were maximized under the assumption of upper bound penetrance proximal to 100%. The significance of these test statistics is derived using a combination of permutation tests (in this case 10,000) without replacement and gene-drop simulation. This was achieved using a total of 1366 individuals as controls. This control dataset includes 1057 individuals from the 1000 genomes project (Phase 1), 184 Danish exomes, 10 individuals from the 10Gen database and 62 additional unaffected controls from the same geographical and cultural region where the extreme pedigrees were ascertained (the 1000 genomes ephase1e has a set of 60 individuals recruited from Medellin, Colombia; the same area of ascertainment of most of these pedigrees).

4.2.5 Further refinement and curating of candidate gene list generated via pVAAST algorithm

Candidate genes reported as containing a significant pVAAST test statistic (based on LOD score and case-control allele frequencies) were further filtered as follows: of all the genes that were significant, those containing a CLRTp score greater than the mean and with a LOD score greater than 3.0 were retained for further downstream analysis. This is conditional on the fact that variants within these genes had an allelic frequency of < 5% in the ExAC (Exome Aggregation Consortium), dbSNP and the 1000 Genomes database. Afterwards, additional filtration was conducted based upon annotations from variant effect predictors (including: Polyphen2, SIFT, MutationTaster, and FATHMM), as implemented in the DNA-seq analysis package SVS (SNP and Variation Suite) Version 8.3.4 (Golden Helix, Bozeman, USA), as well as PROVEAN (Protein Variant Effect Analyser) [39-45]. Mutations predicted to be damaging from at least one of these variant effect predictor algorithms are considered as plausible disease candidates for further downstream analysis.

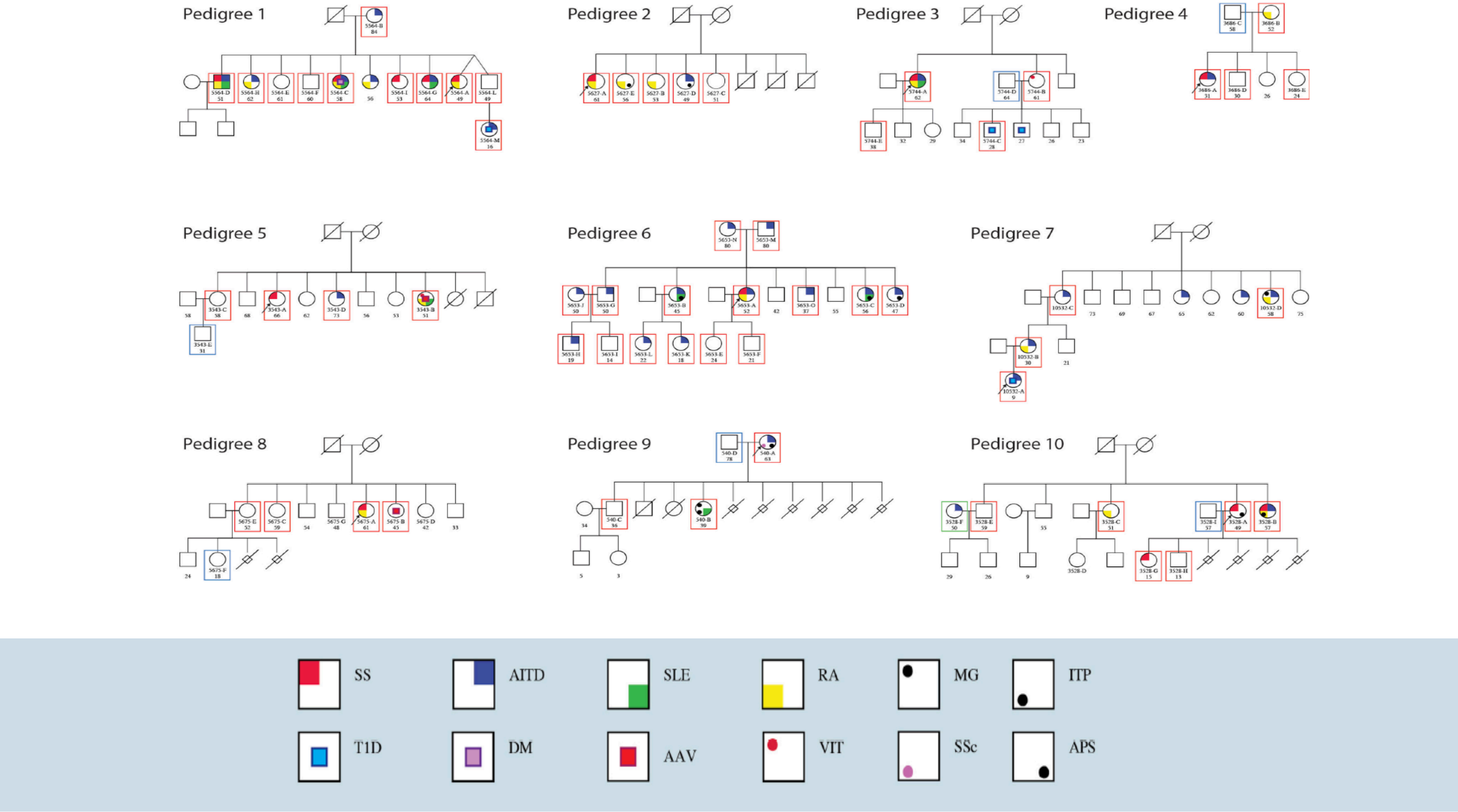

Figure 4.1: Pedigrees from the Paisa community segregating familial autoimmunity subject to WEC and linkage analysis. Phenotype abbreviations are for diseases corresponding to those listed in table 4.2.

4.2.6 Sanger sequencing and variant quality scores

Sanger sequencing was performed for the validation of the gene variants reported herein as previously described [30]. Briefly, a flanking region around each sequence variant site was amplified by PCR. We used the following PCR conditions for the amplification of all the amplicons: (1) initial activation step of 3 min at 94 C, (2) 40 cycles as follows: 45 s of denaturing at 94 C, 30 s of annealing between 53 and 62 C (according to the primer pair, Supplementary Table 3), 60 s of extension at 72 C, and (3) final extension step of 10 min at 72 C. A 15-mL aliquot of the PCR product was analysed by electrophoresis in a 1.5% agarose gel to confirm the expected size of the amplicons. Afterwards, 85 mL of each PCR product was purified with the Qiaquick nucleotide removal kit (Qiagen, Valencia, CA, USA) following the manufacturer’s guidelines. Thereafter, the purified PCR products were spectrophotometrically quantified with a NanoDrop ND-1000 (Wilmington, DE, USA), and sent for Sanger sequencing to the Australian Cancer Research FoundationdBiomolecular Resource Facility (BRF) at the John Curtin School of Medical Research. Bidirectional sequencing of PCR amplicons were carried out by using Big DyeTM chemistry (Big Dye Terminator, Version 3.1; Applied Biosystems, Foster City, CA, USA) with the sequencing primers reported in Supplementary Table 3 of our manuscript in the Journal of Autoimmunity. The sequencing protocol was followed according to the BRF standard operative procedures. In addition to the Sanger sequencing, variant Phred and Mapping Quality scores are also obtained, using IGV [46-47]. For the variant under consideration, the Phred score will quantify the probability of an incorrect base call. Meanwhile, mapping quality determines the likelihood of an incorrect alignment of the sequence read containing the identified variant(s) [46-47].

4.2.7 Metacore pathway and network analysis

To identify potentially enriched polyautoimmunity related physiological pathways involving candidate genes from the pVAAST data, network analyses were performed. For constructing networks and pathways, were examined with the 'Shortest Paths' algorithm implemented within the MetaCore software suite (Version 6.24, Build 67895, Thomson Reuters, New York, USA) [48]. Details regarding some the differences between the various algorithms can be found in the MetaCore Manual. These procedures allowed us to obtain a heuristic integration of maps and networks and rich ontologies for diseases based on the potential biological functions of these candidate genes.

4.3 Results

Significant signals of linkage were obtained for variants harbored in five genes: *SRA1*, *MLL4*, *ABCB8*, *PLAUR*, and *DHX34* (Table 4.3). When conducting permutation tests (number of permutations / 10,000), each of these genes ascertained a p-value of < 0.005 for the CLRTp test statistic. *SRA1*, *ABCB8* and *PLAUR* exhibited a maximisation of the LOD score (and consequently the CLRTp test statistic) under the recessive model for linkage analysis. In the case of *DHX34* and *MLL4*, the greatest maximisation occurred under the dominant model. In addition, the input parameters suggest that LOD scores are maximized for candidate variants within these genes at a disease allele penetrance between 70 and 99%. This result can account for the presence of potential phenocopies within this cohort.Sanger sequencing, as shown in Figure 4.2, corroborated these variants. Also, mapping quality and Phred scores from IGV quantified the accuracy of the base calls and read alignments, as shown in table 4.4 [46-47]. These results indicated that the likelihood of incorrect base calls and alignments is is <0.1% and 0.01% respectively.

Gene	Variant (rs id)	LOD Score	Linkage Model	pVAAST CLRTp Score	Permutation Based P-Value	Alleles Ref/Alt
SRA1	5:139931629-Ins (rs5871740)	5.5	Recessive	214.030	0.0001	C/CTCG
SRA1	5:139931629-SNV (rs202193903)	5.5	Recessive	214.030	0.0001	C/G
ABCB8	7:150744528	4.1	Recessive	27.359	0.00903	G/T
ABCB8	7:150744370	4.1	Recessive	27.359	0.00903	CGT/-
DHX34	19:47883126-SNV (rs151213663)	3.8	Dominant	46.873	0.0003	C/T
PLAUR	19:44153100 (rs4760)	3.6	Recessive	20.031	0.004	A/G
MLL4	19:36218440-SNV (rs186268702)	3.4	Dominant	137.17	0.0001	G/A

Table 4.3: LOD score and pVAAST test statistics of genes harbouring genetic variants with significant values of linkage to MAS in the 10 families. LOD scores are calculated based on the Elston-Stewart algorithm, pVAAST scores are gene-based and P-values are derived from permutation (without replacement) and gene-drop simulation.

Gene	Variant	Mapping Quality	Phred Score
MLL4	19:36218440-SNV (rs186268702)	42	34
DHX34	19:47883126-SNV (rs151213663)	42	36
ABCB8	7:150744528-SNV	42	32
ABCB8	7:150744370-SNV (rs71712832)	42	31
SRA1	5:139931629-Ins (rs7871740)	42	33
SRA1	5:139931629-SNV (rs202193903)	40	31
PLAUR	19:44153100-SNV (rs4760)	42	33

Table 4.4: Mapping quality and Phred scores of chosen variants generated from Linkage and CLRT in pVAAST. Mapping quality indicates the strength of the alignment. I.e. the likelihood of incorrect alignment of sequence reads harbouring a given genetic variant, with the reference sequence. Phred score quantifies the likelihood of an incorrect base call at the position of identification for the variant. Calculations of these scores are based on the log10 scale.

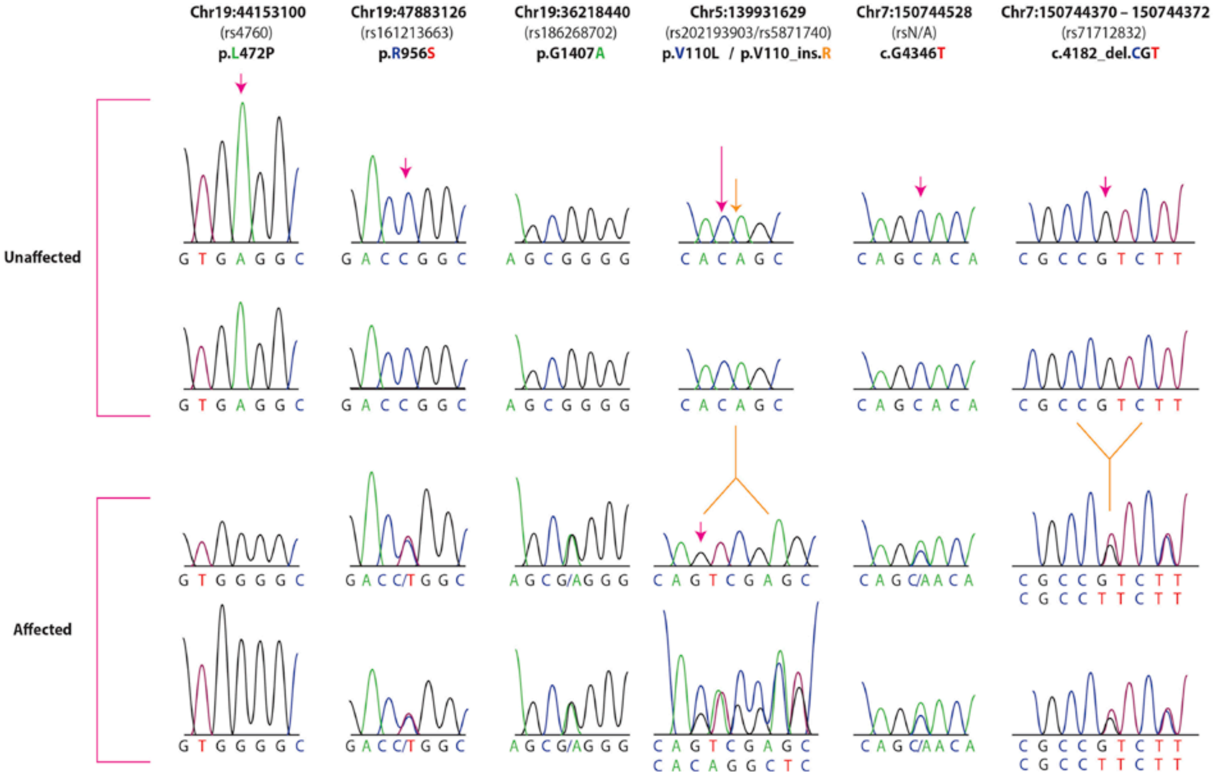

Figure 4.2. Chromatograms of the results of the Sanger sequencing of unaffected (top rows) and affected individuals (bottom rows) segregating variants harbored in genes in significant genetic linkage with polyautoimmunity. The bold subheading on the top indicates the chromosome (Chr) number and the variant position for each one of the studied genes; from left to right they represent *PLAUR*, *DHX34*, *MLL4*, *SRA1*and *ABCB8* (last two panels). Also each row represents the DNA of a separate individual. The available reference SNP (rs) numbers are also shown between parentheses. Below, it is shown the position of the changes either within the protein (p) or the cDNA (c) sequence for the 5 coding and 2 non-coding variants, respectively. In each coding variant the left capital letter indicates the aminoacid occurring in the unaffected individuals, and the right capital letter shows either the aminoacid change (rs4760, rs151213663, rs186268702 and rs202193903) or the insertion (ins) of an arginine (R) in the rs5871740 variant that occurs in the affected individuals. With regards to the non-coding variants, the capital letters indicate a nucleotide change (rsN/A) or a trinucleotide deletion (del) in the rs71712832 variant. In both coding and non-coding variants each capital letter color corresponds to that of the nucleotide involved in the variation as shown in Sanger chromatograms. A short sequence is shown under each individual chromatogram. In case of heterozygosity, there are two sequence reads representing the maternal and parental DNA. The red and orange arrows signal the nucleotide position where the variations occur. The orange inverted and non-inverted Y-shaped pointers depicts the expansion and contraction of the sequence occurring as consequence of the insertion and deletion, respectively, in affected individuals.

Genotypes of individuals carrying the corresponding alternate allele are given in table 4.5. Table 4.6 shows the number of affected and unaffected carriers respectively. The data here also shows that all of these variants are rare or novel, according to the 1000 Genomes and dbSNP databases, and that SRA1 has the highest number of affected carriers, across all analysed families. This gives an accurate reflection of the fact that this candidate gene generated the highest LOD score

Gene	Variant	Individuals with Variant (Genotype, * (a = homozygous, b = heterozygous)
MLL4	19:36218440-SNV (rs186268702)	FE30(b), FE33(b), HUM1374(b), HUM1375(b), HUM1376(b)
DHX34	19:47883126-SNV (rs151213663)	FE6(b), FE8(b), FE12(b), FE13(b)
ABCB8	7:150744528	FE8(b), FE9(b), FE14(b)
ABCB8	7:150744370 (rs71712832)	FE6(b), FE7(b), FE8(b), FE9(b), FE10(b), FE14(b), FE16(b), FE53(b)
SRA1	5:139931629-Ins (rs7871740)	FE13(a), FE14(a), FE16(a), FE75(a) FE6(a), FE79(a), FE80(a), FE9(a), FE12(a), FE52(a), FE53(a), FE82(a), FE83(a) HUM1385(a), FE10(b), FE11(b), FE31(b), HUM1374(b), HUM1375(b), HUM1376(b) FE97(b), FE30(b) HUM1381(b), FE8(b)
SRA1	5:139931629-SNV (rs202193903)	FE6(a), FE8(b), FE9(b), FE14(b), FE11(b), FE16(b), FE12(b), FE79(b), FE82(b), FE83(b), FE51(b), FE52(b), FE53(b), FE43(b)
PLAUR	19:44153100-SNV (rs4760)	FE8(a), FE12(a), HUM1379(a), HUM1380(a), FE79(a), FE82(a), FE6(a), FE7(b), FE9(a), FE10(b), FE14(b), FE16(a), FE13(b), FE53(a)

Table 4.5: Genotypes of candidate variants in genes in affected and unaffected individuals giving significant LOD scores. Data is presented for individuals across all 10 analysed families.

Variant/Gene	Number of Affected Individuals (Across Families)	Number of Unaffected Individuals (Across Families)	dbSNP allele frequency (build 141)	1K allele Frequency (Phase 3)	ExAC Frequency
5:139931629-Ins (rs7871740) *SRA1*	14 homozygous 5 heterozygous	4 heterozygous	Absent	Absent	Absent
5:139931629-SNV (rs202193903) *SRA1*	1 homozygous 11 heterozygous	2 heterozygous	Absent	Absent	Absent
19:36218440-SNV (rs186268702) *MLL4*	5 heterozygous	0	0.184%	0.1% (real allele freq, other based on sample counts)	0.07%
19:47883126-SNV (rs151213663) *DHX34*	4 heterozygous	0	0.587%	0.24%	0.4718%
7:150744528 *ABCB8*	3 heterozygous	0	Absent	Absent	Absent
7:150744370 *ABCB8*	6 heterozygous	2 heterozygous	30.367%	31.73%	Absent
19:44153100-SNV (rs4760) *PLAUR*	2 (homozygous), 8 (heterozygous)	5 heterozygous	9.248%	6.85% (1.2% homozygotes)	12.25% (Homozygotes 1.9%)

Table 4.6: Number of affected and unaffected individuals with mutations in candidate genes from the pVAAST analysis, and allele frequencies in external databases (dbSNP, 1000 Genomes and ExAC).

In addition to the high LOD scores and significant pVAAST statistics generated from the aforementioned analysis, these genes may also have substantial relevance in biological pathways and networks, thereby influencing key processes that underpin the physiological basis for autoimmunity in these patients (Table 4.7). In particular *SRA1* and *PLAUR* are seemingly both involved in the negative regulation of apoptosis (P-value = 1.791^{e-3}), as well as negative regulation of cysteine type endopeptidase activity (Pvalue = 1.287e^{-5}). As indicated in the Metacore analysis in Fig. 4.3 and Table 4.7, these processes seemingly occur as a consequence of the SF1 transcription factor binding to and subsequently activating the SRA1 protein. This in turn activates the ESR1, which then subsequently has a seemingly inhibitory effect on the protein encoded by *PLAUR* (Table 4.7 and Fig 4.3).

Gene(s)	Network Algorithm	GeneGo Process Annotation	Annotation Process P-Value	Network Nodes
PLAUR *SRA1*	Shortest Paths	Negative regulation of apoptosis	1.791e-13	SF1 SRA1 ESR1 (nuclear) PLAUR
PLAUR *SRA1*	Shortest paths	Negative Regulation of Cysteine type endopeptidase activity	1.287e-5	SF1 SRA1 ESR1 (nuclear) PLAUR
PLAUR *SRA1*	Shortest Paths	Positive Regulation of Phosphorylation	1.592e-20	SF1 SRA1 ESR1 (nuclear) PLAUR
PLAUR	Shortest Paths	Negative Regulation of Proteolysis	1.761e-4	ESR1 (nuclear) PLAUR
PLAUR	Shortest Paths	Urokinase Plasminogen Activator Signalling	2.404e-3	ESR1 (nuclear) PLAUR

Table 4.7. GeneGo ontologies for biological process descriptions produced by the Metacore network analysis for the candidate gene list. Table gives the nodes corresponding to the process descriptions as well as the P-value.

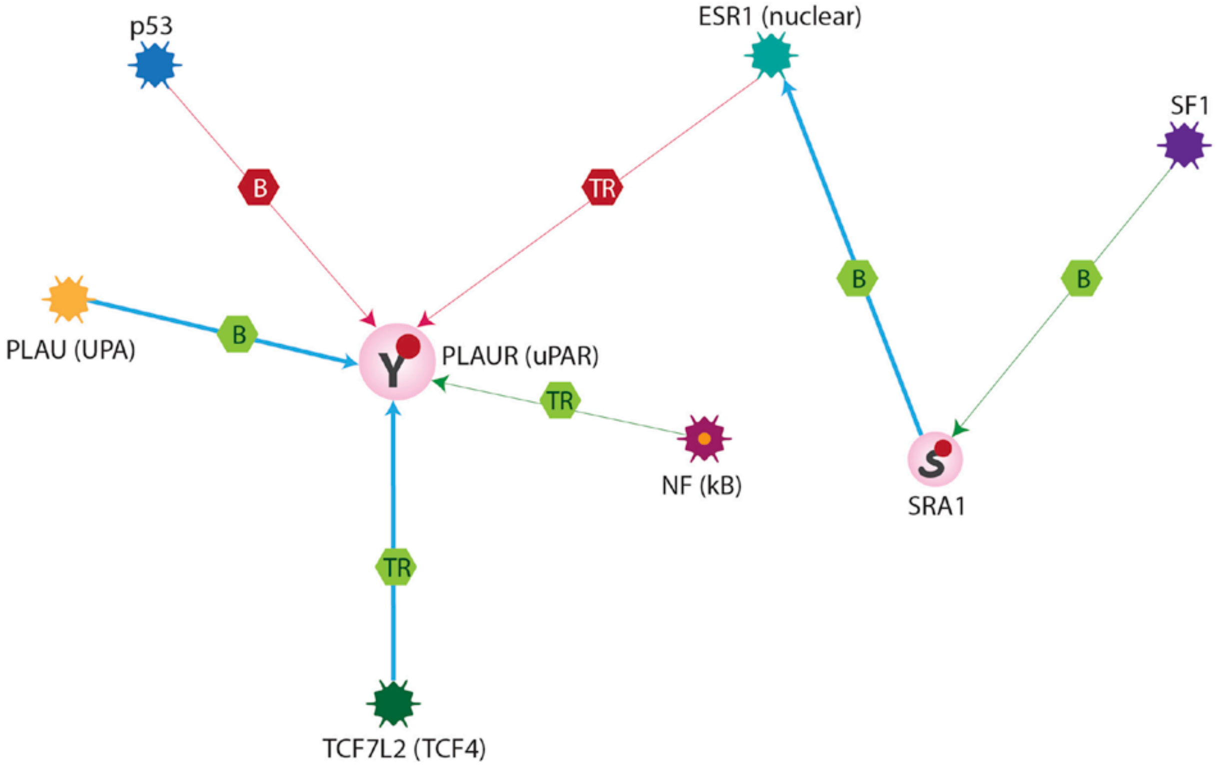

Figure 4.3: Metacore network analysis of candidate genes from the pVAAST analysis. Network is generated by the ‘Shortest Paths’ algorithm. Proteins encoded by genes from candidate list are generated by circular red dot. Abbreviations for the mechanism of interaction between each node for each of the respective proteins is as follows: TR = transcriptional regulation, B = binding. In addition, effects and mechanisms of these types of interactions are given by the following colour coordination: red = inhibition, green = activation**.**

4.4 Discussion

As a whole, we are presenting significant evidence of linkage of functional exonic variants harbored in five genes that cosegregate with polyautoimmunity in extreme pedigrees clustering autoimmune phenotypes. The power of exome sequencing is further enhanced by the fact that this study was done in families, most of them belonging to a genetic isolate, a circumstance that increases genetic and environmental homogeneity. The strongest candidate of the five genes identified is *SRA1* with a LOD score of 5.48. No unaffected individual was heterozygous for both of the identified variants, harbored in this locus. The 3 base insertion at chr5:139931629 (which was denoted to be damaging by the Provean variant effect prediction algorithm, as a consequence of an Arginine insertion between the Valine and Alanine at this position) was not present in any unaffected individual. The fact that *SRA1* has been enriched according to the linkage analysis and pVAAST test statistics in affected pedigree members is further supported by the aforementioned network analysis. The relationship between *SRA1* and the immune system remains to be studied in more detail. However, it is noteworthy to mention that in mouse studies SRA1 levels of mRNA expression were found to be relatively high in mouse spleen, which presumably exerts proinflammatory actions [49]. In fact, *SRA1* KO mice showed significantly reduced expression of a subset of inflammatory-related genes, including tumor necrosis factor-a (TNF-a) and monocyte chemotactic protein-1 (MCP-1), that was accompanied by decreased levels of blood TNF-a [49]. Another candidate gene is *PLAUR* (LOD = 3.62), which encodes the urokinase plasminogen activator receptor (uPAR). When the endogenous ligand urokinase plasminogen activator (uPA) binds to uPAR, it triggers the conversion of plasminogen into plasmin, an active serine protease that is involved in key pathophysiological mechanisms occurring in cancer [50]. Interestingly, uPA and uPAR occur in immune cells, in particular on activated T-cells [51] and monocytes [52], and several of the uPAeinduced effects are independent from plasminogen such as regulation of cell migration, angiogenesis, and adhesion [53]. Also, uPAR expression is enhanced by proangiogenic as well as proatherogenic growth factors and cytokines such as IL-1, suggesting its involvement in inflammatory and proliferative processes [54]. The Metacore network analysis of candidate genes from the pVAAST analysis is depicted in Fig. 4.3. The fact that the analysis revealed both *PLAUR* and *SRA1* are involved in the negative regulation of apoptosis and cysteine type endopeptidase activity suggests they have a potentially important contribution to the autoimmune pathophysiology within these patients. Hypothetically, the occurrence of the rare variants harbored in *SRA1* and *PLAUR* could contribute to the development of autoimmunity by dysregulating apoptosis. Thus, a putative explanation of this network would be that the transcription factor stereoidogenic factor-1 (SF1) increases the expression of SRA1, which in turn can positively modulate anti-apoptotic pathways mediated by estrogen receptor-1 (ESR1), via a binding interaction [55-56]. For instance, downstream anti-apoptotic pathways activated by ESR1 could counter PLAUR-induced apoptosis in line with the negative interaction depicted between ESR1 and PLAUR in the network analysis described in Fig. 3. Additionally, in this network, the tumor suppressor protein p53 is hypothesized to negatively regulate PLAUR, presumably by its non-apoptotic actions [57]. On the other hand PLAUR receives stimulatory feedback from its ligand (PLAU), and two transcription factors named transcription factor 7-like 2 (TCF7L2) and nuclear factor-kappa B (NF-kB), which in turn could positively influence PLAUR-mediated apoptosis. The P-values indicate that the association between the corresponding nodes and these GeneGo processes is highly unlikely to have arisen by chance. Furthermore, based on the results from the pVAAST and Metacore test statistics, it can be hypothesized that the negative regulation of apoptosis may be disrupted as a consequence of deficient cysteine type endopeptidase activity due to one or both mutations in each of these respective genes. Therefore, a deficiency in negative regulation of apoptotic signaling, may lead to enhanced destruction of non-foreign cells, underpinning the basis of the ADs carried by these 32 affected patients. The crucial role of apoptosis in autoimmunity has been studied since more than 20 years [58-60].

Thus, the possible functional relationship between *SRA1* and *PLAUR* and autoimmune-related apoptosis should be investigated in future studies. Although the remaining genes identified from the pVAAST analysis did not show evidence of pathway or physiological significance in the Metacore analysis, this does not rule out their potential relevance in autoimmunity. A brief description of their biological relevance is provided below. *MLL4* encodes an enzyme named histone methyl transferase that methylates ’Lys-40 of histone H3, which is a key epigenetic modification involved in gene activation. This enzyme can cause diand trimethylation to H3 histones interacting with transcription start sites during gene transcription [61-62] and monomethylations to H3 histones associated with enhancer sequences [63]. *MLL4* appears to be involved in different types of cancers such as endometrial, large intestine, glioma and liver carcinomas [64]. Mutations harboured in its paralog gene, namely *MLL2*, have been associated with Kabuki syndrome, which is a complex multisystem developmental disorder characterized by craniofacial, intellectual and cardiac defects [65]. Remarkably, it has been recently reported that patients with Kabuki syndrome carry mutations within *MLL2* presented in humoral immune diseases and in some cases, autoimmunity [66]. In the case of *MLL4*, empirical data is also present for its potential relevance in autoimmunity. It has been found that the methyltransferase encoded by *MLL4* seemingly binds to PTIP (Pax Transactivation Domain Interacting Protein) [67]. PTIP in turn controls for class switch recombination on the immunoglobulin heavy chain locus. Thus a mutation in *MLL4*, may affect the binding of the methyltransferase with PTIP, thereby potentially resulting in defects in class switching. Consequently, this evidence suggests that *MLL4* might also be involved in key pathophysiological aspects underlying autoimmunity.

ABCB8 encodes a multi-pass protein (ABCB8) that is an integral component of the inner mitochondrial membrane. ABCB8 is an ATP-dependent transporter involved in mitochondrial iron export, which is crucial for normal cardiac function. Mice, with deficient expression of ABCB8 in their heart showed aberrant iron homeostasis, increased mitochondrial damage and developed cardiomyopathy [68]. ABCB8 is a member of the MDR/TAP subfamily. Members of this subfamily are involved in antigen processing and presentation by pumping degraded cytosolic peptides across the endoplasmic reticulum into the membrane-bound compartment where class I molecules assemble [69-70]. The identified variants chr7:150744528 and chr7:150744370 both are located in the E2F1 binding site (http://htd.cbi.pku.edu.cn). By themselves the mutations likely decrease ABCB8 expression, reduce mitochondrial potential, render the cells susceptible to redox stress, and promote cell death. The effects of the mutations may hence converge with cell cycle arrest of infected cells (e.g., by Parvovirus B19) [71]. Thus, the possibility exists that mutations of ABCB8 could contribute to autoimmunity by altering cell death and antigen processing and presentation to the adaptive immune system.

Finally, *DHX34* encodes an RNA helicase (DHX34). In vitro studies suggested that DHX34 is involved in nonsense-mediated decay (NMD), a surveillance process that degrades aberrant mRNAs that harbour premature stop codons and also might regulate the abundance of RNAs [72]. DHX34 belongs to the DEAD box protein family; another RNA helicase from this family such as DHX33 has been reported to play a crucial role in sensing viral RNA in myeloid dendritic cells [73]. The possible biological relationships between DHX34 and autoimmunity, remains to be elucidated.

4.4.1 Study’s limitations

We acknowledge the lack of functional genetics as a shortcoming of our study. In addition and as we have mentioned elsewhere: hybridization probes are not available for all annotated exons within the gold standard databases. Also, exome sequencing is not able to detect mutations in non-coding DNA that alter gene function by various regulatory mechanisms and enhancer effects. Such variants are emerging as important contributors to genetic disease and they occur in > 98% of the human genome, which is missed by exome capture [3]. Thus, it is recognized that all genes associated with polyautoimmunity were not included in our study and that new associated genes may be discovered as updated information becomes available. It is worth mentioinng that the dissection of major genes, harboring mutations predisposing to polyautoimmunity, in this manuscript, does not discard interactions of these loci with environmental causes (i.e. the autoimmune ecology) [74], neither does it the switch of the polyautoimmunity phenotype on by infectious diseases. With the available methods for analysis of pedigrees it is very challenging to dissect epigenetic effects. The twins and casecontrol cohorts are better suited to characterize these nonhereditary factors than the use of a limited number of extended and multigenerational pedigrees that maximize the best strategies of physical gene mapping.

4.5 Conclusions

In summary, our linkage analysis in combination with the pVAAST composite likelihood ratio test has successfully identified 5 candidate genes that account for the observed autoimmune phenotypes in extreme pedigrees with polyautoimmunity. The strongest candidate from a statistical and biological relatedness point of view was by far *SRA1*. It is hoped that further functional studies can validate the postulation for the contribution of these genes in ADs.

References

[1] G. Gibson. Rare and common variants: twenty arguments. *Nat Rev Genet*, 13(2):135-45, 2011.

[2] J.M. Anaya. The diagnosis and clinical significance of polyautoimmunity. *Autoimmun Rev,* 13(4-5):423-26, 2014.

[3]. J.M. Anaya. Common mechanisms of autoimmune diseases (the autoimmune tautology). *Autoimmun Rev,* 11(11):781-84, 2012.

[4] J.M. Anaya, J. Castiblanco, A. Rojas-Villarraga A, et al. The multiple autoimmune syndromes. A clue for the autoimmune tautology. *Clin Rev Allergy Immunol,* 43(3):256-64, 2012.

[5] J. Castiblanco, J.C. Sarmiento-Monroy, R.D. Mantilla, et al. Familial aggregation and segregation analysis in families presenting autoimmunity, polyautoimmunity and multiple autoimmune syndrome. *J. Immunol Res,* 572353, doi: 10.1155/2015/572353 2015.

[6] J. Cardenas-Roldan, A. Rojas-Villararga and J.M. Anaya. How do autoimmune diseases cluster in families? A systematic review and meta-analysis. *BMC Med,* 73, https://doi.org/10.1186/1741-7015-11,73 2013.

[7] J.I. Velez, S.C. Chandrasekharappa, E. Henao, et al. Pooling/Bootstrap GWAS (pbGWAS) identifies new loci modifying age of onset in PSEN1 p.Glu280Ala Alzheimer's Disease. *Mol. Psychiatry,* 18(5):568-75.

[8] M. Camargo, D. Rivera, L. Moreno, et al. GWAS reveals new recessive loci associated with non-syndromic facial clefting. *Eur J Med Genet,* 55(10):510-14, 2012.

[9] M. Arcos-Burgos, M. Jain, M.T. Acosta, et al. A common variant of the latrophilin 3 gene, LPHN3, confers susceptibility to ADHD and predicts effectiveness of stimulant medication. *Mol. Psychiatry,* 15(10):1053-66, 2010.

[10] M.L. Marazita, A.C. Lidral, J.C. Murray JC, et al. Genome scan, fine-mapping, and candidate gene analysis of non-syndromic cleft lip with or without cleft palate reveals phenotype-specific differences in linkage and association results. *Hum. Hered,* 68(3):151-70, 2009.

[11] M.L. Bravo, C.Y. Valenzuela and O.M. Arcos-Burgos. Polymorphisms and phyletic relationships of the paisa community from Antioquia (Colombia). *Gene Geogr,* 10(1):11-7, 1996.

[12] J.A. Gómez-Puerta JA, and R. Cervera. Diagnosis and classification of the antiphospholipid syndrome. *J Autoimmun,* 48-49:20-25, 2014.

[13] A.J.Czaja. Current concepts in autoimmune hepatitis. *Ann Hepatol,* 4(1):6-24, 2005.

[14] R.S. Lindsay, and A.D. Toft. Hypothyroidism. *Lancet,* 349(9049):413-417, 1997.

[15] L. DeGroot, and J.L. Jamenson. Thyroid Gland (Part III). In Endocrinology, 4th Ed. Editors. Philadelphia: Saunders, 2000.

[16] A. Bohan, and J.B. Peter. Polymyositis and dermatomyositis. *N Engl J Med,* 292(7):344-347, 1975.

[17] Report of the Expert Committee on the Diagnosis and Classification of Diabetes Mellitus. *Diabetes Care,* 20(7):1183-1197, 1997.

[18] R.Y. Leavitt, A.S. Fauci, D.A. Bloch, et al. The American College of Rheumatology 1990 criteria for the classification of Wegener’s granulomatosis. *Arthritis Rheum,* 33(8):1101-1107, 1990.

[19] A. Jaretzki, R.J. Barohn, R.M.Ernstoff, et al. Myasthenia gravis: Recommendations for clinical research standards. *Neurology,* 55:16–23, 2000.

[20] British Committee for Standadrds in Haematology Genral Haematology Task Force. Guidelines for the investigation and management of idiopathic thrombocytopenic purpura in adults, children and in pregnancy. *Br J Haematol,*120(4):574-596, 2000.

[21] B. Babior. The Megaloblastic Anemias. *In Williams’ Hematology6th Ed,* E Beutler, et Al, Editors. New York: McGraw-Hill; 6th Ed 2000.

[22] W.I. McDonald, A. Compston, G. Edan, et al. Recommended diagnostic criteria for multiple sclerosis: guidelines from the International Panel on the diagnosis of multiple sclerosis. *Ann Neurol,* 50(1):121-127, 2001.

[23] M.M. Kaplan. Primary biliary cirrhosis. *N Engl J Med,* 335(21):1570-1580, 1996.

[24] I. Freedburg. *In Fitzpatrick’s Dermatology in General Medicine, 5th Ed*. New York: McGraw-Hill 1999.

[25] F.C. Arnett, S.M. Edworthy, D.A. Bloch, et al. The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. *Arthritis Rheum,* 31(3):315-324, 1988.

[26] A. Masi, G. Rodnan, T. Medsger. Preliminary criteria for the classification of systemic sclerosis (scleroderma). Subcommittee for scleroderma criteria of the American Rheumatism Association Diagnostic and Therapeutic Criteria Committee. *Arthritis Rheum,* 23(5):581-590, 1980.

[27] C. Vitali, S. Bombardieri, R. Jonsson R et al. Classification criteria for Sjögren's syndrome: a revised version of the European criteria proposed by the American-European Consensus Group. *Ann Rheum Dis,* 61(6):554-558, 2002.

[28] M.C. Hochberg. Updating the American College of Rheumatology revised criteria for the classification of systemic lupus erythematosus. *Arthritis Rheum,* 40(9):1725, 1997.

[29] P. Amerio, M. Tracanna, P. De Remigis, et al. Vitiligo associated with other autoimmune diseases: polyglandular autoimmune syndrome types 3B+C and 4. *Clin Exp Dermatol,* 31(5):746-749, 2006.

[30] A. McKenna, M. Hanna, E. Banks et al. The Genome Analysis Toolkit: a Map-Reduce framework for analysing next-generation DNA sequencing. *Genome Res,* 20(9):1297-1303, 2010.

[31] M. DePristo, E. Banks, R. Poplin, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet,* 43(5):491-8, 2011.

[32] G.A. Van der Auwera, M. Carneiro, C. Hartl, et al. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices. *Curr Prot Bioinformatics*, 43:11.10.1-33. Doi: 10.1002/0471250953.bi1110s43, 2013.

[33] A.S. Johar, J.M. Anaya, D. Andrews, et al. Candidate gene discovery in autoimmunity by using extreme phenotypes, next generation sequencing and whole exome capture. *Autoimmun Rev,* 14(3):204-209, 2015.

[34] A.S. Johar, C. Mastronardi, A. Rojas-Villarraga, et al. Novel and rare functional genomic variants in multiple autoimmune syndrome and Sjogren's syndrome. *J. Trans. Med*, 13:173, doi:10.1186/s12967-015-0525-x, 2015.

[35] G. Paz-Filho, M.C.S. Boguszewski, C.A. Mastronardi, et al. Whole exome sequencing of extreme morbid obesity patients: translational implications for obesity and related disorders. *Genes*, 5(3):709-25, 2014.

[36] T.M. Darlington, R. Pimentel, K. Smith, et al. Identifying rare variants for genetic risk through a combined pedigree and phenotype approach: application to suicide and asthma. *Transl. Psychiatry*, 4:e471, doi: 10.1038/tp.2014.111, 2014.

[37] H. Hu, J.C. Roach, H. Coon, et al. A Unified Test of Linkage Analysis and Rare-Variant Association for Pedigree Sequencing Data. *Nat Biotechnol*, 32(7):663-9, 2014.

[38] B. Kennedy, Z. Kronenberg, H. Hu, et al. Using VAAST to Identify Disease-Associated Variants in Next-Generation Sequencing Data. *Curr Prot Hum Genet,* 81: 6.14.1-25. doi: 10.1002/0471142905.hg0614s81, 2014.

[39] I.A. Adzhubhei, S. Schmidt, L. Peshkin, et al. A method and server for predicting damaging missense mutations. *Nat Methods,* 7(4):248-9, 2010.

[40] S. Henikoff and P.C. Ng. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res*, 31(13): 3812-3814, 2003

[41] J.M. Schwarz, C. Rodelsperger, M. Schuelke, and D. Seelow. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat. Methods,* 7(8):575-6, 2010.

[42] J.M. Schwarz, D.N. Cooper, M. Schuelke, and D. Seelow. MutationTaster2: mutation prediction for the deep sequencing age. *Nat. Methods,* 11(4):361-2, 2014.

[43] Y. Choi, and A.P. Chan. Provean web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics,* 31(16):2745-7, 2015.

[44] Y. Choi, G.E. Sims, S. Murphy, et al. Predicting the Functional Effect of Amino Acid Substitutions and Indels. 7(10):e46688, doi:10.1371/journal.pone.0046688, 2012.

[45] Bozeman MT: Variant Classification. In: SNP and Variation Suite Manual version 8.7.2, Copyright 2014, accessed Feb 2015.

[46] H. Thorvaldsdottir, J.T. Robinson, and J.P. Mesirov. Integrative Genomics Viewer (IGV): high performance genomics data visualisation and exploration. *Brief Bioinform,* 14(2):178-92, 2013.

[47] J.T. Robinson, H. Thorvaldsdottir, W. Winckler, et al. Integrative Genomics Viewer. *Nat Biotechnol,* 29(1):24-26, 2011.

[48] Metacore Version 6.24, Build 67895, Thomson Reuters, New York, USA

[49] S. Liu, L. Sheng, H. Miao, et al. SRA gene knockout protects against diet-induced obesity and improves glucose tolerance. *J. Biol Chem,* 289(19):13000-9, 2014.

[50] M.J. Duffy. The urokinase plasminogen activator system: role in malignancy. *Curr Pharm Des,* 10(1):39-49, 2004.

[51] A. Nykjaer, B. Moller, R.F. Todd 3rd, et al. Urokinase receptor. An activation antigen in human T lymphocytes. *J Immunol,* 152(2):505-16, 1994.

[52] J. Humphries, J.A. Gossage, B. Modarai, et al. Monocyte urokinase-type plasminogen activator up-regulation reduces thrombus size in a model of venous thrombosis. *J. Vasc Surg,* 50(5):1127-34, 2009.

[53] T. Chavakis, A.K. Willuweit, F. Lupu, et al. Release of soluble urokinase receptor from vascular cells. *Thromb Haemost,* 86(2):686-93, 2001.

[54] M. Baran, L.N. Mollers, S. Andersson, et al. Survivin is an essential mediator of arthritis interacting with urokinase signalling. *J. Cell Mol Med,* 13(9B):3997-808, 2009

[55] C. Chiueh, S. Lee, T. Andoh, and D. Murphy. Induction of antioxidative and antiapoptotic thioredoxin supports neuroprotective hypothesis of estrogen. *Endocrine,* 21(1):27-31, 2003.

[56] S.Y. Lee, T. Andoh, D.L. Murphy, and C.C. Chiueh. 17beta-estradiol activates ICI 182,780-sensitive estrogen receptors and cyclic GMP-dependent thioredoxin expression for neuroprotection. *FASEB J,* 17(8):947-8, 2003.

[57] A. Tedeschi, and S. Di Giovanni. The non-apoptotic role of p53 in neuronal biology: enlightening the dark side of the moon. *EMBO Rep,* 10(6):576-83, 2009.

[58] N. Ogawa, H. Dang, L. Kong, et al. Lymphocyte apoptosis and apoptosis-associated gene expression in Sjogren's Syndrome. *Arthritis Rheumatol,* 39(11):1875-85, 1996.

[59] W.M. Kuhtreiber, T. Hayashi, E.A. Dale, and D.L. Faustman. Central role of defective apoptosis in autoimmunity. *J. Mol. Endocrinol,* 31(3):373-99, 2003.

[60] A. Okuma, K. Hoshino, T. Ohba, et al. Enhanced apoptosis by disruption of the STAT3-IkappaB-zeta signaling pathway in epithelial cells induces Sjogren's syndrome-like autoimmune disease. *Immunity,* Mar 21;38(3):450-60, 2013.

[61] B.E. Bernstein, T.S. Mikkelsen, X. Xie, et al. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell,* 125(2):315-26, 2006.

[62] A. Barski, S. Cuddapah, K. Cui, et al. High-resolution profiling of histone methylations in the human genome. *Cell,* 129(4):823-37, 2007.

[63] N.D. Heintzman, R.K. Stuart, G. Hon, et al., Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet,* 39(3):311-8, 2007.

[64] R.C. Rao, and Y. Dou. Hijacked in cancer: the KMT2B (MLL) family of methyltransferases. *Nat. Rev. Cancer,*15(6):334-46, 2015.

[65] A.D.C. Paulussen, A.P.A. Stegmann, M.J. Blok, et al., MLL2 mutation spectrum in 45 patients with Kabuki syndrome. *Hum. Mutat,* 32(2):E2018-E25, 2011.

[66] A.W. Lindsley, H.M. Saal, T.A. Burrow, et al., Defects of B-cell terminal differentiation in patients with type-1 Kabuki syndrome. *J Allergy Clin Immunol,* 137(1):179-87, 2016.

[67] J.A. Daniel, and A. Nussenzweig. The AID-induced DNA damage response in chromatin. Mol. Cell 50(3):309-21, 2013.

[68] Y. Ichikawa, M. Bayeva, M. Ghanefar, et al., Disruption of ATP-binding cassette B8 in mice leads to cardiomyopathy through a decrease in mitochondrial iron export. *Proc Natl Acad Sci U S A,* 109:4152-7, 2012.

[69] A. Sturm, P. Cunningham, and M. Dean. The ABC transporter gene family of Daphnia pulex. BMC genomics 10:170, doi 10.1186/1471-2164-10-170, 2009.

[70] K. Katakura, H. Fujise, K. Takeda, et al. Overexpression of LaMDR2, a novel multidrug resistance ATP-binding cassette transporter, causes 5-fluorouracil resistance in Leishmania amazonensis. *FEBS Lett,* 561(1-3):207-12, 2004.

[71] Z. Wan, N. Shi, N. Wong, et al. Human parvovirus B19 causes cell cycle arrest of human erythroid progenitors via deregulation of the E2F family of transcription factors. *J. Clin Invest,* 120(10):3530-44, 2010.

[72] D. Longman, N. Hug, M. Keith, et al. DHX34 and NBAS form part of an autoregulatory NMD circuit that regulates endogenous RNA targets in human cells, zebrafish and Caenorhabditis elegans. *Nucleic Acids Res,* 41(17):8319-31.

[73] Y. Liu, N. Lu, B. Yuan, et al., The interaction between the helicase DHX33 and IPS-1 as a novel pathway to sense double-stranded RNA and RNA viruses in myeloid dendritic cells. *Cell Mol Immunol*, 11(1):49-57, 2014.

[74] J.M. Anaya, C. Ramirez-Santana, M.A. Alzate, et al. The autoimmune ecology. *Front Immunol*, 7:139, doi: 10.3389/fimmu.2016.00139.

Chapter 5

Genetics of Population Isolates: Comparisons of Genomic Variation in the Paisa Population with other Cohorts from Europe and South America

Abstract

Having identified candidate variants contributing to sporadic and familial autoimmunity, we now perform population genetics comparisons of the Paisa cohort with others part of the 1000 Genomes project. The purpose of this is to elaborate differences in genomic patterns that render population isolates, in particular the Paisa community, suitable for genetic epidemiological studies. Specifically, we obtained pairwise linkage disequilibrium values of all variants in the Paisa cohort and compared these to Finnish, British, Italian, Iberian Spanish, Utah, Peruvian and Puerto Rican populations of the 1000 Genomes database. We found that the Paisa cohort exhibits higher overall LD rankings compared to all populations except for the Peruvian cohort, based on the Wilcoxon Rank Sum Test. Also, when conducting Z proportion tests for the differences in rare variation percentages (i.e. proportion of variants <5% in the 1000 Genomes Project), the Paisa cohort significantly exceeded all cohorts except for the Puerto Ricans. Both these outcomes are reflective of demographic and population genetic history over time.

5.1 Introduction

A genetic isolate represents populations separated by geographical and/or cultural barriers that results in minimal admixing between individuals from external populations [1]. Such populations often arise through founder effects and/or population bottlenecks as a consequence of and along with these barriers. One of the most well known examples of this phenomenon, is that of the Paisa community, residing in Norther Colombia (including the departments of Antioquia, Risaralda, Caldas and Quindio) [1].

Although individuals of the Paisa region have mixed genetic lineages tracing back to the Spanish conquistador and Amerindian ancestry, this population has formed a genetic isolate over time. The founding population of the Paisa descended from genetic admixture from Iberian Spanish Conquistadors from the regions of Andalusia, Extramadura as well as early settlers from the Basque region and Sephardic Jewish individuals escaping religious persecution, dating back to the 15th century [1-3]. According to historical archives, the current population of the Paisa isolate are descended as a result of subsequent genetic admixture between these European colonizers with Amerindian females, supported by past admixture analyses. Over time, the regulations imposed by the Spanish monarchy as well as the geographic barriers within the Paisa region, resulted in the current Paisa individuals containing reduced Amerindian, and greater European ancestry, resulting in the world's currently largest known genetic isolate [1-3]. This is supported by previous genetic ancestral studies, suggesting that more than 94% of Y-chromosomes in the Antioquian population trace back to European lineages, whilst the remainder point towards African and Amerindian ancestry at proportions of 5% and 1% respectively [4]. Furthermore, more than 90% of the mtDNA gene pool pertains to Amerindian lineages. This sex-based bias in ancestry concurs with the historical fact that most of the non Native American migrants were males. In the descendent admixed individuals, there is still noticeable allelic structure of Amerindian ancestry within admixed mestizo individuals (in particular females) in the Paisa region [4].

Unlike other areas of South America, there is relatively little evidence of African ancestry (predominantly thought to have been derived from historical slave trade), amongst the Paisa population [5]. African ancestry is present amongst Paisa individuals. However, this may have been subsequently facilitated due to genetic exchange between Spain and North Africa, as both were under Arab rule during overlapping time periods. Data to support this this is from short tandem repeat (STR) microsatellite comparisons, suggesting that Andalusians have shorter genetic distances from Northwest Africans, compared to many other European as well as Non-European populations [5]. Furthermore, it was found in this particular study, that the overall genetic distanced, evaluated by Delaunar Triangulation and Neighbour-Joining algorithms was much closer between European and African as a whole, compared to those extracted from single populations compared between each continent [5]. Therefore, it can be interpreted that there has been previous instances of small, but noticeable gene flow from Northwest Africa to Andalusia, the latter of which (in turn) made substantial contributions to the ancestry of modern Paisas. To support this theory, must note that gradients can be observed between Spanish, Antioquians and North African individuals, possibly reflecting gene flow patterns during Arab occupation of Iberia [5]. These historical events, can also explain the prevalence of Sephardic Jewish ancestry in current Antioquians [4-5]. It was around this time, where expulsion of Jewish individuals was taking place in the Spanish community, many of whom migrated to North Africa and the Americas, thereby potentially explaining substantial degrees of allele sharing along regions of the Y chromosome between Sephardic Jews and current Antioquians, providing evidence that the former had evident influence of shaping the ancestry of the Paisa community [4-5].

In addition to the Andulusian and Sephardic Jewish ancestry, allele sharing of Y chromosome loci has also been observed amongst Antioquians with Basque and Catalan populations. Once again, this is consistent with documented history, with alleles observed exclusively in Basque individuals, exhibiting a frequent (>5%) prevalence in the Antioquian population. Thus, these patterns of admixture and immigration from diverse population groups, have all contributed to shaping the population genetics of the Paisa genetic isolate over time [1-5].

The demographic and genetic history of the Paisa population has substantial implication for genetic studies. There has been a recent shift from GWAS based approaches of the common-disease common-variant hypothesis, into developing and analytical strategies for identification of rare-variants as disease candidates [6-7]. This is because common variants can’t always account for the missing phenotypic heritability for a particular trait or disease. In addition to the compensation of low effect size by aggregating multiple rare variants into a single multi-site genotype, the power of these analyses are also enhanced through the use of genetic isolates such as the Paisa population [8]. This is due to their genetic homogeneity, resulting in a greater potential enrichment of rare variants as a subsequent effect of evolutionary forces such as founder effects, bottlenecks, genetic drift and unique admixture patterns [1, 6-8]. Furthermore, it has been previously hypothesised that isolated populations with admixture among individuals with divergent ancestry during the founder time are more likely to have higher levels of linkage disequilibrium, between a given set of variants [1]. Empirical evidence indicates, that this leads to a potentially increased power for mapping and identification of candidate disease genes [9-11].

In this study our objective is to determine the level of linkage disequilibrium and proportion of rare variation within the Paisa community, relative to other admixed populations that are either admixed (non-isolates) or arose from a unique founder effect (i.e. from a single ethnic group). In order to achieve this, various population groups (harboured within the continents ancestral origin of the current Paisa population, as determined by historical and genetic records) from the 1000 Genomes Project will be used. This is aimed to quantitatively elucidate the value of population isolates in association, linkage or any genetic epidemiological study.

5.2: Individuals and Methods

5.2.1 Samples Used and Preparation of VCF Files

Variant data for analyses was downloaded from the 1000 Genomes (Phase 3, hg19) Consortium [12]. These included populations from the Americas are: Colombians in Medellin (CLM), Peruvians in Lima (PEL), Puerto Ricans in Puerto Risco (PUR). Europeans populations include: Iberian Spanish (IBS), Utah individuals with Central European Ancestry (CEU), Finnish in Finland (FIN), British from England and Scotland (GBR) and Toscani in Italy (TSI). In total 794 individuals across 8 different populations in the 1000 Genomes database were analysed [12]. Note: The MXL (Mexicans in Los Angeles) are not part of this analysis, as their exact origins in Mexico, or their extent of genetic admixture with Non-Latino groups in the USA is not known. VCF files from the 1000 Genomes Consortium from each of these populations (for all autosomal chromosomes) were extracted and filtered for genotyping quality, using the information provided in the GATK best practices pipeline, as well as the gvf conversion tool. Variants with genotyping quality (Phred score) greater than or equal to 30 (representing >99.9% probability that the nucleotide base call is correct), were retained for downstream analysis [13-18]. Also, variants were filtered on the basis of whether they occur in exonic or intronic regions, with the aid of the Variant Effect Predictor. All variants occurring in coding or splice regions were included for analysis. Amongst the coding variants, only those that were non-synonymous were incorporated. Those in intronic regions were excluded [19].

5.2.2 Linkage Disequilibrium Analyses

LD was calculated for each population, using the PLINK software package, by Purcell et al. Due to the fact that the r-squared (r^2) measure of LD is frequency dependent, common variants were filtered (>5% individual frequency in the population) and subsequently used for this analysis [20]. Matrices were constructed, and the magnitudes of LD values between all pairs of variants within the matrix were compared (by ranking) between Paisa and other populations, using the Wilcoxon Rank Sum Test [21-24]. This gives the likelihood that any given LD value from the Paisa matrix will be greater than each of the other population groups included in this analysis. Pairwise Wilcoxon Rank Sum Tests were conducted for the 1000 Genomes Paisa individuals against each chosen population.

5.2.3 Rare Variant Proportions Analyses

Using the Pyhton programming language, the proportion of rare variants for each population, (defined in this case as those with less than 5% Frequency in the total 1000 Genomes population) was calculated [25]. This threshold for rare variation is based on the threshold set by Gibson et al, stating that more than 1/3 of exonic variants consist of minor allele frequencies of less than 5% in a given population [25-26]. Significance of the differences in rare variant proportions were evaluated by implementing a Z-prportion test, avalialable through the Epitools package in R version 3.2.5 [27].

5.3 Results

5.3.1 Linkage Disequilibrium Comparisons Between Populations

The significance of differences in the relative magnitude of LD between populations is determined from the P-value calculated from the Wilcoxon Rank Sum statistic, as in table 5.1. It was found that the rank sum statistic in the Paisa individuals was higher than all populations (Wilcoxon P-value = 0.0004705 against British population, P-value < $2.2e^{-16}$ for all other pairwise comparisons) except for the Peruvians in Lima. The Peruvians in Lima had a significantly greater rank sum statistic than all of the other populations, in each of the pairwise comparisons (P-value < 2.2e^{-16}) [21-24].

Cohort 1 (C1)	**Cohort 2 (C2)**	**Wilcoxon Rank Sum Test Statistic (C1)**	**Wilcoxon Rank Sum Test (C2)**	**P-value (Hypothesis: C1 > C2)**
CLM	CEU	2.197316e+14	2.043272e+14	<2.2 x 10-16
CLM	GBR	2.085402e+14	2.053634e+14	0.004705
CLM	TSI	2.2121e+14	1.948166e+14	< 2.2 x 10^{-16}
CLM	FIN	2.41472e+14	2.27722e+14	< 2.2 x 10^{-16}
CLM	IBS	1.756964e+14	1.557886e+14	< 2.2 x 10^{-16}
CLM	PUR	2.5938e+14	2.41186e+14	< 2.2 x 10^{-16}
PEL	CLM	1.43957e+14	1.342066e+14	< 2.2 x 10^{-16}
PEL	CEU	1.381798e+14	1.199198e+14	< 2.2 x 10^{-16}
PEL	GBR	1.29987e+14	1.206172e+14	< 2.2 x 10^{-16}
PEL	TSI	1.389102e+14	1.142724e+14	< 2.2 x 10^{-16}
PEL	IBS	1.031602e+14	9.14122e+13	< 2.2 x 10^{-16}
PEL	FIN	1.519298e+14	1.336676e+14	< 2.2 x 10^{-16}
PEL	PUR	1.536323e+14	1.415656e+14	< 2.2 x 10^{-16}

Table 5.1: Wilcoxon Rank Sum Test for 1000 Genomes Phase 3 variants from Populations compared against the Paisa (CLM) and Peruvian Cohorts. For each row, the populations compared in the pairwise tests are given in Columns 1 and 2 respectively. (CLM = Colombians in Medillin, CEU = Central Europeans in Utah, GBR = British in England and Scotland, TSI = Toscani in Italy, FIN = Finnish in Finland, IBS = Iberian Spanish in Spain PUR = Puerto Ricans in Puerto Rico, PEL = Peruvians in Lima). The Wilcoxon Rank Sum value for each cohort in every pairwise test is given. Significance of the difference in the magnitude of the LD values (from the matrices constructed by plink software) in terms of their relative rankings are given in the final column (P-value). This P-value is for the hypothesis that random LD values extracted from the first cohort are greater than those in the second cohort (C1 > C2).

In order to adjust for potential bias in LD comparisons as a consequence of unequal population sizes, results of this analysis are also given for 85 randomly individuals randomly extracted from each population. These are presented in table 5.2. This is to ensure that all populations are of equal size when applying the Wilcoxon Rank Sum Test on their LD values. The reason for applying this test on 85 individuals for each subset is that this was the number of individuals present in the PEL population. We see in this case that the Paisa population (CLM) again had a significantly greater rank statistics than all other populations (P-value <

$2.2e^{-16}$ for all pairwise comparisons) except for the PEL cohort. In turn, the rank sum statistics for the PEL cohort were greater than all other cohorts compared against (P-value < $2.2e^{-16}$).

Cohort 1 (C1)	Cohort 2 (C2)	Wilcoxon Rank Sum Test Statistic (C1)	Wilcoxon Rank Sum Test (C2)	P-value (Hypothesis: C1 > C2)
CLM	CEU	1.872024e+14	1.808598e+14	$< 2.2 \times 10^{-16}$
CLM	GBR	1.877942e+14	1.829036e+14	$< 2.2 \times 10^{-16}$
CLM	TSI	1.86989e+14	1.804066e+14	$< 2.2 \times 10^{-16}$
CLM	FIN	2.07647e+14	2.024176e+14	$< 2.2 \times 10^{-16}$
CLM	IBS	1.538702e+14	1.48819e+14	$< 2.2 \times 10^{-16}$
CLM	PUR	2.177824e+14	2.155802e+14	$< 2.2 \times 10^{-16}$
PEL	CLM	1.30064e+14	1.240102e+14	$< 2.2 \times 10^{-16}$
PEL	CEU	1.228502e+14	1.177616e+14	$< 2.2 \times 10^{-16}$
PEL	GBR	1.23484e+14	1.19904e+14	$< 2.2 \times 10^{-16}$
PEL	TSI	1.227116e+14	1.174624e+14	$< 2.2 \times 10^{-16}$
PEL	IBS	1.009778e+14	9.68946e+13	$< 2.2 \times 10^{-16}$
PEL	FIN	1.362702e+14	1.318086e+14	$< 2.2 \times 10^{-16}$
PEL	PUR	1.529494e+14	1.403666e+14	$< 2.2 \times 10^{-16}$

Table 5.2: Wilcoxon Rank Sum Test for 1000 Genomes Phase 3 variants from Populations compared against the Paisa (CLM) and Peruvian Cohorts, with each set having an equal number (85) of individuals. For each row, the populations compared in the pairwise tests are given in Columns 1 and 2 respectively. (CLM = Colombians in Medillin, CEU = Central Europeans in Utah, GBR = British in England and Scotland, TSI = Toscani in Italy, FIN = Finnish in Finland, IBS = Iberian Spanish in Spain PUR = Puerto Ricans in Puerto Rico, PEL = Peruvians in Lima). The Wilcoxon Rank Sum value for each cohort in every pairwise test is given. Significance of the difference in the magnitude of the LD values (from the matrices constructed by plink software) in terms of their relative rankings are given in the final column (P-value). This P-value is for the hypothesis that random LD values extracted from the first cohort are greater than those in the second cohort (C1 > C2).

5.3.2 Rare Variant Proportion Comparisons Between Populations

Given the shift in genetic studies for study of rare variants as mentioned above, we have also obtained results for the proportion of rare variants (<%5 frequency in the 1000 Genomes Project), that are present in each of the ascertained populations. Results of all rare-variant proportions for each population are given in table 5.3 [26-27]. Out of all the populations scanned for rare variants, the CLM population had the highest proportion of rare variants and the pairwise Z scores were higher than all of the groups analysed (P-value <0.05) with the exception of the Puerto Ricans. These results corroborate well with the demographic structure and history for each population group.

Cohort 1	**Cohort 2**	**Z –Proportion Statistic**	**P-value (C1 > C2)**
CLM	CEU	55.601	< 0.0002
CLM	GBR	78.75	< 0.0002
CLM	PUR	-23.166	0.0231
CLM	FIN	95.001	< 0.0002
CLM	IBS	37.152	0.0017
CLM	TSI	47.18	< 0.0002

Table 5.3: Rare variant proportion comparisons between the Paisa and Peruvian cohorts with other populations of the 1000 Genomes project phase 3. Identity of cohort for comparison is given in columns 1 and 2. Differences in proportions are given by a Z-value, followed by the significance of this score in the 3rd and 4th columns respectively.

5.4 Discussion

Our results suggest that Paisa (CLM) individuals have a higher proportion of rare variants. Also, they have a greater probability of harbouring mutations with higher ranked LD values relative to most other 1000 Genomes populations in these analyses (when LD values in each matrix are chosen at random as per the protocol of the Wilcoxon Rank Sum Test). This supports the theory stated by Arcos-Burgos and Muenke, that genetic isolates descended from populations undergoing admixture during the founder time are more likely to exhibit higher levels of LD, compared to populations with unique founder effects from non-divergent population ethnic groups [1]. Previously sampled isolates also exhibited similar patterns, whereby LDU map lengths were shorter in younger isolates with divergent founders (i.e. Antioquian cohort) compared to outbred populations or older isolated cohorts, from Europe [28]. As previously mentioned, this is the case with the Paisa, as their demographic history is shaped by admixture between Iberian as well as Amerindian founders. Also founding time in the Paisa (~ 450 years) is much later than the other analysed cohorts in this study

When comparing the LD values via the Wilcoxon Rank Sum Test of the Paisa, with all 7 other population groups, we find that results show correlation with population history, according to the theory stated by Muenke and Arcos-Burgos [1]. Firstly, the CEU population is originally descended from a group of Mormons in the 1840s, arising from a relatively small, and substantially isolated population [29-30]. These individuals were almost exclusively of Northern and Western European Ancestry. This is determined by previous studies, indicating significant correlations in SNP frequencies between Irish populations and HapMap individuals. The authors from this study also found that the CEU had the highest LD (r2) correlation with the Irish population (>0.95), relative to those from Finland and the UK. Furthermore, CEU individuals and those from Southern England have been shown to be almost indistinguishable in terms of Fst and PCA results. Hence, the current CEU population did not descend from admixture between largely divergent founders, unlike the Paisa community [31-32]. Many of these individuals, related to the Mormons, many of whom in turn are descended from single common ancestors.

Like the CEU, genetic evidence also points towards a unique founder effect amongst the TSI population. Although rulers from Rome and other Italian regions colonized Tuscany over time, Etruscan DNA dominates the majority of this region's genetic ancestry, according to Y chromosome haplotype studies. Ancestral studies of mtDNA lineages support the hypothesis that the Etruscans (one of the first groups to have been established in Italy) population development can be traced back to local origins [33]. Furthermore, Italians share many ancestral genetic components amongst individuals across the country (albeit in different proportions) [34-36]. In fact, Haplogroup R1b on the Y-chromosome is present in the highest frequency, relative to other haplotypes in majority of the Italian population. Amongst all populations analysed, Tuscan regions have the highest proportion of this haplogroup [34-36]. Genetic maps constructed for Europe indicate that Italy has substantially distinct genomic components compared to the remainder of Europe [37]. This phenomenon occurred, mainly due to large alpine mountain chains, as well as the Mediterranean Sea, acting as geographic barriers to restricting migration and gene flow (from outside the country) for many generations since the establishment of this population. Thus, it has been previously concurred among geneticists that historical external migrations into the Italy, had little effect (apart from ancient Greek and Roman migrations) in causing major alterations of the genetic composition of Italian populations [38-40].

As well as geographic barriers, the Black Plague, striking Tuscany twice in medieval history also led to population bottlenecks. Consequently, this led population underwent multiple founder effects [41-42]. Any traces of genetic admixture between Etruscans and other population groups within Italy, would have been depleted. All these complex historical and geographic factors point to limited ethnic divergence amongst the founding ancestors of the current TSI population [1, 41-42]. Although, modern migration patterns have facilitated greater admixture between Tuscan and other Italian populations, this mainly results in gene flow of haplotypes that are already shared amongst many Italian groups [34-36]. Also, any influences of LD due to cross-migration between these populations would have been reduced from chromosome recombination, over many generations. Hence, (as previously mentioned) this founder effect is not the same as the Paisa, whereby initial admixture led to overrepresentation of newly introduced genetic variation (in the founder time) over successive generations, and LD effects are still present and will persist for many generations, before they are depleted by allelic recombination [1-2, 4]. These phenomena may all account for the substantially lower LD ranks in the TSI (and CEU) cohorts compared to the Paisa individuals.

Having accounted for these differences, one must note, that Finnish as well as British and Iberian Spanish populations also descend from ancestors of diverse ethnic groups. In other words, historical data indicates that these cohorts arose through divergent founder lineages [43]. In the case of the Finns, the dual origin model represents their emergence from 2 different geographical locations (Uralic speakers from Ladoga Lake as well as migrants arriving from south of the Gulf of Finland [43]. These 2 groups together formed a small founder population, whose growth was not rapid due to war, famine and disease. This was followed by expansion (of the descendent population) into the North, West and East of the Country.

In addition to the Finnish population, genetic isolation has also seemingly played a major role in British populations as reflected in their fine-scale population genetic structure [44]. Also, the regions of Orkney and Kent have exhibited different levels of ancestry arising from Scandinavian and Anglo-Saxon migrations into the British Isles [44].

Lastly, in the case of the Iberian Spanish populous, colonization also had a substantial effect on their genetic structure, during the Arab conquest of Spain as previously mentioned [5]. However (particularly in the case of the Finns and British cohort), these initial admixture events during the establishment of these populations occurred many generations ago (> 1000 years). Therefore, the effects of LD would once again have been reduced over many generations as a subsequent effect of genetic recombination over time [1]. Moreover, as the IBS cohort was collected from all autonomous communities of Spain, one can assume that these individuals descend from populations established from the period of Arabic rule and thereafter, mainly to due religious and ethnic-based autocracy, resulting in many population bottlenecks and subsequent founder effects from multiple conflicts. This would have created a relatively uniform genetic architecture amongst Spanish populations, as opposed to the admixed/multi-founder effect of the Paisa [45]. As already mentioned, whatever limited admixture took place would have been reduced by recombination. Thus, once again in accordance with Arcos-Burgos and Muenke, that provides an explanation for the significantly higher LD ranks of the Paisa over the FIN, GBR and IBS cohorts [1].

Thus our results had greater LD rank sum statistics in the Paisa over each of the analysed European populations, coinciding strongly with spatial and temporal demographics in each case. However, the Paisa population, had overall significantly lower rank sum statistics compared to the PEL cohort, once again pointing to the latter’s unique geographic and genetic history. The Peruvians in Lima follow similar historical demographics to the Paisa community. They too have experienced the effects of Iberian admixture with Native American populations [46]. Over time, there have been high levels of migration and gene flow within Lima from various ethnicities. However, previous studies involving ancestry specific PCA (ASPCA) are more closely related to the Aymara and Quechua populations, (who form the population majority among Native American groups in the Andes Highlands), than other Native American Andean groups such as the Huilliche, Inga and Yaghan, as well as those in the Southern Amazonian basin [46]. The Aymara and Quechua populations in turn seemingly had high effective population sizes and gene flow between themselves, suggested by Reynolds’ genetic distance and Median-Joining (MJ) networks identifying their shared haplotype lineages [46-47]. Nevertheless, according to the allelic structure and differentiation identified in the aforementioned ASPCA, the Quechua and Aymara still seemingly experienced genetic isolation from other Native American groups, due to geographic obstacles of the Andes [46].

Following the Pre-Columbian era, Peru also experienced a multi-founder effect, from Iberian colonization, similar to the Paisa community. Nonetheless, the last major migration and admixture pulse from the Iberian Peninsula and Native Americans (who in turn established the founder population) was 9 generations ago, according to tract length distribution studies. This is 2 generations later than the Paisa community [46]. This can account for the higher LD ranking levels in the Peruvians over the Paisa and the other 1000 Genomes American and European populations part of our study, concurrent with the multi-founder effect theory in population isolates [1].

Not only is the Paisa isolate useful for their genetic studies due to the above mentioned linkage disequilibrium evidence, but they also harbour substantial rare variation, thereby aiding disease studies [1]. Given the migratory, isolation and admixture patterns across all populations, it is of no surprise that the proportion of rare variation in the Paisa significantly exceeds all of the European cohorts, and the Peruvians in Lima [1]. In addition, when comparing the Paisa and Puerto Rican populations, one must consider that the latter are one of the world's most admixed populations. This is not only due to contribution of genetic variation from Europeans and Native Americans, but also the introduction of a large colonial labour from Africa, during the invasion of the Spanish empire. 80% of Africans who migrated to the Americas for this purpose disembarked at Puerto Rico and other Caribbean Islands. However, these divergent groups did not form a genetic isolate, (potentially leading to lower observed LD magnitudes as per Wilcoxon scores). As can be inferred from the current genomic ancestry of Puerto Rico genetic flow across the island was strongly maintained. Mitochondrial and Y-chromosome DNA analysis showed that more than 15% of genetic ancestry could be traced to African and Native American individuals respectively [48-49]. These proportions along the Y-chromosome are substantially greater than the Paisa isolate [2-4]. Also mtDNA sampled from women part of the Puerto Rican diaspora contain > 80% of ancestry markers from African, Native American and European populations. The maximum percentage of women with ancestral markers derived only from one of these ethnicities was 5% [50]. Additionally Puerto Rico has experienced continuous immigrant influx since its initial occupation from various European nations such as France, Ireland, Italy, Germany and Scotland as part of the Royal Decree of Graces, 1815 [51]. This was followed by further European migration post WW2 and influx from the United States during the country’s strong economic development in the 20th century. All these factors promote the introduction of new genetic variation, thereby resulting in an increased likelihood of identifying rare variants.

5.5: Conclusion

Thus, it is evident from the above LD and rare-variant analyses that individuals recruited for DNA sampling from the Paisa community, offer a powerful analytical tool for genetic epidemiology. Potentially, these results suggest that within this cohort, there is a greater likelihood of identifying rare variants harboured in regions with higher levels of relative LD magnitude compared to most European and American populations in the 1000 Genomes database. It must be noted that the 1000 Genomes population doesn’t represent all variation across both these continents. Nevertheless, these results still provide strong quantitative evidence of the value of the Paisa isolate in genetic analysis due to their multi-founder effect.

References

[1] M. Arcos-Burgos, and M. Muenke. Genetics of Population Isolates. *Clin Genet,* 61(4):233-247, 2002.

[2] M.L. Bravo, C.Y. Valenzuela, and O.M. Arcos-Burgos. Polymorphisms and phyletic relationships of the Paisa community from Antioquia (Colombia). *Gene Geogr,* 10(1):11-7, 1996.

[3] N.R. Mesa, M.C. Mondragón, I.D. Soto, et al. Autosomal, mtDNA, and Y-Chromosome Diversity in Amerinds: Pre- and Post-Columbian Patterns of Gene Flow in South America. *Am. J. Hum. Genet,* 67(5):1277-1286, 2000.

[4] L.G. Carvajal-Carmona, I.D. Soto, N. Pineda, et al. Strong Amerind/White Sex Bias and a Possible Sephardic Contribution among the Founders of a Population in Northwest Colombia. *Am. J. Hum. Genet*, 67(5):1287-1295, 2000.

[5] E. Bosch, F. Calafell, A. Perez-Lezaun, et al. Genetic structure of northwest Africa revealed by STR analysis. *Eur J Hum Genet*, 8(5):360-6, 2000.

[6] D.J. Liu, and S.M. Leal. A Novel Adaptive Method for the Analysis of Next-Generation Sequencing Data to Detect Complex Trait Associations with Rare Variants Due to Gene Main Effects and Interactions. *PLoS Genet*, 6:e1001156. https://doi.org/1.1371/journal.pgen.1001156, 2010.

[7] S. Lee, G.R., Abecassis, M. Boehnke, and X. Lin. Rare-Variant Association Analysis: Study Designs and Statistical Tests. *Am. J. Hum. Genet*, 95(1):5-23, 2014.

[8] L. Trauta, T. Hautala, S. Hämäläinen, et al. Enrichment of rare-variants in population isolates: single AICDA mutation responsible for hyper-IGM syndrome type 2 in Finland. *Eur J Hum Genet*, 24(10):1473-1478.

[9] J.I. Vélez, F. Lopera, D. Supelveda-Falla, et al. APOE*E2 allele delays age of onset in *PSEN1* E280A Alzheimer's Disease. *Mol Psychiatry*, 21(7):916-924.

[10] J.I. Vélez, C.A. Mastronardi, H.R. Patel, et al. Conference: 65th Annual Meeting of the American Society of Human Genetics Annual Meeting, Baltimore, MD, US, October, 6-10, 2015, DOI:10.13410/RG.2.2.30663.50085.

[11] J.I. Vélez, F. Lopera, H.R. Patel, et al. Mutations modifying sporadic Alzheimer's disease age of onset. *Am J Med Genet B Neuropsychiatr Genet*, 171(8):1116-1130, 2016.

[12] The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526(7571):68-74, 2015.

[13] A. McKenna, M. Hanna, E. Banks et al. The Genome Analysis Toolkit: a Map-Reduce framework for analysing next-generation DNA sequencing. *Genome Res*, 20(9):1297-1303, 2010.

[14] M. DePristo, E. Banks, R. Poplin, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet,* 43(5):491-8, 2011.

[15] G.A. Van der Auwera, M. Carneiro, C. Hartl, et al. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices. *Curr Prot Bioinformatics*, 43:11.10.1-33. Doi: 10.1002/0471250953.bi1110s43, 2013.

[16] B. Kennedy, Z. Kroennburg, H. Hu, et al. Using VAAST to Idnetify Disease-Associated Variants in Next-Generation Sequencing Data. *Curr Prot Hum Genet,* 81: 6.14.1-25. doi: 10.1002/0471142905.hg0614s81, 2014.

[17] M. Yandell, C. Huff, H. Hu, et al. A probabilistic disease-gene finder for personal genomes. *Genome Res,* 21(9):1529-42, 2011.

[18] H. Hu, J.C. Roach, H. Coon, et al. A Unified Test of Linkage Analysis and Rare-Variant Association for Pedigree Sequencing Data. *Nat Biotechnol*, 32(7):663-669, 2014.

[19] W. McLaren, L. Gil, S.E. Hunt et al. The Ensembl Variant Effect Predictor. *Genome Biol,* 17:122 doi:10.1186/s13059-016-0974-4, 2016.

[20] S. Purcell, B. Neale, K. Todd-Brown et al. PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am. J. Hum. Genet,* 81(3):559-75, 2007.

[21] H.B. Mann, and D.R. Whitney. On a Test of Whether one of Two Random Variables is Stochastically Larger. *Ann. Math. Stat*, 18(1):50-60, 1947.

[22] W.J. Conover and R.L. Iman. Rank Transformations as a Bridge Between Parametric and Nonparametric Statistics. *Am. Stat*, 35(3):124-29, 1981.

[23] D.F. Bauer. Constructing confidence sets using rank statistics. *JASA,* 67(339):687-90, 1972.

[24] M.A. Hollander, and D.A. Wolfe. *Nonparametric Statistical Methods*: 2nd Ed. New York: John Wiley & Sons, 1999.

[25] Python Software Foundation Python Language Reference, Version 2.7. Available at http://www.python.org

[26] G. Gibson. Rare and common variants: twenty arguments. *Nat Rev Genet,* 13(2):135-45, 2012.

[27] R.C. Sprinthall. *Basic Statistical Analysis*: 9th Ed. Pearson Education, 2011.

[28] S. Service, J. DeYoung, M. Karayiorgou et al. Magnitude and distribution of limkage disequilibrium in population isolates and implications for genome-wide association studies. *Nat Genet,* 38(5):556-60, 2006.

[29] T. McLellan, L.B. Jorde, and Skolnick, M.H. Genetic distances between Utah Mormons and related populations. *Am. J. Hum Genet*, 36(4):836-57, 1984.

[30] The International HapMap Consortium. The International HapMap Project. *Nature*, 426(6968):789-96, 2003.

[31] C.T. O’Dushlaine, D. Morris, V. Moskvina, et al. (2010). Population Structure and Genome-wide patterns of variation in Ireland and Britain. *Eur J Hum Genet,* 18(11):1248-54, 2010.

[32] C.T. O’Dulshaine, C. Dolan, M.E. Weale, et al. An assessment of the Irish population for large-scale genetic mapping studies involving epilepsy and other complex diseases. *Eur J Hum Genet*, 16(2): 176-83, 2008.

[33] S. Ghirotto, F. Tassi, E. Fumagilli, et al. Origins and Evolution of Etruscans' mtDNA. *PLoS ONE*, *8*, e55519. https://doi.org/10.1371/journal.pone.0055519, 2013.

[34] G Fiorito, C. Di Gaetano, S. Guarrera, et al. The Italian Genome reflects the history of Europe and the Mediterranean basin. *Eur J Hum Genet*, 24(7):1056-62, 2016.

[35] C. Di Gaetano, F. Voglino, S. Guarrera, et al. An Overview of the Genetic Structure within the Italian Population from Genome-Wide Data. *PLoS ONE*, 7:e43759, https://doi.org/10.1371/journal.pone.0043759, 2012.

[36] F. Brisighelli, V. Álvarez-Iglesias, M. Fondevila, et al. Uniparental Markers of Contemporary Italian Population Details on Its Pre-Roman Heritage. *PLoS ONE*, *7*:e50794. doi:10.1371/journal.pone.0050794, 2012.

[37] O. Lao, T.T. Lu, M. Nothnagel M., et al. Correlation between Genetic and Geographic Structure in Europe. *Curr Biol*, 18(16):1241-8, 2008.

[38] P. Ralph, and G. Coop. The geography of recent genetic ancestry across Europe. *PLoS Biol*, 11: e1001555 doi:10.1371/journal.pbio.1001555, 2013.

[39] C. Di Gaetano, N. Cerutti, F. Crobu, et al. Differential Greek and northern African migrations to Sicily are supported by genetic evidence from the Y chromosome. *Eur J Hum Genet*, 17(1):91-9, 2008

[40] S. Tofanelli, F. Brisighelli, P. Anagnostou, et al. The Greeks in the West: Genetic signatures of Hellinic colonization in southern Italy and Sicily. *Eur J Jum Genet*, 24(3):429-36, 2016.

[41] O.J. Bendedictow. The Black Death, 1346-53: The Complete History (Boydell and Brewer), 2004.

[42] C.M. Cipolla. Fighting the Plague in Seventeenth Century Italy (Madison: University of Wisconsin Press), 1981.

[43] S.R. Wang, V. Agarwala, J. Flannick, et al. Simulation of Finnish Population History, Guided by Empirical Genetic Data, to Assess Power of Rare-Variant Tests in Finland. *Am. J. Hum. Genet*, 94(5):710-20, 2014.

[44] S. Leslie, B. Winnie, G. Hellenthal, et al. The fine-scale genetic structure of the British Population. *Nature*, 519(7543):309-14, 2015.

[45] S.M. Adams, E. Bosch, P.L. Barlesque et al. The Genetic Legacy of Religious Diversity and Intolerance: Paternal Lineages of Christians, Jews and Muslims in the Iberian Peninsula. *Am. J. Hum. Genet*, 83(6):725-36, 2008.

[46] J.R. Homburger, A. Moreno-Estrada, Gignoux, et al. Genomic Insights into the Ancestry and Demographic History of South America. PLos Genet. 11, e1005602. https://doi.org/10.1371/journal.pgen.1005602, 2015.

[47] J.R. Sandoval, D.R. Lacerda, M.S.A. Jota, et al. The Genetic History of Indigenous Populations of the Peruvian and Bolivian Altipano: The Legacy of the Uros. PLoS ONE. *8,* e73006. https://doi.or/10.1371/journal.pone.0073006, 2013.

[48] H. Tang, S. Choudry, R. Mei, et al. Recent genetic selection in the ancestral admixture of Puerto Ricans. *Am. J. Hum. Genet*, 81(3):626-33, 2007.

[49] J.C. Martinéz -Cruzando. The Use of Mitochondrial DNA to Discover Pre-Columbian Migrations to the Caribbean: Results for Puerto Rico and Expectations for the Dominican Republic. KACIKE: *Journal of Caribbean Amerindian History and Anthropology,* ISSN 1562-5028, 2002.

[50] C. Bonilla, M.D. Shriver, E.J. Parra, et al. Ancestral proportions and their association with skin pigmentation and mineral bone density in Puerto Rican women from New York City. *Hum Genet*, 115(1):57-68, 2004.

[51] ArchivesPuertoRico.com 2013. About the Royal Decree of Graces.

Chapter 6: Discussion

6.1: Success of Sporadic and Familial Studies

Rare variants, enriched in sporadic cases and families from a well known genetic isolate, potentially sharing similar immune mediated pathways may be responsible for contributing to observed, extreme autoimmune phenotypes. For many years, given the modest outcomes of the CDCV hypothesis, in accounting for the heritability of complex diseases (especially autoimmunity and MAS), identification of rare variants has been of great interest in genetic analysis [1-3]. This has been facilitated, not only by the development in modern sequencing technologies [4], but also the emergence of new linkage and rare-variant association analysis algorithms [5-6]. The power of these statistical studies has been further boosted, by algorithms combining linkage analysis and association tests into a unified framework, with incorporation of de novo mutations [5-6].

In polyautoimmunity and Multiple Autoimmune Syndrome, our findings of the *LRP1* and *SRA1* genes [7-8] are of special relevance, as they can impact upon early diagnosis and potential development of targeted therapies. Such success can only be achieved through the use of genetic isolates such as the Paisa community, due to their genomic and environmental homogeneity, as previously mentioned [9-10]. This not only benefits Paisa patients, but also other cohorts around the world. This is because the same, potentially pathogenic mutations, overrepresented in the Paisa cohorts may also be disease causing in other populations. However, if they are prevalent in populations with diverse ethnic and genomic backgrounds, they may not be easily detectable due to stratification.

6.1.1: Similarities of Results between Sporadic and Familial Autoimmunity Analyses

Although the list of candidate genes, in both the sporadic case-control and familial linkage analyses yielded different lists of candidate genes, these 2 studies still potentially reveal similar findings on a pathophysiological and functional level. Evidence from pathway and network analysis 3.3 suggest sharing of biological networks and processes amongst identified genes from case-control and linkage analyses from studies of the sporadic and familial cases respectively [7-8]. This was identified via the implementation of the Metacore-based pathway and network analysis, involving data mining from the GeneGo ontology database. Specifically, both cases involved the PLAUR (Plasminogen Urokinase Receptor) protein. As previously mentioned, this protein was shown to be a part of a series of nodal interactions with LRP1 and SRA1 in the candidate gene lists from the sporadic and familial cases respectively, associated to processes that are relevant to autoimmune pathophysiology [7-8]. Furthermore, the processes shared between these nodes in both cases are all similar in nature, as they are centred around the occurrence and/or regulation of apoptotic activity (phagocytosis, negative regulation of apoptosis, endopeptidase activity etc) [7-8, 11-17]. Additionally, given the interactions of both SRA1 and LRP1 with PLAUR [7-8], all 3 of these genes can potentially regulate inflammatory responses by IL-1 and IL-10 signalling pathways, uPAR expression on monocytes and phospholipase A2 activity, (leading to IFN-gamma activation) respectively [7-8, 17-24]. Therefore, understanding of the autoimmune tautology and pathophysiology for phenotypes observed in this cohort is enhanced. It’s true that our studies (involving both linkage and rare-variant association analyses) did not share any overlapping genes, between sporadics and familial cases [7-8]. However they successfully identified mutations that may have important similarities in terms of their biological pathways, networks and mechanisms [25-26], in the context of autoimmunity, polyautoimmunity and MAS.

Nevertheless there is sufficient evidence, supporting the identification of autoimmune genes in these studies, particularly through the combine linkage and association algorithm, implemented upon the pedigrees. This evidence could be strengthened by replication of results in new, untested populations. In addition, the findings could also develop a platform for targeted therapies that target the *SRA1*, *MLL4*, *DHX34*, *ABCB8* and *PLAUR* genes. This will depend upon findings from functional validation, evaluating these mutations as disease causing variants. Assuming that positive results are obtained from such experiments, therapeutic agents may be able to be implemented. For example, gene therapy, targeting specific mutations within these listed genes may be able to successfully initiate growth, development and replication of cells not carrying any of these variants in their DNA sequence. This may ensure that the proteins encoded by these genes are retaining their required homeostatic function. Another option is the development of antigens or inhibiting agents that can accelerate or reduce the apoptotic, endopeptidase or Plasminogen Urokinase activity of these proteins (the latter of which mediates IL1 responses) [7-8, 11-24], depending on whether they are loss or gain of function mutations. Such therapeutics may take time to develop, but these findings provide a strong platform for molecular biologists to understand the approach that needs to be taken limiting the effects of these autoimmune phenotypes.

6.2 Future Directions

The rigorous statistical analysis was successful in combining the algorithmic power of rare-variant association tests and linkage studies to successfully identify candidate autoimmune genes. However, the next step is to implement future strategies and directions to build upon the current findings, developed from our current framework. This can improve the clinical applications of this research, based on the findings we have thus far.

6.2.1 Development of pVAAST Algorithm to Test for Multiple Loci

The advantages of pVAAST have previously been mentioned, due to its ability to incorporate linkage and association analyses (robust to both common and rare variants), test for de novo mutations and include compound heterozygotes as part of the analyses [5-6]. The algorithm provides strong evidence for aforementioned genetic variants contributing to autoimmune phenotypes, but it is not applicable to test likelihood ratios and allele frequency differences of multiple loci simultaneously [5-6, 27-29]. Hence, further developments may be needed to introduce this capability, in order to further understand the importance of gene interactions in these diseases. This can be implemented by the incorporation of two-trait-locus models into the linkage and association algorithms of pVAAST for dominant, recessive and co-dominant modes of inheritance [27-29]. That means, the LOD score calculation, computing the probability of the genotypes and phenotypes at the current and latent loci, conditional on the penetrance/genotype disease probability and allele frequencies will be calculated for 2 candidate loci simultaneously [27-29]. In brief, this works by conditional probabilities of genotypes and phenotypes using the allelic inheritance vector approach. These amendments can be programmed into the source code of the software package. To begin, the loci identified from familial and sporadic cases can be tested in this algorithm on the same dataset, used in our studies thus far.

6.2.2 Meta Analysis of Previous Studies Combined with the Paisa Cohort:

Now that results from well-developed statistical algorithms are obtained, they can then be combined with our previous studies, to extract additional statistical information about potential causative genes in autoimmunity. This further increases the statistical power in identifying and validating disease risk genes. The best option for this will be Genome Scan Meta Analysis (GSMA), combining results from the Paisa dataset, along with those of previous studies in other cohorts [30-31]. GSMA works by dividing the genome into equal bins, and the most significant mutations in each bin (determined by P-value or LOD score) are listed, and the bins are ranked according to these results. Ranks of the variants in each bin are combined across studies to get the sum rank (SR), the significance of which is determined by permutation [30-31]. This gives the probability of observing the same sum rank in descending order across the permuted replicates. To adjust for differences in sample numbers across studies, ranks can be weighted according to the number of cases and families included across all individual analyses [30-31]. The advantage of this algorithm is that candidate disease genes are not determined by the absolute value of the LOD score, but instead are prioritised by ranking and relative significance across all studies [30-31]. It will also adjust for heterogeneity introduced by study designs or population diversity, by calculation of the Q-adjusted statistic, that determines the sum of squared differences between the individual and average bin ranks across each study. Therefore, studies can be included in the meta-analyses that don’t necessarily use the same statistical algorithms [29-31]. This can be applied to both single locus and gene interaction based composite likelihood ratio test as per the combined linkage-association analysis framework (of pVAAST) [5-6].

6.2.3 Alternative Approaches to Association Analyses: Pooling/Bootstrap GWAS and Versatile Gene-Based Association

The strategies suggested above for future analyses will be successful, only when additional multigenerational pedigrees are available. Extensive collection of DNA samples from families of the Paisa isolate has led to successful studies thus far. Nevertheless, one cannot guarantee that such well-phenotyped pedigrees with extensive genealogical records will be available when conducting this research amongst other populations. Thus alternative strategies for association tests need to be considered, especially in cases of small sample sizes. We have previously referred to gene-based variant collapsing statistical calculations, in order to identify disease-causing mutations. However, these methods are primarily designed for rare variants [2, 26]. When analysing common variants, the elevated allele frequencies may lead to spurious associations, and without the availability of pedigrees, linkage analyses are not an option. Therefore, increasing the analytical power of identifying such variants with relatively few samples becomes a challenge. One alternative is that suggested by Velez et al [32], combining pooling and bootstrapping methods with GWAS (pbGWAS)-based case vs. control studies.

In brief, the algorithm involves multiple rounds of bootstrap based resampling of subsetted data. The test statistic is conditional on the difference in allelic frequencies between cases and controls, as well as the variance of these differences, across each pooling step. P-values are then obtained for evaluating the significance of this test statistic, based on the fact that the value will follow a Chi-Square distribution under the null hypothesis. P-values from all stages of pooling are subsequently combined, using Stouffer’s method, which enables P-value mapping, according to a standard normal distribution, as these values will follow this pattern under the null hypothesis [32-34]. The key variables in performing this calculation include: the weight and quantile of the standard normal distribution (obtained from corresponding effect size estimates and individual P-values respectively calculated for all individual test statistics, derived from case-control allele frequency differences across each pooling simulation) [32-34]. In addition, assumption of independence between simulated pairs (of case and control pools respectively) doesn’t always hold. Thus, the parameter of the correlation coefficient is introduced, quantifying the degree of dependence in allelic frequency distributions between the pairs of pooled samples. This value can be estimated (if unknown), based upon the inverse value, variance vectors and quantiles of the standard normal distribution, corresponding to calculated P-values across all pooled case-control subsamples [32-34]. These formulas incorporated in this algorithm, substantially increase its overall power, compared with other commonly used methods [32].

The power of this mathematical structure can be attributed to its weighting approach. That means, during combination of P-values, each study is weighted according to its overall power, obtained from the square standard error of the effect size estimate from each individually pooled subset [32-34]. Empirical observations indicate that when combining studies of both equal and unequal sample sizes, weighted approach of Stouffer’s collaborated P-value are more powerful than algorithms applying unweighted statistics [32-34]. This can account for bias in sample quantities, as effect size info is based on relative differences in magnitude, rather than P-values. The variance of this difference, is a denominator in effect size calculation. This can potentially add weight to studies with larger pools of data as well as those with greater effect size, thereby offsetting sample size bias [32]. To some extent, this can also reduce publication bias, whereby studies with P-values < 0.05 are more likely to be published than those with P-values > 0.05.

Another issue to address when combining P-values is their asymmetry across multiple studies. Certain methods such as Fisher’s Test are more sensitive to smaller than larger P-values, further exacerbated (once again) by differences in sample sizes [33-34]. It must also be noted that high P-values close to 1 are likely to favour the opposite alternate hypothesis. For this reason, even though high and low P-values differ by a substantial margin, they can both arrive at the same conclusion in rejection of the null hypothesis, when separate subsets of samples are combined. [33-34] Once again, this is neutralised by Stouffer’s weighted Z transformation, as weighting procedures are more reliant on effect sizes, as mentioned earlier [33].

Given the mathematical properties of the weighted Z transform method in reducing sources of bias, it is of no surprise that this algorithm generated the best correlation between observed individual and combined P-values from empirical data [32-33]. Additionally, performance evaluation by Velez et al for this algorithm indicated 98% identification significantly associated SNPs (with 99% classification rate accuracy) using smaller sample size, with resampling of multiple pools, compared to traditional GWAS methods recruiting much larger cohorts, in the Welcome Trust Case-Control Consortium (WTCCC) [35]. Thus, it is worthwhile considering its implementation when trying to identify common disease mutations, when only sporadic cases are available, due to its overall efficiency in providing high quality results, whilst minimising the sample size, and therefore the financial cost of these studies.

Advantages of the pooling and bootstrap based GWAS methods are not only limited to evaluation of single variants. Instead, this can also be extended to gene-based statistics. This is evident in the formulas built in the Versatile Gene-Based Association Studies (VEGAS) software package [32, 36]. The analysis involves simulation of n multivariate normally distributed vectors, corresponding to the n number of SNPs in a given gene, represented by the LD matrix of P-values. This matrix is then decomposed, known as Cholesky matrix decomposition. This gives new vectors (obtained from random variables) with multivariate normal distribution [32, 36]. The p-values for the association signal of these vectors are later transformed into vectors with 1 degree of freedom chi square variables. Subsequently, these elements (representing the grouped SNPs within the given gene) are summed and combined into generating the overall test statistic, simulated, by comparing the overall case-control distribution, over multiple resampling rounds. The P-values are then derived from the proportion of simulated gene-based statistics exceeding the observed test case value [32, 36].

The accuracy of significance levels quantified from resampling seemingly matches that obtained from permutation procedures. Also, this algorithm (like the rare-variant collapsing methods) is ideal to reduce the effects of the multiple testing problem, as testing 25,000 genes is likely to reduce the FDR, compared to analysing millions of individual SNPs genome-wide [32]. As well as the time-based efficiency and accuracy of significance calculations to minimise Type 1 error, the analytical approach of VEGAS is also essential to increase the overall power of GWAS studies in terms of its ability to reduce false negatives [32]. The reason is that common variants are often observed to exhibit smaller effects than rare variants, based on current and previous empirical data [32]. Therefore, information from multiple variants from the same gene into a single vector with multiple elements is combined [31, 35]. This may enable identification of multiple mutations that can simultaneously account for a larger proportion of phenotypic variability in a given population, compared to single variants alone.

Furthermore, the VEGAS analysis also offers benefits that are available specifically through the incorporation of common variants in disease studies. Firstly, test statistics can be adjusted by linkage disequilibrium. Due to the fact LD measures (in particular r^2) are often allele frequency dependent, this is an advantageous property [37]. Correction for LD can reduce the likelihood of obtaining synthetic associations from other pathogenic common variants giving false signals [32, 37]. This is especially important in large genes, where multiple variants may be present. Whilst the algorithm may overlook the effect of rare variants, this can be overcome by cross-referencing results from VEGAS analysis with rare-variant collapsing data. This will increase the overall capacity in research to enhance understanding of the genetic etiology of complex diseases, with simultaneous identification of rare and common mutations.

6.2.4 Correlating Genomic Ancestry of Paisa with Population History

Thus far, we have quantified the value of the Paisa isolate as a population descended from multifounder effects, in terms of its LD patterns compared with other cohorts, from the 1000 Genomes Project. Nevertheless, in addition to LD, further analyses on this population's genomic ancestry can be conducted, which is useful for genetic epidemiological purposes. These include: Ancestry Specific Principal Components Analysis (ASPCA), IBD (Identity By Descent) determination within and between populations, and analysing tract lengths to infer population migration and admixture patterns [38-40]. This can be achieved using the 1000 Genomes populations [41] as (or other collected samples) per the European and Latin American populations (including Puerto Rico), analysed in chapter 5. These approaches have been previously used, and can be further expanded to the Paisa population, to provide additionally quantify its advantages as a genetic isolate. Subsequently, analysis of these cohorts can be compared to results obtained from previously analysed South American and European reference panels. European (POPRES) and Latin American reference panels are available [42-43]. However if possible, more localised samples, corresponding to countries of origin and recorded demographic history of each cohort of European populations from the 1000 Genomes project may also be selected. These studies will elucidate the genetic and demographic history of these populations, which can be pivotal in influencing disease susceptibility.

The first possible approach to consider is ASPCA. It is known that the aim of traditional PCA is to project genetic data points onto multidimensional space (each dimension representing a component), in order to illustrate as much variability between individuals as possible [38, 44, 45]. However, in this case, PCA is only performed on components of individual's ancestry that pertain to a given continental or geographic area of origin (I.e. European or Native Amerindian ancestry in this case). This is achieved by masking haplotypes that do not conform to the ancestry component being tested, identified on a locus specific level. Adjustment for missing values is conducted by obtaining the 1st and 2nd order derivatives of the mathematical function that facilitates data entry in the PCA matrix [46]. This is known as the Hessian matrix function [46]. These derived functions can then be used to input entries for missing data points, based on observed matrix values. Comparing this to previous results will potentially distinguish (between Paisa and other cohorts) population genetic differentiation and clustering patterns on the basis of position in PCA multidimensional space [38], for analyses corresponding to European and Native American ancestry components respectively.

Inferences of ancestry differentiation between chosen populations can be further analysed by calculating and comparing tract length distributions, within and between populations [38, 47]. Programs and software packages perform this by implementing an optimization function to infer tract length under given migration/population history models and with observed tract lengths. Therefore, migration models can be subsequently fitted upon local observed and calculated ancestry tract length distributions [38, 47]. This allows patterns of migration, admixture and population ancestry to be inferred, both in terms of the source populations and timing (number of generations ago, that these events occurred). The calculation of tract length distributions is predominantly a Markov chain process. Lengths are determined on the basis of transition rates and probabilities between states (of alleles and genotypes), taking into account genetic drift, variances in fraction of chromosome lengths obtained from migrant plus local populations and ancestry recombination switch points [38, 47]. Probability of allelic states are dependent on rates of recombination and the likelihood of finding ancestry of haplotype p at a given locus, depends on the ancestry proportions of the parental pools [38, 47]. This is supported by analysis of the assortment variance, which works on the premise that not all individuals contribute the same amount of genetic material to descendants [38, 47]. This can be reflected by the decay of heterozygosity over time, whose value can be used to derive the total assortment variance, from which the time of a given migration pulse can be inferred, as expressed in the mathematical models [38, 47].

Studying IBD patterns within and between populations can further support findings from the aforementioned tract length analysis. Therefore, it is possible to corroborate results and provide additional evidence of multifounder effects specifically within the Paisa community [38], which may influence linkage disequilibrium patterns and enrichment of disease susceptibility loci as described from our studies of comparative LD patterns.

Comparison of Wright-Fischer simulated data with Markov Chain predictions, have previously shown strong agreement, with regards to predicted migration rates. Also, Tract lengths under this algorithm were highly significant when applied to external data sets (ASW cohort of HapMap project), after multiple simulations [38, 47]. Therefore, given the observed evidence this model's accuracy, and that it accounts for the aforementioned evolutionary forces (drift, migration), this could be informative about drawing further comparisons of the population genetic history of the Paisa community and other analysed (chapter 5) cohorts [38]. Subsequently, this influences the impact of founder effects, admixture etc. on distribution of potential disease variants in studied populations.

6.2.5 Functional Validation of Identified Candidate Mutations.

Thus far we have shown strong statistical evidence of specific candidate mutations underpinning autoimmunity, both in terms of segregation patterns across families and sporadic cases, plus network analyses. In addition, we have also identified population genetic variation patterns that may affect the prevalence of pathogenic variants in given populations. However, it will also be important to back these findings with functional studies. Whilst network analysis can provide insights into the mechanism by which these genes can lead to disease phenotypes, this too is based purely on statistical algorithms, dependent on the size and specific identities of gene list elements. Therefore mechanistic validation of the mutations, harboured within these genes is required. To begin, it is possible to evaluate the apoptotic, endopeptidase, plasminogen urokinase activity and inflammatory activity of the candidate mutations, as these were key processes identified as being associated (from metacore analysis and additional literature searches) with the corresponding genes [7-8, 11-24]. Additionally, new mechanisms may also be identified in the future, increasing development options for therapeutic targets.

6.3 Conclusion

The incorporation of bioinformatics, statistical, population genetic and biological mechanism evidence in this epidemiological and population genetics study has enabled us to successfully identify *SRA1, LRP1, DHX34, ABCB8, MLL4, PLAUR* as potential contributors to autoimmune disease etiology. We also correlated populationn genomic variation patterns from comparative studies of LD and rare-variant prevalence of the Paisa community and other selected cohorts with their demographic history. These findings are informative about the population genetic history of the Paisa isolate, and how this influences the allelic distribution of disease variants within this cohort. It is hoped that all these results can aid future studies in epidemiology of autoimmune and other complex diseases, as well as facilitate development of targets for therapeutic intervention.

References

[1] G. Gibson. Rare and common variants: twenty arguments. *Nat Rev Genet,* 13(2):135-45, 2012.

[2] D.J. Liu and S.M. Leal. A novel adaptive method for analysis of next generation sequencing data to complex trait associations with rare variants due to gene main effects and interactions. *PLoS Genetics,* 6:e1001156.

[3] S. Lee, G.R., Abecassis, M. Boehnke, and X. Lin. Rare-Variant Association Analysis: Study Designs and Statistical Tests. *Am. J. Hum. Genet,* 95(1):5-23, 2014.

[4] S. Goodwin, J.D. McPherson, and W. Richard McCombie. Coming of age: 10 years of next-generation sequencing technologies. *Nat Rev Genet,* 17(6):333-51, 2016.

[5] M. Yandell, C. Huff, H. Hu, et al. A probabilistic disease-gene finder for personal genomes. *Genome Res,* 21(9):1529-42, 2011.

[6] H. Hu, J.C. Roach, H. Coon, et al. A Unified Test of Linkage Analysis and Rare-Variant Association for Pedigree Sequencing Data. *Nat Biotechnol*, 32(7):663-669, 2014.

[7] A. Johar, J.C. Sarmiento-Monroy, A. Rojas-Villararga, et al. (2016). Definition of mutations in polyautoimmunity. *J. Autoimmun,* 72, 65-72.

[8] A.S. Johar, C. Mastronardi, A. Rojas-Villarraga, et al. Novel and rare functional genomic variants in multiple autoimmune syndrome and Sjogren's syndrome. *J. Trans. Med*, 13:173, doi:10.1186/s12967-015-0525-x, 2015.

[9] M. Arcos-Burgos, and M. Muenke. Genetics of Population Isolates. *Clin Genet*, 61(4):233-247.

[10] M.L. Bravo, C.Y. Valenzuela, and O.M. Arcos-Burgos. Polymorphisms and phyletic relationships of the Paisa community from Antioquia (Colombia). *Gene Geogr*, 10(1):11-7

[11] A. Nilsson, L. Vesterlund, P.A. Oldenborg, et al. Macrophage expression of LRP1, a receptor for apoptotic cells and unopsonized erythrocytes, can be regulated by glucocorticoids. *Biochem Biophys Res Commun,* 417(4):1304-9, 2012.

[12] A. Fernandez-Castaneda, S. Arandjelovic, T.L. Stiles, et al: Identification of the low density lipoprotein (LDL) receptor-related protein-1 interactome in central nervous system myelin suggests a role in the clearance of necrotic cell debris. *J. Biol Chem,* 288(7):4538-48, 2013.

[13] J. Dalli, L.V. Norling, T. Montero-Melendez, et al. Microparticle alpha-2-macroglobulin enhances pro-resolving responses and promotes survival in sepsis. *EMBO Mol Med,* 6(1):27-42, 2014.

[14] S. Mandrekar, Q. Jiang, C.Y. Lee, : Microglia mediate the clearance of soluble Abeta through fluid phase macropinocytosis. J Neurosci 2009; 29:4252-62.

[15] M.H. Biermann, S. Veissi, C. Maueroeder, et al. The role of dead cell clearance in etiology and pathogenesis of Systemic Lupus Erythematosus: dendritic cells as potential targets. *Expert Rev Clin Immunol,* 10(9):1151-64.

[16] I.K. Poon, C.D. Lucas, A.G. Rossi, and K.S. Ravichandran. Apoptotic cell clearance: basic biology and therapeutic potential. *Nat Rev Immunol,* 14(3):166-80.

[17] A. Nykjaer, B. Moller, R.F. Todd 3rd, et al. Urokinase receptor. An activation antigen in human T lymphocytes. *J Immunol,* 152(2):505-16, 1994.

[18] J. Humphries, J.A. Gossage, B. Modarai, et al. Monocyte urokinase-type plasminogen activator up-regulation reduces thrombus size in a model of venous thrombosis. *J. Vasc Surg,* 50(5):1127-34, 2009.

[19] T. Chavakis, A.K. Willuweit, F. Lupu, et al. Release of soluble urokinase receptor from vascular cells. *Thromb Haemost,* 86(2):686-93, 2001.

[20] M. Baran, L.N. Mollers, S. Andersson, et al. Survivin is an essential mediator of arthritis interacting with urokinase signalling. *J. Cell Mol Med,* 13(9B):3997-808, 2009.

[21] K. Zurhove, C. Nakajima, J. Herz, et al. Gamma-secretase limits the inflammatory response through the processing of LRP1. *Sci Signal,* 1(47):ra15, doi: 10.1126/scisignal.1164623, 2008.

[22] J.J. Cush, J.B. Splawski, R. Thomas, et al. Elevated interleukin-10 levels in patients with rheumatoid arthritis. *Arthritis Rheum,* 38(1):96-104, 1995.

[23] J.M. Anaya, P.A. Correa, M. Herrera, et al. Interleukin 10 (IL-10) influences autoimmune response in primary Sjögren's syndrome and is linked to IL-10 gene polymorphism. *J Rheumatol,* 29(9):1874-6, 2002.

[24] W.F. Boron. *Medical Physiology: A cellular and molecular approach.* Elsevier/Saunders. p. 103 ISBN 1-4160-2328-3, Philadelphia, USA, pg.103, 2003.

[25] Y. Itan, S.Y. Zhang, G. Vogt, et al. The human gene connectome as a map of short cuts for morbid allele discovery. *Proc Natl Acad Sci USA,* 110(14):5558-63, 2013.

[26] Y. Itan, M. Mazel, B. Mazel, et al. HGCS: an online tool for prioritizing disease-causing gene variants by biological distance. *BMC Genomics,* 15:256, doi: 10.1186/1471-2164-15-256, 2014.

[27] L. Kruglyak, M.J. Daly, M.P. Reeve-Daly and E.S. Lander. Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am J Hum Genet,* 58(6):1347-63, 1996.

[28] K. Strauch, R. Fimmers, T. Kurz, et al. Parametric and Nonparametric Multipoint Linkage Analysis with Imprinting and Two-Locus-Trait Models: Application to Mite Sensistization. *Am J Hum Genet*, 66(6):1945-57, 2000.

[29] J. Dietter, A. Spiegel, D. an Mey, et al. Efficient two-trait-locus linkage analysis through program optimization and parallelization: application to hypercholesterolemia. *Eur J Hum Genet,* 12(7):542-50, 2004.

[30] P. Asherson, K. Zhou, R.J.L. Anney, et al. (2008). A high-density SNP linkage scan with 142 subtype ADHD sib pairs identifies linkage regions on chromosomes 9 and 16. *Mol. Psychiatry,* 13(5):514-21.

[31] K. Zhou, A. Dempfie, M. Arcos-Burgos, et al. (2008). Meta-analysis of genome-wide linkage scans of attention deficit hyperactivity disorder. *Am. J. Med. Genet. B Neuropsychiatr. Genet,* 147(B)(8):1392-8.

[32] J.I. Velez, S.C. Chandrasekharappa, E. Henao, et al. Pooling/Bootstrap GWAS (pbGWAS) identifies new loci modifying age of onset in PSEN1 p.Glu280Ala Alzheimer's Disease. *Mol. Psychiatry,* 18(5):568-75.

[33] R.A. Fischer. *Statistical Methods for Research Workers.* 4th edition, Oliver and Boyd: London, 1932.

[34] M.C. Whitlock. Combining probability from independent tests: the weighted Z-method is superior to Fisher’s approach. *J Evol Biol,* 18(5):1368-73, 2005.

[35] J.I. Vélez, C.A. Jack, A. Chuah, et al. Cross validation of pooling/resampling GWAS using WTCCC data. *Mol Biol Genet Eng,* 3:1, http://dx.doi.org/10.7243/2053-5767-3-1, 2015.

[36] J.Z. Liu, A.F. McRae, D.R. Nyholt, D.R. Nyholt et al. A Versatile Gene-Based Test for Genome-wide Association Studies. *Am J Hum Genet,* 87(1):139-45, 2010.

[37] J.M. VanLiere and N.A. Rosenberg. Mathematical properties of the r^2 measure of linkage disequilibrium. *Theor Popul Biol*, 74(1):130-37, 2008.

[38] J.R. Homburger, A. Moreno-Estrada, Gignoux, et al. Genomic Insights into the Ancestry and Demographic History of South America. *PLos Genet,* 11:e1005602. https://doi.org/10.1371/journal.pgen.1005602, 2015.

[39] J.R. Sandoval, D.R. Lacerda, M.S.A. Jota, et al. The Genetic History of Indigenous Populations of the Peruvian and Bolivian Altipano: The Legacy of the Uros. *PLoS ONE,* (8):e73006. https://doi.or/10.1371/journal.pone.0073006, 2013

[40] J.R. Sandoval, A. Salazar Granara, O. Acosta et al. Tracing the genomic ancestry of Peruvians, reveals a major legacy of pre-Columbian ancestors. *J Hum Genet,* 58(9):627-34, 2013.

[41] The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526(7571):68-74, 2015.

[42] M.R. Nelson, K. Bryc, K.S. King, et al. The Population Reference Sample, POPRES: a resource for population, disease and pharmacological genetics research. *Am J Hum Genet,* 83(3):347-58, 2008.

[43] J.M. Galanter, J.C. Fernandez-Lopez, C.R. Gignoux et al. Development of a panel of genome-wide of ancestry informative markers to study admixture throughout the Americas. *PLoS Genet,* 8(3):e1002554, doi: 10.1371/journal.pgen.1002554, 2012.

[44] K. Pearson. On Lines and Planes of Closest Fit to Systems in Space. *Philos Mag,* 2(11):559-72, 1901.

[45] A.L. Price, N.J. Patterson, R.M. Plenge, et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet,* 38(8):904-9, 2006.

[46] D.H. Alexander, J. Novembre and K. Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res,* 19(9):1655-64, 2009

[47] S. Gravel. Population genetics models of local ancestry. *Genetics*, 191(2):607-19, 2012.