UNIVERSITY of the WESTERN CAPE

SANBI South African National Bioinformatics Institute

Identification of coding variants associated with familial systemic lupus erythematosus through whole exome sequencing

by

Larry Peter van Vuuren

A thesis submitted in partial fulfilment for the degree of Doctor of Philosophy at the South African National Bioinformatics Institute, Faculty of Science, University of the Western Cape

Supervisor: Dr. Nicki Tiffin

Co-Supervisor: Dr. Junaid Gamieldien

November 2016

DAAD South Africa

Declaration

I, Larry Peter van Vuuren, declare that this thesis titled, ***"Identification of coding variants associated with familial systemic lupus erythematosus through whole exome sequencing"*** and the work presented in it is my own. I confirm that:

• This work was done wholly while in candidature for the degree of Doctor of Philosophy (PhD) at the University of the Western Cape and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at University of the Western Cape or any other educational institution, except where due acknowledgement is made in the thesis.

• Any contribution made to the research by others, with whom I have worked with is explicitly acknowledged in the thesis

Signed: 

Date:

Acknowledgements

All praise, glory and honour to the **LORD** JESUS **CHRIST** without whom nothing is possible.

My sincere appreciation and gratitude are expressed towards the staff members at the South African National Bioinformatics Institute, University of the Western Cape, for the years of support and assistance they have provided.

In particular, I would also like to thank Mr. Peter van Heusden for his expert technical assistance, and my peers Mr. Hocine Bendou, Mr Campbell Rae, Dr. Jean-Baka, Dr. Darlington Mapiye, Mrs. Tracey Calvert-Joshua, Miss. Stephanie Pitts and Dr. Galen Wright, for their help and moral support throughout the years.

I would also like to express my sincere gratitude to my supervisors Dr. Nicki Tiffin and Dr. Junaid Gamieldien for their patience, assistance, knowledge and understanding throughout the duration of the course. I would also like to give special thanks to Dr. Nicki Tiffin, for the friendliness and kindness that she has shown me. Dr. Tiffin is a wonderful person and working with her was truly an eye-opening experience as to what a good manager she actually is. She is very supportive and leads by example. Without all of the above-mentioned, this work would not have been possible.

My sincere appreciation and gratitude are also expressed towards the staff members at the Office for Students with Disabilities, University of the Western Cape, for the years of support and assistance they have provided. In particular, Mrs. Evadne Abrahams, Miss. Carmen Loubser and Miss. Zeena Spannenberg, you guys are the best and please keep up the good work.

To my uncle, Hilton Sydien and aunty, Beverly Sydien (nee: van Vuuren) (my second parents) – Thank you for your encouragement and support throughout my academic career.

To my uncle, Peter Edmund van Vuuren, former principal of Hillside High, now Chief Whip for Basic Education in the Eastern Cape for the Democratic Alliance. Thank you, uncle Edmund; for the efforts that you've exerted in opening the doors of your school when others pushed me away, thus giving me an opportunity to finish my secondary education. I truly appreciate all you have done for me.

I would also like to express my profound gratitude towards my loving mother, brother, sister and extended family – for their guidance, love and support throughout my life. Thank you, mommy you are the best.

I would also like to express my deep thanks and gratitude to my (late) grandparents George and Elizabeth van Vuuren, who have been a source of encouragement and support during my whole studies. Without their love and faith, I would not have been where I am today. For this reason, this thesis is dedicated to them as well.

To my late father, you were a man of integrity and well-liked by numerous people. I only knew you for nine years but I remember as a little boy, you nicknamed me professor and for this reason I am dedicating this degree to you because you spoke this into my life. Your presence is solely missed and it saddens me because you are not around to witness this degree being awarded to me. Thank you daddy and you'll always be my role model.

Lastly, I would also like to thank DAAD for financial assistance during the course of my studies.

Abstract

Systemic lupus erythematosus (SLE) is a chronic inflammatory autoimmune disease characterized by the production of a wide range of autoantibodies directed against self-antigens. SLE can influence almost any organ system and its appearance and course are highly varied, ranging from remission to disease flare. SLE demonstrates a variety of constitutional symptoms, such as the skin, musculoskeletal and mild hematologic involvement. On the other hand, some patients present with primarily hematologic, renal or neuropsychiatric manifestations.

Whole-exome sequencing (WES) now offers the possibility of identifying rare and novel variants responsible for complex disease. This study was undertaken to identify coding variants that may be associated with familial SLE, using WES Next Generation Sequencing (NGS) of members of a South African family with familial SLE (three affected with SLE and two unaffected). Four inheritance models, namely a susceptibility model, a tipping point model, a protective mutation model, and an asymptomatic model were proposed for the analysis because it was not clear whether the mode of inheritance is dominant or recessive. The susceptibility model considers all family members to have a background of SLE susceptibility variants; the tipping point model looks at all variants shared by affected members; the protective mutation model looks at all variants shared by unaffected members, and the asymptomatic model looks at variants shared by the three affected and one unaffected.

WES identified ten novel variants, one each in (*MYH8*, *NYX*, *SERPINB13*, *CD177*, *CD24*, *HSD11B2*, *MERTK*, and *IL36G*), and two in *PRSS1*. The variants found in *NYX*, *SERPINB13* and *IL36G* were predicted to be deleterious using PROVEAN web server protein prediction scores. Our study also reported nineteen rare missense variants with minor allele frequency < 1% one each in (*STAT4*, *C3*, *ISG15*, *PIK3CD*, *TBC1D9*, *AOC3*, *DTX3*, *MORC1*, *SLC9A2*,

TBX6, *TF*, *HP*, *SYNE2*, *MERTK*, *TNFRSF10B*, *GYPC* and *LRRC1*), and two in *LILRA2*. The variants found in *STAT4*, *ISG15*, *TBX6* and *GYPC* were all predicted to be deleterious by both SIFT and PolyPhen2 algorithms. IPA's gene network analysis showed direct interactions between six of our candidate genes (*STAT4*, *TF*, *SYNE2*, *PIK3CD*, *TNFRSF10B* and *ISG15*) and known and differentially expressed SLE genes. Furthermore, we report that some of these candidate genes mapped to known SLE pathways. Some gene candidates (*ISG15*, *STAT4* and *TNFRSF10B*) were also regulated by *STAT3* and *IL6* which have both been implicated in a wide variety of inflammation-associated disease states.

The variants found in *STAT4* and *ISG15* best supported two of the four proposed inheritance models. These models are the tipping point and protective mutation models. Within the proposed tipping point model, the three affected members shared a variant found within a known SLE susceptibility gene *STAT4*. Within the proposed protective mutation model, the two unaffected members shared a variant within the *ISG15* gene which, research has shown, can be both a promoter and inhibitor of SLE.

The *in-silico* approach utilized in this study may contribute significantly to elucidate the mechanism's that underlies SLE. This WES study has made a fundamental contribution to understanding the genetics of SLE, based on a South African family with a rare inherited form of the disease.

Table of Contents

List of tables

List of figures

List of abbreviations

ACR	American College of Rheumatology
AD	autoimmune disease
APC	antigen-presenting cells
bp	base pairs
BWA	Burrows-Wheeler transform based alignment algorithm
cM	centiMorgans
CNV	copy number variation
dbSNP	Single Nucleotide Polymorphism Database
DE	differentially expressed
DisGeNET	database of gene-disease associations
DM	diabetes mellitus
DNA	deoxyribose nucleic acid
EBNA-1	Epstein-Bar virus nuclear antigen 1
EBV	Epstein-Barr virus
EVS	Exome Variant Server
ExAC	Exome Aggregation Consortium
FATHMM	functional analysis through hidden Markov Models
GATK	Genome Analysis Toolkit
GERP	Genomic Evolutionary Rate Profiling
GO	Gene Ontology
GVF	Genome Variation Format
GWAS	genome-wide association studies
HGNC	HUGO Gene Nomenclature Committee
HGP	Human Genome Project
IPA	Ingenuity Pathway Analysis

JAK/STAT	Janus kinase/signal transducers and activators of transcription
LOD	logarithm of the odds
MAF	minor allele frequency
MN	Minnesota
MRL	Murphy Roths Large
MS	multiple sclerosis
NGS	next-generation sequencing
NR	not reported
PA	Pennsylvania
PacBio	Pacific Biosciences
PBMC	peripheral blood mononuclear cells
PCR	polymerase chain reaction
PE	paired-end
PGM	Personal Genome Machine
RXR	retinoid X receptor
SAM	Sequence Alignment/Map
SCA	Sickle-cell anemia
SE	single-end
SIFT	Sorting Intolerant from Tolerant
SLE	systemic lupus erythematosus
SNP	single nucleotide polymorphism
SOLID	Sequencing by Oligo Ligation Detection
TAPER	Tool for Automated selection and Prioritization for Efficient Retrieval
Th1	T helper 1
Th2	T helper 2
TR	thyroid hormone receptor
VAAST	Variant Annotation Analysis and Search Tool

VAT	Variant Annotation Tool
VCF	variant call format
VEP	Variant Effect Predictor
VQSLOD	variant quality score log-odds
VQSR	Variant Quality Score Recalibration
VST	Variant Selection Tool
WES	Whole Exome Sequencing
WI	Wisconsin
µg	microgram

Chapter 1: Introduction

1 Introduction to Systemic lupus erythematosus (SLE)

1.1 Background

The origins of lupus date back to the thirteenth century when a physician named Rogerius outlined for the first time erosive facial lesions, that resembled the bite of a wolf (*lupus* in ancient Latin means wolf). From the middle ages to the midst of the nineteenth century, the fundamental clinical illustrations of lupus were dermatologic, as depicted by Bateman, Cazenave and Kaposi (Blotzer, 1983; Smith and Cyr, 1988). In 1833, Cazenave used the term, *erythema centrifugum*, to characterize cutaneous lesions that are now known as discoid lupus, and in 1846 the butterfly distribution of the facial rash was described by von Hebra (Blotzer, 1983; Smith and Cyr, 1988). In the year 1872, Kaposi was the first to describe systemic appearance of lupus, such as subcutaneous nodules, arthritis with synovial hypertrophy of both small and large joints, lymphadenopathy, fever, weight loss, anemia and central nervous system involvement (Kaposi, 1872), and Kaposi's findings were later confirmed by Oslek (Oslek, 1904) and Jadassohn (Jadassohn, 1904). In 1948, Malcolm Hargraves and colleagues made a phenomenal scientific breakthrough when they discovered the lupus erythematosus (LE) cell in the bone marrow of patients with acute disseminated LE (Hargraves *et al*., 1948), as well as the false-positive test for syphilis (Moore *et al*., 1957) and the immunofluorescent test for antinuclear antibodies (Friou, 1957).

Systemic lupus erythematosus (SLE) is a chronic inflammatory autoimmune disease characterized by the production of a wide range of autoantibodies directed against self-antigens (Chiorazzi, 1987). SLE can influence almost any organ system and its appearance and course are highly mutable ranging from languid to eruptive (Bartels and Muller, 2011). Primarily, SLE demonstrates a variety of constitutional symptoms, such as skin (Figure 1.1),

musculoskeletal and mild hematologic involvement (Table 1.1) (Gladman *et al*., 1999; Schur, 2003; Husby, 1999). On the other hand, some patients present with primarily hematologic, renal or neuropsychiatric manifestations (Schur, 2003). Patients with SLE are also at high risk of contracting coronary artery disease (Manzi *et al*., 1997; Jonsson *et al*., 1989; Rahman *et al*., 1999). Respiratory and urinary system infections, are quite common in patients with SLE, making it difficult to differentiate from flares of lupus activity (Edworthy, 2001; Schur, 2003). The clinical expressions of SLE are essentially the same in children and adults (Lehman *et al*., 2004). In two definitive studies (Singh *et al*., 1997; Marini and Costallat, 1999), of children with the disease, the most persistent manifestations were fever, rash, arthritis, alopecia and renal involvement. In comparison to adults, children have a higher incidence of malar rash, anemia, leukocytopenia (Rood *et al*., 1999) and severe manifestations such as neurologic or renal involvement (Carreno *et al*., 1999).

Figure 1.1: Malar rash, the most common cutaneous manifestation of systemic lupus erythematosus (Gill *et al*., 2003).

Table 1.1: Clinical Features of SLE (Schur, 2003; Gilboe and Husby, 1999).

Affected Percentage organ system	**Percentage organ system of patients**	**Signs and symptoms**
Constitutional	50 to 100	Fatigue, fever (in the absence of infection), weight loss
Skin	73	Butterfly rash, photosensitivity rash, mucous membrane lesion, alopecia, Raynaud's phenomenon, purpura, urticaria, vasculitis
Musculoskeletal	62 to 67	Arthritis, arthralgia, myositis
Renal	16 to 38	Hematuria, proteinuria, cellular casts, nephrotic syndrome
Hematologic	36	Anemia, thrombocytopenia, leukopenia
Reticuloendothelial	7 to 23	Lymphadenopathy, splenomegaly, hepatomegaly
Neuropsychiatric	12 to 21	Psychosis, seizures, organic brain syndrome, transverse myelitis, cranial neuropathies, peripheral neuropathies
Gastrointestinal	18	Nausea, vomiting, abdominal pain
Cardiac	15	Pericarditis, endocarditis, myocarditis
Pulmonary	2 to 12	Pleurisy, pulmonary hypertension, pulmonary parenchymal disease

Concord guidelines provided by the American College of Rheumatology (ACR) outline the basis for the accurate and standardized diagnosis of SLE. The initial recommendations published by ACR in 1982 were updated in 1997 and contain 11 diagnostic categories (Table 1.1.1). The presence of any four of these criteria, either concurrently or consecutively, confirms the diagnosis of SLE (Ben-Menachem, 2010).

Table 1.1.1: ACR Classification Criteria for SLE (Tan *et al*., 1982; Hochberg, 1997).

The diagnosis of systemic lupus erythematosus requires the presence of four or more of the following 11 criteria, serially or simultaneously, during any period of observation.
1. Malar rash: fixed erythema, flat or raised, over the malar eminences, tending to spare the nasolabial folds
2. Discoid rash: erythematous, raised patches with adherent keratotic scaling and follicular plugging; possibly atrophic scarring in older lesions
3. Photosensitivity: skin rash as a result of unusual reaction to sunlight, as determined by patient history or physician observation
4. Oral ulcers: oral or nasopharyngeal ulceration, usually painless, observed by physician
5. Arthritis: non-erosive arthritis involving two or more peripheral joints, characterized by swelling, tenderness, or effusion
6. Serositis: pleuritis, by convincing history of pleuritic pain, rub heard by physician, or evidence of pleural effusion; or pericarditis documented by electrocardiography, rub heard by physician, or evidence of pericardial effusion
7. Renal disorder: persistent proteinuria, > 500 mg per 24 hours (0.5 g per day) or > 3+ if quantitation is not performed; or cellular casts (may be red blood cell, hemoglobin, granular, tubular, or mixed cellular casts)
8. Neurologic disorder: seizures or psychosis occurring in the absence of offending drugs or known metabolic derangement (e.g., uremia, ketoacidosis, electrolyte imbalance)
9. Hematologic disorder: hemolytic anemia with reticulocytosis; or leukopenia, < 4,000 per mm3 (4.0 109 per L) on two or more occasions; or lymphopenia, < 1,500 per mm3 (1.5 109 per L) on two or more occasions; or thrombocytopenia, < 100 103 per mm3 (100 109 per L) in the absence of offending drugs
10. Immunologic disorder: antibody to double-stranded DNA antigen (anti-dsDNA) in abnormal titer; or presence of antibody to Sm nuclear antigen (anti-Sm); or positive finding of antiphospholipid antibody based on an abnormal serum level of IgG or IgM anticardiolipin antibodies, a positive test result for lupus anticoagulant using a standard method, or a false-positive serologic test for syphilis that is known to be positive for at least 6 months and is confirmed by negative Treponema pallidum immobilization or fluorescent treponemal antibody absorption test
11. Antinuclear antibodies: an abnormal antinuclear antibody titer by immunofluorescence or equivalent assay at any time and in the absence of drugs known to be associated with drug-induced lupus

1.2 Epidemiology of SLE

SLE is more prevalent in females than in males, with a female to male ratio of 9:1 (Figure 1.2).

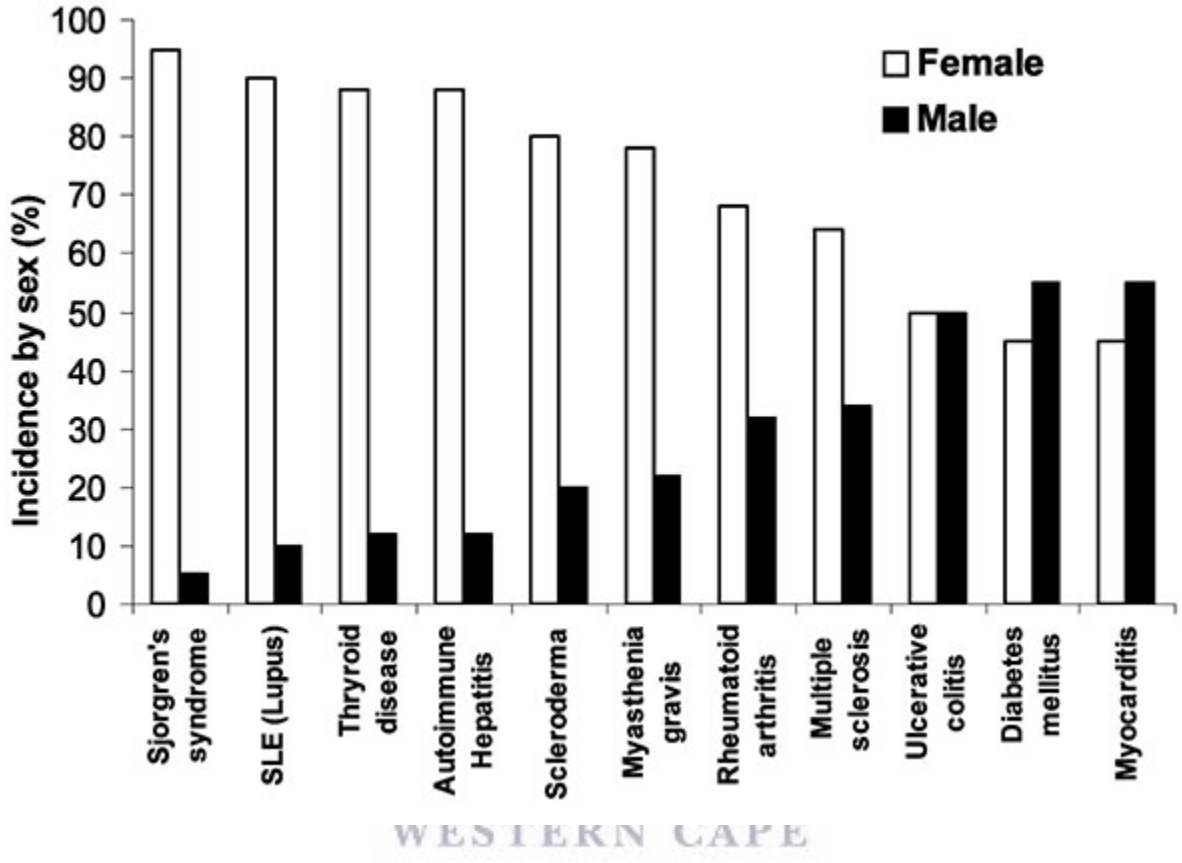

Figure 1.2: Major autoimmune diseases, comparing the prevalence of disease in women (white bar) to the prevalence in men (black bar) by percentage (Whitacre, 2001).

Global SLE incidence rates ranges from approximately 1 to 10 per 100,000 persons (Table 1.2) (Uramoto *et al*., 1999; McCarty *et al*., 1995; Naleway *et al*., 2005; Peschken and Esdaile, 2000; Nossent, 1992; Vilar and Sato, 2002; Stahl-Hallengren *et al*., 2000; Jonsson *et al*., 1990; Voss *et al*., 1998; Nossent, 2001; Johnson *et al*., 1995; Hopkinson *et al*., 1993; Hopkinson *et al*., 1994; Nightingale *et al*., 2006; Somers *et al*., 2007; Gudmundsson and Steinsson, 1990; Lopez *et al*., 2003; Alamanos *et al*., 2003; Anstey *et al*., 1993), and prevalence normally ranges from 20 to 70 per 100,000 (Table 1.2.1) (Uramoto *et al*., 1999; Naleway *et al*., 2005; Peschken and Esdaile, 2000; Nossent, 1992; Stahl-Hallengren *et al*.,

2000; Jonsson *et al*., 1990; Voss *et al*., 1998; Nossent, 2001; Johnson *et al*., 1995; Hopkinson *et al*., 1993; Hopkinson *et al*., 1994; Gudmundsson and Steinsson, 1990; Lopez *et al*., 2003; Alamanos *et al*., 2003; Anstey *et al*., 1993; Maskarinec and Katz, 1995; Chakravarty *et al*., 2007; Boyer *et al*., 1991; Hochberg, 1987; Samanta *et al*., 1991; Molokhia and McKeigue, 2000; Gourley *et al*., 1997; Al-Arfaj *et al*., 2002; Bossingham, 2003; Segasothy and Phillips, 2001; Hart *et al*., 1983).

Table 1.2: Incidence of SLE in Adults, by Location, in Studies Spanning 1975 to 2000.

Author (Reference), Country (Area), Study Period[a]	**Total Rate per 100,000 per year**[b] **(*n*)**	**Female Rate per 100,000 per year**[b] **(*n*)**
Americas		
Uramoto, United States, Minnesota (MN), 1980 to 1992	5.6 (48)	9.4 (42)
McCarty, United States, Pennsylvania (PA), 1985 to 1990	2.4 (191)	African Americans 9.2 (45) Whites 3.5 (129)
Naleway, United States, Wisconsin (WI), 1991 to 2001	5.1 (44)	8.2 (36)
Peschken, Canada (Manitoba), 1980 to 1996	First Nations ~ 3.5 (49) Whites ~ 1.2 (177)	—
Nossent, Curaçao 1980 to 1989	Afro-Caribbean 4.6 (68)	7.9 (60)
Vilar, Brazil, 2000	8.7 (43)	14.1 (38)
Europe		
Stahl-Hallengren, Sweden, 1981 to 1991	1981-86 4.5 (38) 1987-91 4.5 (41)	1981-86 5.4 (32) —
Voss, Denmark, 1980 to 1994[c]	1980-84 1.0 (—) 1985-89 1.1 (—) 1990-94 2.5 (—)	— — —
Nossent, Norway, 1978 to 1996	2.9 (83)	5.1 (73)
Johnson, United Kingdom (Birmingham), 1991	Total 3.8 (33) — — —	Total 6.8 (31) Afro-Caribbean 22.8 (6) Asian 29.2 (8) Whites 4.5 (17)
Hopkinson, United Kingdom (Nottingham), 1989 to 1990	Total 4.0 (23) Afro-Caribbean 31.9 (3) Whites 3.4 (19)	Total 6.5 (19) — —
Nightingale, United Kingdom, 1992 to 1998	3.0 (390)	5.3 (349)
Somers, United Kingdom, 1990 to 1999	4.7 (1638)	7.9 (1374)
Gudmundsson, Iceland, 1975 to 1984	3.3 (76)	5.8 (67)
López, Spain, 1998 to 2002	2.2 (116)	3.6 (102)
Alamanos, Greece, 1982 to 2001	1.9 (178)	—
Oceania		
Anstey, Australia, 1986 to 1990	Aboriginal 11	—

[a]Nightingale and Somers used the United Kingdom General Practitioner Research Database; Gudmundsson used hospital records, and all other studies used various type medical records for case ascertainment. [b]Age-adjusted rates provided when available; group-specific estimates provided when based on 2 or more cases. —, data not reported. [c]Voss included a total of 107 patients, but the number per time period was not provided.

Table 1.2.1: Prevalence of SLE in Adults, by Location, in Studies Spanning 1975 to 2000.

Author (Reference), Country (Area), Study Period[a]	**Total Rate per 100,000**[b] **(*n*)**	**Female Rate per 100,000**[b] **(*n*)**
Americas		
Uramoto, United States (MN), 1992	130 (—)	—
Maskarinec, United States, Hawaii, 1989	Total 42 (454)	Total 74 (401) Non-whites 78 (315) Whites 71 (86)
Chakravarty, United States California, PA, 2000	California 108 (—) Pennsylvania 150 (—)	Total 184 (—) African American 406 (—) Hispanic 139 (—) Asian, Pacific Island 93 (—) Whites 164 (—) Total 253 (—) African American 694 Hispanic 245 (—) Asian, Pacific Island 103 (—) Whites 203 (—)
Naleway, United States (WI), 2001	79 (64)	132 (54)
Boyer, United States Alska, 1991	112 (9)	166 (8)
Peschken, Canada (Manitoba), 1996	Total 22 (257) First Nations 42 (49) Whites 21 (177)	— — —
Nossent, Curaçao, 1989	Afro-Caribbean 48 (69)	84 (63)
Europe		
Stahl-Hallengren, Sweden, 1986, 1991	1986 42 (44) 1991 68 (41)	— —
Voss, Denmark, 1994	22 (104)	38 (93)
Nossent, Norway, 1995	50 (89)	89 (79)
Hochberg, United Kingdom, 1982	7 (20)	13 (20)
Johnson, United Kingdom (Birmingham), 1991	Total 28 (242) Afro-Caribbean 112 (50) Asian (Indian) 47 (36) Whites 21 (155)	Total 50 (227) Afro-Caribbean 197 (48) Asian (Indian) 97 (35) Whites 36 (143)
Hopkinson, United Kingdom (Nottingham), 1990	Total 25 (147) Afro-Caribbean 207 (21) Asian (Indian) 49 (7) Asian (Chinese) 93 (2) Whites 20 (117)	Total 45 (136) — — — —
Samanta, 1989, United Kingdom (Leicester)	Total 26 (50) Asian 64 (19) Whites 20 (31)	Total – Asian (Indian) 73 (13) Whites 32 (26)
Molokhia, United Kingdom (London), ages 15 to 64, 1999		Afro-Caribbean 177 (72) West African 110 (20) Whites 35 (66)
Gourley, Ireland, 1993	25 (415)	—
Gudmundsson, Iceland, 1984	36 (86)	62 (77)
Lopez, Spain, 2002	34 (367)	58 (324)
Alamanos, Greece, 2001	38 (193)	67 (170)

Middle East		
Al-Arfaj, Saudi Arabia, 1992	19 (2)	37 (2)
Oceania		
Bossingham, Australia, 1996 to 1998	Total 45 (108) Aboriginal 93 (26)	— —
Segasothy, Australia, 1999	Aboriginal 74 (18) Whites 19 (6)	— —
Anstey, Australia, 1991	Aboriginal 52 (13)	Aboriginal 100 (13)
Hart, New Zealand, 1980	Total 18 (136) Polynesians 51 (34) Whites 15 (96)	— — —

[a]Chakravarty used state hospitalization databases, adjusted for an estimate of the proportion of SLE patients hospitalized annually. Hochberg used the United Kingdom General Practitioner Research Database. Gudmundsson used hospital records. Al-Arfaj used a survey with follow-up examination, and all other studies used various type medical records for case ascertainment. [b]Age-adjusted rates provided when available; group-specific estimates provided when based on 2 or more cases. —, data not reported.

The age of onset of SLE varies, 65% of SLE patients display disease symptoms between the ages of 16 and 55 (Ballou *et al*., 1982), 20% of them manifest symptoms before the age of 16 and the remaining 15% after the age of 55 (Font *et al*., 1998). In countries outside Africa, individuals of African or Asian ancestry are more at risk of developing SLE, and the incidence/prevalence of the disorder is much higher in these ethnic groups compared to Caucasians (Cervera *et al*., 2009). Socio-economic factors such as poverty and lack of education may also contribute to the high cumulative incidence rates of SLE amongst non-whites (Alarcon *et al*., 2004). SLE prevalence differs along what is known as a tropical gradient, with the highest figures in temperate regions and lowest in the tropics (Wadee *et al*., 2007), and this anomaly may be attributed to tropical infectious diseases such as malaria. Some research suggests that a specific gene variant, known as Fcγ receptor RIIB, is actively linked to an increased risk for SLE development and that this variant also makes a person more resistant to malaria. They further suggested that Fcγ receptor RIIB would be useful in areas of the world where malaria is rife such as Africa or Asia, and would also persist in the DNA of individuals whose ancestors came from malaria regions (Brandt, 2010). It is difficult,

however, to determine the true effect of under-diagnosis on perceived differences in incidence (Tiffin *et al*., 2013).

More or less 10% of SLE patients have a relative that also has SLE, and first-degree relatives of SLE patients have an expanded probability of having a non-SLE autoimmune disease (AD) compared to the general population (Wong and Tsao, 2006; Silverman and Eddy, 2011). Numerous studies have compared disease incidence or prevalence among relatives of patients with a variety of ADs to the disease frequency among relatives of a selected control group or to estimates from the general population. Table 1.2.2 encapsulates data from studies of first-degree relatives (parents, siblings and children) (Altobelli *et al*., 1998; Dahlquist *et al*., 1989; Cederholm and Wibell, 1991; Midgard *et al*., 1996; Robertson *et al*., 1996; Sadovnick *et al*., 1988; Jankovic *et al*., 1997; Lawrence *et al*., 1987; Strom *et al*., 1994; Nagata *et al*., 1995; Koumantaki *et al*., 1997; Jones *et al*., 1996; del Junco *et al*., 1984; Lin *et al*., 1998; Ginn *et al*., 1998; Sakkas *et al*., 1995; Foster *et al*., 1993). Strong affiliations (odds ratio ranging from 5-10) are observed in studies of type I diabetes, Grave's disease, discoid lupus, and SLE.

Table 1.2.2: Studies of familial association's with ADs in first-degree relatives.

Disease study (ref)	**Location**	**Design, data source [a] number of patients**	**Familial association with the same disease [b]**	**Familial association with another disease [b]**
Type 1 diabetes mellitus (DM)				
Altobelli *et al*.	Italy	case-control (CC), Q (parents), 136	4.0(1.610.2)	Type II DM: 1.6(0.92-2.8)
Dahiquist *et al*.	Sweden	CC, Q (parents), 339	7.8 (3.6-16.8)	Type II DM: 2.1 (0.35-14.3)
Cederholm and Wibell	Sweden	CC, Q (parents), 161	7.0 (4.2-11.9)	Type II DM: 2.5 (1.4-4.4)
multiple sclerosis (MS)				
Midgard *et al*.	Norway	CC, Q (parents), 155	12.6 (1.7-552)	AD:[c] 1.2 (0.81-1.7)
Robertson *et al*.	United Kingdom	Cohort, exams, 674	9.2	not reported (NR)
Sadovnick *et al*.	Canada	Cohort, exams, 815	30-50	NR

Graves' disease Jankovi *et al.*	Serbia	CC, Q (patients), 100	7.2(0.85-60)	NR
Discoid lupus Lawrence *et al.*	United Kingdom	CC, exams, 37	7.2 (2.9-17.6)	SLE: 8.9 (1.3 99)
SLE				
Strom *et al.*	United States	CC, G (patients), 195	2.0 (0.6-7.0)	AD:[d]2.3 (1.2-4.6)
Nagata *et al.*	Japan	CC, G (patients), 282	NR	AD:[e]5.2 (1.1-25)
Lawrence *et al.*	United Kingdom	CC, exams, 36	3.5 (2.2-142)	NR
rheumatoid arthritis (RA)				
Koumantaki *et al.*	Greece	CC, Q (patients), 126	4.4 (1.7-11.1)	NR
Jones *et al.*	United Kingdom	CC, exams, 207	1.6 (0.3-8.7)	NR
del Junco *et al.*	United States	Cohort, records, exams, 78	1.7 (1.0-2.9)	NR
Lin *et al.*	United States	CC, records, 29	15.5 (2.0-122)	AD:[f]3.6 (1.2-14.5) RA and others:[f]1 1.4 (2.5-47)
Myositis				
Ginn *et al.*	United States	CC, 0 (relatives), 21	NR	AD:[g] 7.9 (2.-21.9)
Systemic sclerosis Sakkas *et al.*	Greece	CC, 0 (patients), 166	NR	Cancer: 3.8 (2.2-6.7)
Sjogren syndrome Foster *et al.*	United Kingdom	CC, 0 (patients), 42	1.9 ($p < 0.01$)	Clinical thyroid disease: 6.6 (3.5-12.3) AD:[h] 2.5

0, questionnaire or interview; ref, reference; SLE. 'Questionnaire or interview asked either of patients or of their relatives; exams = physical examination of relatives reported to have the disease of interest; records = medical record review of relatives reported to have the disease of interest. [b]Odds ratio or risk ratio and 95% confidence interval or p-value. [e]RA, psoriasis, goitre, DM. [d]RA, inflammatory bowel disease, SLE, and other AD. 'Collagen diseases, including SLE. [f]Autoimmune thyroid disease, Type I DM, rheumatic fever, ankylosing spondylitis, myasthenia gravis. #Includes autoimmune thyroid disease, RA, Type I DM, psoriasis, Sjogren syndrome, pernicious anemia, Takayasu arteritis, ulcerative colitis, hemolytic anemia, dermatomyositis, idiopathic thrombocyptopenic purpura, and other autoimmune diseases. [k]Type I DM, RA, pernicious anemia, SLE. Statistical significance not reported; odds ratio based on 7 cases in 140 relatives of probands compared to estimated population prevalence of 2%.

1.3 Aetiology of SLE

Defining the patho-aetiology of SLE remains difficult because associated risk factors are very diverse. Genetic factors, environmental factors and female sex, along with defective immune regulatory mechanisms such as the clearance of apoptotic cells, all contribute to the pathogenesis of SLE (Figure 1.3). These factors lead to an inevitable break in immunological tolerance (Bertsias *et al*., 2012). The depletion of immune tolerance, elevated antigenic load, excess T cell help, defective B cell suppression and the displacement of T helper 1 (Th1) to T

helper 2 (Th2) result in immune responses leading to B cell hyperactivity and the production of pathogenic autoantibodies (Mok and Lau, 2003).

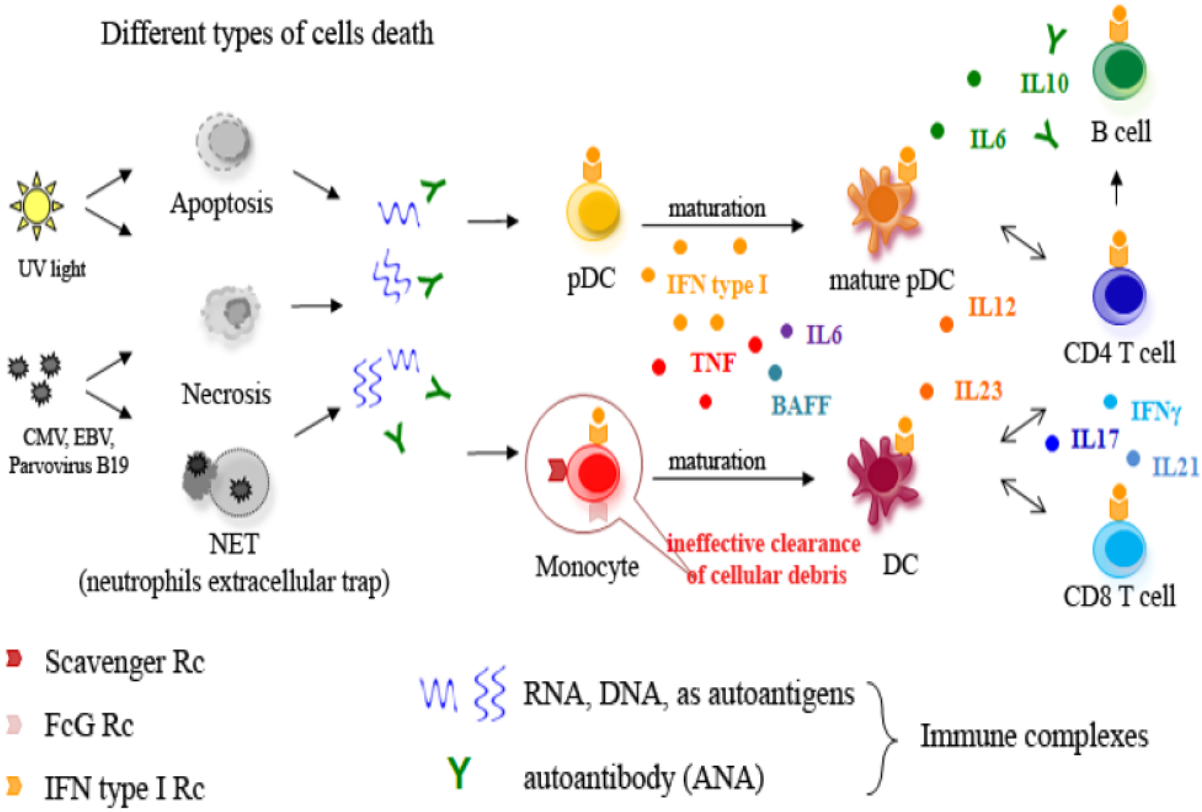

Figure 1.3: A model outlining the pathogenesis of SLE, showing that elevated amounts of apoptosis-related endogenous nucleic acids prompts the production of IFNα and induces autoimmunity by breaking self-tolerance through stimulation of antigen-presenting cells. Once activated, immune reactant's such as immune complexes intensifies and prolongs the inflammatory response (Bertsias *et al.*, 2012).

1.4 Autoimmunity and Th1/Th2 cytokine balance

Autoimmunity is a condition which is provoked by the immune system launching an attack on self-molecules due to the decline of immunologic tolerance to auto-reactive immune cells (Smith and Germolec, 1999). In the past, autoimmunity has been regarded as being synonymous with the development of clinical disease. It has become obvious that this is not always the case and that it is possible to draw a distinction between what could be described as either destructive or non-destructive autoimmunity. Destructive autoimmunity is affiliated with the development of clinical disease, whereas autoimmune responses that are non-destructive do not lead to disease. It is now evident that the relative contribution of Thl/Th2

CD4+ T cells to the evolving autoimmune response is correlated with the expression of autoimmunity as either a destructive or a non-destructive process (Charlton and Lafferty, 1995).

The Th1/Th2 balance hypothesis emanated from investigations in mice of two subtypes of CD4 T-helper cells varying in cytokine secretion patterns and other functions (Mosmann *et al*., 1986). The conception subsequently was later exercised on human immunity (Mosmann *et al*., 1989), where it was found that Th1 cells discharged pro-inflammatory cytokines such as IL2, IFNγ and TNFα, and as a result of these mediators Th1-polarized responses are highly protective against infections especially the intracellular pathogens, because of the ability of Th1-type cytokines to activate phagocytes and intensify the cellular response. On the other hand, Th2 cells discharged anti-inflammatory cytokines such as IL4, IL5 IL9 and IL13, and induces the *in situ* survival of eosinophils (through IL-5), stimulate the production of B lymphocytes leading to huge amounts of antibodies, including IgE (through IL-4 and IL-13), as well as the growth and degranulation of mast cells and basophiles (through IL-4 and IL- 9) (Figure 1.4).

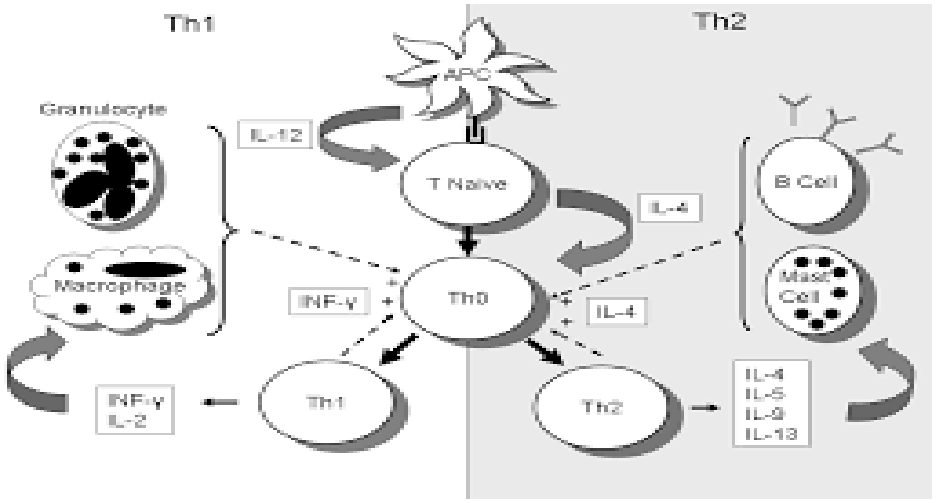

Figure 1.4: Diagrammatic representation of the differentiation into Th1 or Th2 cells from naive cells (Urra and De La Torre, 2012).

In the above figure, antigen-presenting cells (APC) interplay with undifferentiated cells secreting specific cytokines that promote differentiation towards Th1 or Th2 cells. IFNγ discharged by Th1 cells and IL4 formed by Th2 cells act as their own growth factors and cross-regulate the other differentiation. Two features designate the Th1/Th2 balance: first, each cell subset produces cytokines that serve as their own growth factor (autocrine effect), and second, the two subsets discharge cytokines to cross-regulate each other's development. Polarization to a subtype or another relies largely on the APC and exposure to the antigen. This process is controlled by the microenvironment of cytokines resulting in the antigen presentation of APC to T naive cells (Urra and De La Torre, 2012). A Th1/Th2 disproportion with excessive Th1 predominance results in organ-specific inflammatory diseases such as RA, multiple sclerosis (MS) and type 1 diabetes, whereas Th2 predominance has been described in allergy and systemic AD (Abbas *et al*., 1996). The roles of Th1 and Th2 cytokines in the pathogenesis of SLE are unclear. In patients with SLE, Th2 cytokines are increased whereas Th1 cytokines are decreased (Kunman and Steinberg, 1995; Ogawa *et al*., 1992). Therefore, SLE was originally considered to be a Th2 predominant disease. However, various findings contravene this hypothesis, for instance, that IFNγ levels in the sera of patients with SLE are fundamentally elevated and that there is an association between the

severity of SLE and the concentration of IFNγ secreted (Al-Janadi *et al*., 1993). All these discoveries state that both Th1 and Th2 responses are equally important in the pathogenesis of lupus-associated tissue injury. SLE is a disease that comprises dysregulation of a wide range of cytokines and some SLE patients with arthritis have more elevated IFN*γ* levels than other patients; and on the contrary, patients with serositis have higher IL-4 levels (Chang *et al*., 2002). Moreover, SLE patients with nephritis have higher Th1 cytokines in serum and urine than non-nephritis patients (Chan *et al*., 2006). A more significant variation in the Th1/Th2 balance in peripheral blood exists between World Health Organization class IV and V lupus nephritis. Th1 cells are prevalent in class IV but not in class V (Akahoshi *et al*., 1999). In class V, the number of penetrating Th1 cells are reduced with a large percentage of CD4 T cells producing IL4 in the peripheral blood (Masutani *et al*., 2001). Th1 or Th2 dominance depends on the stage of the disease.

1.5 The contribution of genes and genetic factors to SLE

Siblings of SLE patients are more or less 30 times at risk of developing SLE in comparison with individuals without an affected sibling (Bertsias *et al*., 2012). Identical twins with SLE are concordant for disease in about 25% of cases and are therefore discordant (i.e., where one twin has SLE and the other does not), in about 75% of cases (Schur, 1995). Genetic susceptibility influences the progression of SLE in a number of ways (Moser *et al*., 2009). Occasionally this is caused by a deficiency of a single gene (e.g. C1q) (Tsokos and Kammer, 2000; Moser *et al*., 2009), but it more commonly appears to result from the combined effects of numerous genes. In fact, modern searches for lupus genes, largely through candidate gene-association studies, genome-wide microsatellite, and single nucleotide polymorphism (SNP) association scans, have clearly indicated that SLE is a disease with multiple genetic inheritances and no single causative gene (Sestak *et al*., 2007). Therefore, the accumulation of several genes and the contributions of each allele (odds ratio ~1.5) are required to

significantly increase the risk of SLE. The amalgamation of risk alleles that leads to susceptibility and the mechanisms by which they regulate autoimmunity is not fully understood. As a matter of fact, most SNPs affiliated with SLE fall within non-coding regions of DNA and represent markers of co-segregated alleles, and many SNPs are linked to genes considered to be involved in the immune response. Over the past few years, genome-wide association studies (GWAS) have extensively increased the number of candidate genes affiliated with SLE. These genes have variable functions. Some, for instance, *IRF5*, *STAT4*, osteopontin, *IRAK1*, *TREX1* and *TLR8* are involved in nucleic acid sensing and interferon production (Abelson *et al*., 2008; Armstrong *et al*., 2009; Graham *et al*., 2007; Jacob *et al*., 2009; Kariuki *et al*., 2009; Lee-Kirsch *et al*., 2007), other genes are involved in T cell (*PTPN22*, *TNFSF4*, *PDCD1*) or B cell (*BANK1*, *BLK*, *LYN*) signaling pathways (Graham *et al*., 2008; Hom *et al*., 2008; Lu, 2009; Zikherman *et al*., 2009). The *PTPN22* gene also regulates lymphocyte activation (Zikherman *et al*., 2009). *BCL6* is the lineage-specific transcription factor of follicular helper T cells, a T cell subset that affords help to B cells in germinal centers (Nurieva *et al*., 2009). Recently, many lines of investigations also suggested that *IRF5* and *STAT4* gene polymorphisms are closely associated with disease onset of SLE (Xu *et al*., 2013). *STAT4* has also been associated with RA, whereas *PTPN22* was linked to both RA and diabetes, yet other genes seem to specifically increase the risk of SLE. A current large-scale replication study supported some of the above-named associations and systematically identified *TNIP1*, *PRDM1*, *JAZF1*, *UHRF1BP1* and *IL10* as risk loci for SLE (Gateva *et al*., 2009). Although reassuring, the loci identified thus far could only account for approximately 15% of the heritability of SLE (Manolio *et al*., 2009). The identification of candidate genes and alleles exhibits an essential step in the understanding of SLE pathogenesis, the relative significance of each gene in the global disease process and their contributions to phenotype and severity is still poorly understood. Future studies need to

address the pathways affected by genes associated with SLE and their relationships with T and B cell functional disruptions.

1.6 Environmental factors spark the onset of SLE

The contribution of the environment to the expression of SLE is well substantiated. Candidate environmental stimulators of SLE include ultraviolet light, DNA methylation and infectious or endogenous viruses or viral-like elements (Bertsias *et al*., 2012). Epigenetic variations such as DNA methylation have been proven to be associated with SLE, exposure to ultraviolet light has been acknowledged as a major hazard in clinical disease, and diverse environmental toxins, such as smoking, have also been taken into account as risk factors for SLE in epidemiological studies (Chambers *et al*., 2008). Viral infections, such as parvovirus B19 and cytomegalovirus, are prevalent in patients with SLE (Ramos-Casals *et al*., 2008). As such, there have been numerous debates that have focused on the concept that a viral infection could trigger SLE (Aslanidis *et al*., 2008). The high prevalence of the Epstein-Barr virus (EBV) in the adult population made it challenging to draw any decisive conclusions about causation, but conclusive evidence that EBV foreran SLE development was presented in a study in which serum samples from patients were examined before and after lupus development. It was found that all patients developed antibodies to the EBV protein Epstein-Bar virus nuclear antigen 1 (EBNA-1) before developing the hallmark SLE autoantibodies (e.g. anti-Ro) and SLE disease (McClain *et al*., 2005). A highly elevated level of EBV seropositivity has also been reported in pediatric SLE patients compared with healthy children (James *et al*., 1997) and a raised EBV viral titer was observed in adult SLE patients, assumed to be a consequence of a T cell disruption (Kang *et al*., 2004). Although chronic viral infection can advance to T cell debilitation, viruses have also been involved in contributing to autoimmunity through molecular imitations. Some viral proteins are identical to self-antigens and therefore trigger specific immune responses that can cross-react with

self-antigens. In addition, EBNA-1 frequently crossreacts with self-antigen Ro. Self-antigen Ro is a regular target of autoantibodies (Toussirot and Roudier, 2008). Molecular imitations are also seen with bacterial and parasitic epitopes. Over the past 30 years, the treatment of SLE was essentially based on a number of traditional drugs like corticosteroids, antimalarials, azathioprine and cyclophosphamide. Nonetheless, these drugs are rapidly being displaced due to the introduction of novel drug compounds. Some of these novel agents have been successfully utilized in other diseases, while others are being specifically devised to inhibit the immune irregularities seen in SLE (Mosca *et al*., 2001).

1.7 Understanding the genetic contributors to disease

The history of genetics roots back to the 19th century when, in 1865, Gregor Mendel, a monk in an Augustinian monastery, identified the laws of inheritance in garden peas, a feature that was omitted until "Mendelism" was reawakened in 1900 (Rimoin *et al*., 2007). Over the next half century, genetics advanced as a basic science, with a focus on *Drosophila*, the mouse, and corn as experimental systems. Most human studies were based on biostatistics and population-based mathematical analyses. Nonetheless, during this time, Mendelian inheritance was defined in multiple disorders, such as albinism, brachydactyly, and symphalangism (Keeler, 1953; Bell, 1951). A scientific approach to human genetics emanated in 1948 with the endowment of the American Society of Human Genetics (Weiss and Ward, 2000). In 1983, due to advancements in the field of molecular genetics, the Huntington disease gene was the first to be mapped to a human chromosome without any pre-existent indication of the gene location (Bates, 2005). Since then, over the following two decades, advances in human genetics have made considerable progress in genome analysis techniques leading to the discovery of a remarkable number of human disease genes. This wealth of information has also reported that the traditional difference between Mendelian and complex diseases might sometimes be obscured. Genetic and mutational data on a rising

number of disorders have depicted how phenotypic effects can develop from the combined activity of alleles in many genes (Badano and Katsanis, 2002).

1.7.1 Familial genetic linkage studies in SLE

Prior to modern genotyping and genomic approaches, linkage analysis was used to identify genetic markers that may segregate with disease-causing genes. Linkage analysis is based on the premise that genes lying close to each other are less likely to segregate during meiosis. The method searches for known genetic markers that co-segregate with the inherited disease phenotype. Families with clear Mendelian inheritance of SLE are rare, making such studies uncommon for SLE. In the early late 80s, Bias and colleagues were the first to perform human linkage studies in SLE (Bias *et al*., 1986), they used clinical and laboratory manifestations of autoimmunity as an intermediate phenotype. Their segregation data best fit a model of autosomal dominant inheritance, nonetheless, the poorly informative markers used, the insufficiency of pedigree material available to study and the apparent complexity of the underlying genetics colluded against producing significant linkage results. Researchers at the Oklahoma Medical Research Foundation conducted a genetic linkage analysis of familial SLE defined by nucleolar immunofluorescence patterns that were evaluated using six screening models of inheritance (Moser *et al*., 1998). Microsatellite genotyping data at 307 loci were utilized to screen with the two-point logarithm of the odds (LOD) scores using a maximum likelihood model-based linkage analysis. A LOD score of 5.07 was attained for marker D11S2002 on chromosome 11q14 among families of African-American ancestry. This LOD score rose above the accepted threshold of vested linkage (Center, 1995). This effect was then expanded to LOD = 5.62 using a dominant model of inheritance, a disease frequency of 0.07, 100% homogeneity, and penetrance values of 95% and 99% for males and females, correspondingly. The multipoint analysis yielded a maximal LOD score (LOD multi = 4.64) at 82 centiMorgans (cM) from the p telomere of chromosome 11 using the same

dominant model. To further corroborate the presence of linkage in the genomic neighbourhood of D11S2002, they genotyped two additional markers, D11S937 and D11S1887, which were 5.5 cM centromeric and 6 cM telomeric to D11S2002 correspondingly? Validated linkage results to this region with a LOD = 4.86 and LOD = 2.93 for markers D11S937 and D11S1887 respectively, were done using a dominantly inherited model with 92% penetrance in females and 49% penetrance in males. Other effects that exceeded the threshold for suggestive linkage (LOD 1.9) are summarized in (Table 1.7.1). Whilst this study identified broad genetic loci that may be associated with the disease, it could not address underlying genetic mechanisms for disease occurrence.

Table 1.7.1: Chromosomal regions with a screening two-point LOD score >1.9 in pedigrees multiplex for SLE and stratified by an SLE affected with a nucleolar antinuclear antibody pattern (Sawalha *et al*., 2002).

Chromosome region	**Genetic marker**	**Ethnic group**	**LOD score**
1q21-22	Fc_RIIA	EA	2.58
1q21-22	Fc_RIIA	ALL	2.89
1q21-22	Fc_RIIA	EA	1.99
1q23.2	1S1677	ALL	2.06
1q23.2	1S1677	EA	2.46
10q24.33	10S1239	AA	1.95
11p11	11S1985	AA	2.47
11p11	11S1985	ALL	2.86
11q14	11S2002	AA	5.62*
11q14	11S2002	ALL	3.46*
11q23.3	11S4464	EA	1.91
11q23-24	11S912	EA	1.99

*Effect exceeds the threshold for established linkage (LOD 3.3). This effect was maximized with a dominant model with 95% penetrance in males and 99% in females. ALL = All pedigrees; EA = European-American pedigrees; AA = African-American pedigrees.

1.7.2 Microarray genotyping and expression studies in SLE

Genetic association studies in SLE have been advancing since before 1971 when HLA-B8 was found to be associated with the disease (Grumet *et al*., 1971). During the 70s and 80s of the last century, mouse models of spontaneous lupus-like (NZB × NZW) F1 hybrids, BXSB mice (carrying the disease-accelerating Yaa gene on the Y chromosome) (Merino *et al*., 1992; Subramanian *et al*., 2006; Pisitkun *et al*., 2006), Murphy Roths Large (MRL)/lpr mice (MRL mice homozygous for a fas mutation) (Chu *et al*., 1993; Watanabe-Fukunaga *et al*., 1992) or MRL/gld mice (MRL mice homozygous for a fasL mutation) (Lynch *et al*., 1994; Takahashi *et al*., 1994) were established (Andrews *et al*., 1978; Theofilopoulos and Dixon, 1985; Cohen and Eisenberg, 1991; Izui *et al*., 1995). Research based on these mice revealed that a number of genes, loci and pathways were directly affiliated with SLE in both mouse and human species (Nguyen *et al*., 2002; Vyse and Kotzin, 1996; Santiago-Raber *et al*., 2004; Li and Mohan, 2007; Morel, 2010). Nonetheless, for the first time in the midst of 2002-2003, researches began using deoxyribose nucleic acid (DNA) microarray technology to profile genes expressed in autoimmune/inflammatory diseases. This approach was used to study: SLE (Bennett *et al*., 2003; Baechler *et al*., 2003; Crow and Wohlgemuth, 2003; Han *et al*., 2003), Juvenile Idiopathic Arthritis (Pascual *et al*., 2005; Allantaz *et al*., 2007; Ogilvie *et al*., 2007; Fall *et al*., 2007; Barnes *et al*., 2009), MS (Achiron *et al*., 2007; Singh *et al*., 2007), RA (Edwards *et al*., 2007; van der Pouw *et al*., 2007; Lequerre *et al*., 2006; Batliwalla *et al*., 2005), Sjogren's syndrome (Emamian *et al*., 2009), diabetes (Kaizer *et al*., 2007; Takamura *et al*., 2007), inflammatory bowel disease (Burczynski *et al*., 2006), psoriasis and psoriatic arthritis (Stoeckman *et al*., 2006; Batliwalla *et al*., 2005), inflammatory myopathies (Greenberg *et al*., 2005; Baechler *et al*., 2007), scleroderma (Tan *et al*., 2006; York *et al*., 2007), vasculitis (Alcorta *et al*., 2007) and anti-phospholipid syndrome (Potti *et al*., 2006).

In 2003, Rus and colleagues recorded data from a study of gene expression in peripheral blood mononuclear cells (PBMC) from 21 SLE patients and 12 controls (Rus *et al*., 2002). The microarray assay they used (Panorama Cytokine Gene Array membranes; Sigma Genosys, Inc), comprised 375 genes enriched in cytokines, chemokines, cell surface receptors, and other immune-system cell surface molecules, including adhesion molecules. Despite the Rus study not being able to provide an opportunity to detect the expression of genes that were not seemingly directly related to immune system function, it put a spotlight on several genes that had not been studied previously in detail in SLE. Overall, the data presented in this study supported the value of the microarray approach for detecting genes differentially expressed among PBMC from lupus patients. Maas and colleagues, released a second early report, pertaining to PBMC microarray data from a small number of patients with SLE (n = 9), RA (n = 9), type I DM (n = 5) or MS (n = 4), along with nine control subjects before and after immunization with influenza vaccine (Maas *et al*., 2002). They found that unfractionated PBMC yielded data that seemed to be reproducible and statistically significant, and furthermore, they found that the characterization of the cell make-up of PBMCs showed that a mutable presence of mononuclear cell populations could not account for differential gene expression. Through advances in technology, it has become apparent that microarray gene-expression studies were rapidly being succeeded by sequenced-based methods, which had the potential to detect and evaluate rare transcripts without prior knowledge of a particular gene and could provide information regarding alternative splicing and sequence variation in identified genes (Wold and Myers, 2008; Wang *et al*., 2009).

1.7.3 Sanger Sequencing

The contemporary origins of DNA sequencing began in 1977, when Sanger and colleagues as well as Maxam and Gilbert, formulated methods to sequence DNA by chain termination and fragmentation techniques, respectively (Sanger *et al*., 1977; Maxam and Gilbert, 1977). This

method was termed Sanger sequencing and revolutionized biology by providing the tools to elucidate complete genes and later entire genomes. Even though the Sanger method was still regarded by the research community as the gold standard for sequencing, it had numerous limitations (Men *et al*., 2008). A major drawback of the Sanger method for larger sequence outputs was the use of gels which acted as sieving separation media for the fluorescently labeled DNA fragments (Adelson *et al*., 2005). A rather low number of samples could be analyzed in parallel (Ahmadian *et al*., 2000a), and the overall automation of the sample prep methods was problematic (Ahmadian *et al*., 2000b). DNA fragments had to be cloned into bacteria for larger sequences (Ansorge, 2009), cost of sequencing was high (Bains and Smith, 1988), sequencing errors were frequent and level of sensitivity (generally estimated at 10-20%) was incapable of detecting clinically relevant low-level mutant alleles (Benkovic and Cameron, 1995). The method was also unable to analyze complex diploid genomes at low cost, and *de novo* genome assembly was challenging (Drmanac *et al*., 1989; Espinosa *et al*., 2003).

When the Human Genome Project (HGP) was launched in 1990, Sanger sequencing became the method of choice (Lander *et al*., 2001). However, this method underwent multiple improvements, ultimately leading to the complete sequence of 3 billion base pairs (bp) contained within a human genome (Marzillier, 2013). It took 13 years and an estimated $3.8 billion to complete HGP (Tripp and Grueber, 2011). For this reason, the National Human Genome Research Institute introduced a funding program with the intention of downsizing the cost of human genome sequencing to US $1000 in ten years (Schloss, 2008). This accelerated the development and marketing of next-generation sequencing (NGS) technologies, as opposed to the automated Sanger method.

1.7.4 Inauguration of Next Generation Sequencing (NGS)

The advent of NGS technologies in the marketplace has transformed the way we can pursue avenues of basic, applied and clinical scientific research. The power of NGS is analogous to the early days of polymerase chain reaction (PCR) in its transformative effect on genetic research. The major enhancement offered by NGS is its potential to produce huge amounts of data cheaply, in excess of one billion short reads per instrument run. This attribute has broadened the realm of experimentation beyond just determining the order of bases. The ability to sequence the whole genome of multiple related organisms has permitted large-scale comparative and evolutionary studies to be performed that were inconceivable just a few years ago. The widespread application of NGS was the re-sequencing of human genomes to improve our understanding of how genetic variations affect health and disease. This array of NGS attributes also made it likely for numerous platforms to exist together in the marketplace, with some having clear advantages for certain applications over others (Branton *et al*., 2008). The efficacy and pervasive availability of NGS platforms has significantly expanded the scale of many DNA-sequencing applications, from detecting SNPs (Dalca and Brudno, 2010) or copy number variations (CNV) (Alkan *et al*., 2011), to assembling (novel) genomes or transcriptomes (Flicek and Birney, 2009), developing quantitative RNA-sequencing analysis (Pepke *et al*., 2009), or detecting epigenetic changes (Meaburn and Schulz, 2011). NGS technology permits sequencing short fragments of DNA across the entire genome, generating single-end (SE) or paired-end (PE) reads of 50-700 bp. The reads usually require some pre-processing conversion steps. The emergent raw DNA-sequencing read data is then analyzed following two computational macro-processes: mapping and assembling (Dalca and Brudno, 2010), quality control, quality score recalibration, realignment in problematic regions of the genome, and (Alkan *et al*., 2011) advanced steps focused on variant calling (SNPs, insertions-deletions (Indels) and CNVs) and annotation. A typical

NGS workflow is illustrated in Figure 1.7.4. A crucial initial step in this process involves preparation of a "library" comprising DNA fragments ligated to platform-specific oligonucleotide adapters. The input nucleic acid can be genomic DNA, standard or long-range PCR amplicons, or cDNA (McKernan *et al*., 2009; Wheeler *et al*., 2008; Margulies *et al*., 2005). High-quality DNA in ample quantity is the basis for any effective sequencing experiment. For many sequencing applications, about 1 to 5 microgram (μg) of purified DNA is required, an amount that may not always be reachable. However, whole genome amplification is frequently employed to increase the amount of DNA for genotyping (Lovmar and Syvanen, 2006). The DNA samples to be sequenced are first transformed into one of two main types of sequencing libraries, fragment libraries or mate-pair libraries. The initial step in the preparation of a sequencing library involves shredding the DNA sample, usually using sonication or nebulization. For the preparation of fragment libraries, sequencing adapters are ligated to both ends of the DNA fragments, followed by PCR amplification using primers complementary to the adapters. The amplified fragments are then sequenced either from one end SE or from both ends PE. Paired reads grants more accurate alignment to a reference genome, and are also very useful to untangle repeats (Berglund *et al*., 2011).

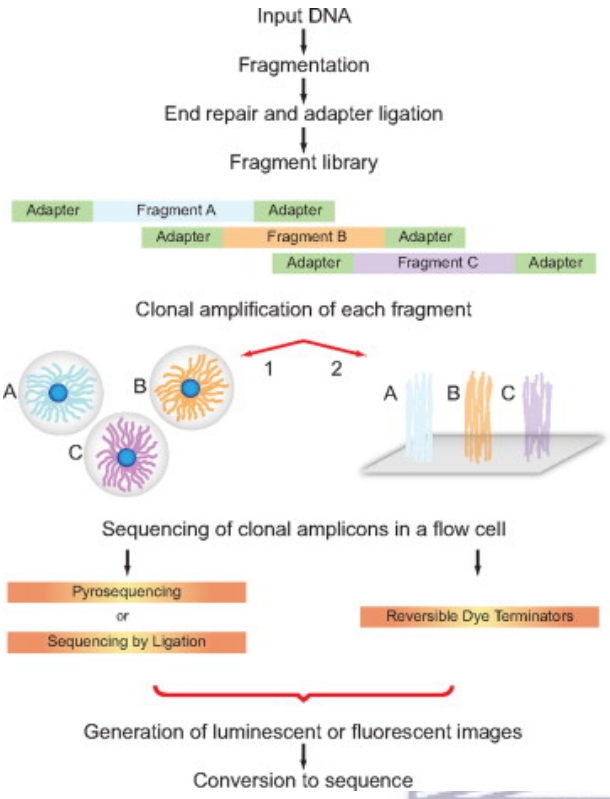

Figure 1.7.4: NGS process steps (Voelkerding *et al*., 2010).

The first NGS platform that became commercially available in 2005 was the pyrosequencing method by 454 Life Sciences (now Roche) (Margulies *et al*., 2005). Pyrosequencing technology was developed at the Royal Institute of Technology and was the first substitute to conventional Sanger sequencing for *de novo* DNA-seq (Gharizadeh *et al*., 2003). This technique had potential advantages of accuracy, flexibility, parallel processing and was easily automated. One year later, the Solexa/Illumina second generation NGS platform was made public (Illumina attained Solexa in 2007). In 2007, Applied Biosystems (now Life Technologies) launched another second generation NGS platform named Sequencing by Oligo Ligation Detection (SOLiD) (Valouev *et al*., 2008). The Illumina and SOLiD platforms produced much larger numbers of reads than 454 (30 and 100 million reads, subsequently) but the reads generated were only 50-100 bp long. In 2010, Ion Torrent (now Life Technologies) launched the Personal Genome Machine (PGM). This platform was refined by Jonathan Rothberg, the founder of 454, and was similar to the 454 system. Higher speed,

lower cost, and smaller instrument size were all hallmarks of PGM. The first PGM produced up to 270 Mb of sequence with up to 100 nucleotide reads, much shorter than those generated by 454. Additional third generation NGS platforms have been developed, such as Qiagen-intelligent bio-systems sequencing-by-synthesis (Ju *et al*., 2006), Polony sequencing (Shendure *et al*., 2005), and a single molecule detection system (Helicos BioSciences) (Pushkarev *et al*., 2009). In this modern system, the template DNA was not amplified before sequencing, which placed this method at the frontier between NGS and the so-called third-generation sequencing technologies. Third-generation methods also allowed the identification of single molecules and as an additional frequent attribute, sequencing occurs in real time (Schadt *et al*., 2010). The forerunner in this field is presently Pacific Biosciences (PacBio). Their first instrument, the PacBio RS, emerged in 2010 and produced a multitude of long kilobase reads (Eid *et al*., 2009). The long reads made this technology optimal for the completion of *de novo* genome assemblies. PacBio is based on the analysis of natural DNA synthesized by a single DNA polymerase. It involves the incorporation of phosphate-labeled nucleotides which is processed to base-specific fluorescence detected in real time. The sequencing runs generally takes minutes or hours instead of days. Table 1.7.4 illustrates many of the characteristics of the most recently utilized NGS technologies.

Table 1.7.4: An evaluation of various NGS platforms.

Platform	Library/ Preparation	NGS chemistry	Read length (bases)	Run time	Accuracy	Cost per megabase	Pros	Cons	Biological applications	Refs
Roche/ 454 GS FLX Titanium	Emulsion PCR	Pyro-sequencing	400	10 h	99.5%	$84.39	Longer reads improve mapping in repetitive regions; fast run times	High reagent cost; high error rates in homopolymer repeats	Bacterial and insect genome *de novo* assemblies; medium scale (<3 Mb) exome capture; 16S in metagenomics	Margulies *et al*., 2005
Illumina/ Solexa	Bridge PCR	Reversible terminators	100	4-9 days	>98.5%	$5.97	Currently the most widely used platform in the field	Low Multi-plexing capability of samples	Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in meta-genomics	Solexa Home page
SOLiD	Emulsion PCR	Sequencing by ligation	50	7-14 days	99.94%	$5.81	Two-base encoding provides inherent error correction	Long run times	Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in meta-genomics	Applied Biosystems Home Page
Polonator	Emulsion PCR	Sequencing by ligation	26	5 days	>99%	~$1	Least expensive platform; open source to adapt alternative NGS chemistries	Users are required to maintain and quality control reagents; shortest NGS read lengths	Bacterial genome resequencing for variant discovery	Shendure *et al*., 2005

HeliScope	Single molecule	Reversible terminators	32	8 days	>99%	~$1	Non-bias representation of templates for genome and seq-based applications	High error rates compared with other reversible terminator chemistries	Seq-based methods	Helicos Home Page
Pacific Biosciences (target release: 2010)	Single molecule	Real-time	964	N/A	N/A	N/A	Has the greatest potential for reads exceeding 1 kb	Highest error rates compared with other NGS chemistries	Full-length transcriptome sequencing; complements other resequencing efforts in discovering large structural variants and haplotype blocks	Schadt *et al.*, 2010

With the continuous advancement of NGS technologies, DNA sequencing costs have been massively downsized (Table 1.7.4). Now, it is viable to sequence all known genes for a single individual with a suspected genetic disease or complex disease predisposition (Pareek *et al*., 2011).

1.7.5 Contributions of NGS to understand the genetic factors that underlie human disease

During the completion of the HGP (Lander *et al*., 2001), the largest providers had vested factory-like sequencing facilities using capillary sequencers aided by complex robotics and infrastructure. Even these limitless operations, however, were not applicable for studying variation across individuals because the technology was actually a scaled-up version of a concept established 25 years earlier, namely Sanger sequencing (Sanger *et al*., 1977). On the contrary, technology development for human genetics was focused on procedures to effectively genotype individuals at known SNPs. A distinct DNA sequence probe could target the known SNP position, and other alleles could be distinguished using techniques such as mass spectrometry (Tang *et al*., 1999), or by imaging fluorescently labelled-primers (Shen *et al*., 2005). Associations between deviation in DNA sequence and deviation in human phenotypes were made even prior to the discovery of the genetic code, not to mention DNA sequencing technology. Sickle-cell anemia (SCA) became the initial 'molecular disease' when the underlying biology was proposed (Pauling *et al*., 1949), and then endorsed (Ingram, 1957), that the biochemical variations between sickle and healthy hemoglobin were caused by a difference in the underlying sequence of amino acids. It took decades, for the discovery of mutations that gave rise to Mendelian (single-gene) diseases such as SCA to become main-stream. A unification of large clinical collections of affected families, the advancement of statistical methods for linkage analysis (Elston and Stewart, 1971), and the design of molecular techniques for genotyping polymorphic sites in the human genome (Botstein,

1980) generated an explosion of Mendelian gene mapping. This assembly of improvements in various scientific fields allowed the first of multiple transitions in disease genetics from a broadly theoretical enterprise to a data-driven, experimental science. Sadly, these within-family linkage studies were unsuccessful when applied to complex diseases. A theoretical solution to the dilemma in finding complex disease genes was proposed when it was proven that comparing the allele frequencies of variants across the genome amongst thousands of cases and controls would be well-suited to identify common alleles of small effect, which were unseen to even the largest linkage studies (Risch and Merikangas, 1996). This methodology led to the introduction of the GWAS era and, once more, a major diversion in the practice of disease genetics was put into effect by a timely combination of the accessibility of large collections of patient DNAs, new cheap technologies for genotyping hundreds of thousands of SNPs, and methods for analyzing and elucidating this avalanche of novel data (Burton *et al*., 2007). Despite the fact that GWAS identified thousands of statistically impeccable associations, it swiftly became clear that this approach alone would not be able to explain the full range of genetic susceptibility to complex disease. For example, researchers studying Crohn's disease, a common form of inflammatory bowel disease, victoriously identified 71 associated loci, but unfortunately these only explained 23% of the heritability of the disease (Franke *et al*., 2010). Studies on other diseases, such as diabetes, were even less successful in explaining heritability (Voight *et al*., 2010). The stage was set for yet another regeneration of the techniques for understanding the genetic causes of human disease. The International HapMap Project (Consortium, 2005) sanctioned the GWAS era by implementing an elaborate inventory of common SNPs in the genome, as well as their patterns of linkage disequilibrium, in diverse populations from around the globe. This knowledge allowed the development of arrays of a few hundred thousand SNPs that secured nearly all of the frequent variation (variants with a minor allele frequency (MAF) >0.05) in

European populations (Barrett and Cardon, 2006), but was low-priced enough to be run on thousands of samples. Furthermore, the earliest benefits of high-throughput sequencing came from sequence-based reference datasets of population variation, largely the 1000 Genomes Project (Consortium, 2010). The 1000 Genomes pilot project data from 179 samples was already operational and had been integrated into the study of complex diseases and traits in two predominant ways: as enhanced imputation reference sets and in the development of next-generation genotyping arrays.

NGS has already revolutionized the study of Mendelian disease by counteracting the laborious process of finding the causative mutation via linkage analysis in affected families, followed by fine-mapping and Sanger sequencing of positional candidate genes. Alternatively, it was possible to sequence directly all the exomes - the portion of the genome containing known genes - of individuals with Mendelian diseases, and compare them to exome sequence from unaffected controls. Whole Exome Sequencing (WES), a prominent NGS method, has been outstanding when applied to such diseases, where most causal alleles interrupt protein-coding (exonic) sequences (Stenson *et al*., 2009). Though often regarded as a cheaper alternative to whole-genome sequencing that does not analyse non protein-coding regions of the genome where regulatory motifs might lie, WES has also been utilized to study complex diseases in situations where coding variation is inclined to play a major role. For example, highly penetrant CNVs have been shown to play a crucial role in the risk of autism and other clearly polygenic neurodevelopmental phenotypes (Glessner *et al*., 2009; Pinto *et al*., 2010). WES in autistic individuals and their parents lead to the discovery of potentially severe *de novo* mutations (Neale *et al*., 2012; O'Roak *et al*., 2012; Sanders *et al*., 2012). WES in families with numerous individuals affected by a complex disease may also be an adequate approach for finding disease genes because it would be possible to identify family-specific variants that would be unseen by traditional across-family linkage (Hinrichs and

Suarez, 2011). Specifically, the rare, highly pervasive mutations might well exist in complex diseases but be spread across dozens of genes. For instance, this technique was used to discover a severe mutation causing low High-density lipoprotein-cholesterol in a Canadian family with 75 members (Reddy *et al*., 2012). WES was also used by Ellyard and colleagues, in a 4-year-old girl with early-onset SLE. They identified a rare, homozygous mutation in the three prime repair exonuclease 1 gene (*TREX1*) that was predicted to be highly deleterious (Ellyard *et al*., 2014).

1.7.6 Bioinformatics and NGS

NGS has been extensively endorsed by the research community (Mardis, 2011), and is rapidly being invoked clinically, driven by awareness of its diagnostic utility and improvements in quality and speed of data procurement (Brownstein *et al*., 2014). Nonetheless, with the ever-expanding rate at which NGS data is propagated, it has become essential to revamp the data processing and analysis workflow in order to bridge the gap between big data and scientific discovery. In the matter of deep whole human genome comparative sequencing (re-sequencing), the analytical process to go from sequencing instrument raw output to variant discovery depends upon numerous computational steps. This analysis process could take days to complete, and the resulting bioinformatics overhead represents a significant impediment as sequencing costs dwindle and the rate at which sequence data is produced continues to grow rapidly. There are two phases of NGS data analysis, namely primary and secondary analysis. Primary analysis generally defines the process by which instrument-specific sequencing measures are transformed into FASTQ files comprising the short read sequence data and sequencing run quality control metrics. Secondary analysis involves alignment of these sequence reads to a human reference genome and detection of variations between the patient sample and the reference. The most frequently employed secondary analysis approach assimilates five sequential steps. These include: initial

read alignment (Gonzaga-Jauregui *et al*., 2012), elimination of duplicate reads (deduplication) (Mardis, 2011), local realignment around known indels (Brownstein *et al*., 2014), recalibration of the base quality scores (Cock *et al*., 2010), and variant detection as well as genotyping (DePristo *et al*., 2011). The final output of this process, yields a variant call format file, which is then ready for tertiary analysis, where clinically relevant variants are discovered. Of the stages of human genome sequencing data analysis, secondary analysis is by far the most computationally demanding. This is as a result of the size of the files that must be altered, for establishing optimum alignments of millions of reads to the human reference genome, and sequential alignments for variant calling and genotyping. Abundant software tools have been generated to carry out secondary analysis steps, each with conflicting strengths and weaknesses. Of the multitudinous aligners available (Schbath *et al*., 2012), the Burrows-Wheeler transform based alignment algorithm (BWA), is most frequently used due to its precision, speed and ability to output Sequence Alignment/Map (SAM) formats (Li and Durbin, 2009). Picard and SAMtools are ordinarily used for the post-alignment processing steps and outputs SAM binary (BAM) format files (Li *et al*., 2009). Considerable statistical methods have been devised for variant calling and genotyping in NGS studies (Nielsen *et al*., 2011), with the Genome Analysis Toolkit (GATK), being the most prominent (DePristo *et al*., 2011). The bulk of NGS studies integrate BWA, Picard, SAMtools and GATK to pinpoint and genotype variants (Gonzaga-Jauregui *et al*., 2012). In spite of all these incredible advancements, there remain scientific questions pertaining to sample enrichment, sequencing methodologies and variant discovery and calling algorithms, which still require careful scrutiny in order to validate the analytical steps of NGS techniques for clinical applications. Nonetheless, the fundamental predictive challenge lies within the elucidation of the clinical significance of the variants observed in a given patient, and their significance for family members and for other patients. Every step in the variant clarification

process has impediments and complications, such as their contributions to false positive and negative results. There is no single piece of information enough on its own to make concrete conclusions regarding the pathogenicity and disease causality of a given variant (Quintáns *et al*., 2014).

Researchers must rely on a variety of evidences such as case-control, segregation, family history, or other statistical studies for direct association of variant to the disease (Goldgar *et al*., 2008).

1.8 Aim of Thesis

The aim of this study is to identify coding variants that may be associated with familial systemic lupus erythematosus, using whole exome NGS of family members of a South African family with familial SLE.

1.8.1 Objectives

1.8.1.1 To characterize and define possible models of inheritance of SLE in a South African family with familial lupus.

1.8.1.2 To harvest blood DNAs from five family members, three affected with SLE and two unaffected, using a PaxGENE DNA protocol and to ship samples to Otogenetics for whole-exome sequencing using next generation technologies.

1.8.1.3 To analyze raw sequence data to identify genetic variants in all individuals sequenced.

1.8.1.4 To identify possible aetiological variants that may be contributing to SLE in the family members, under the different hypothesized models of inheritance defined for this analysis.

Chapter 2: Inheritance models underlying variant analysis and prioritisation

2.1 Mendelian Inheritance

Diseases may be classified based on the mode of disease transmission. Fundamental arrangements have characterized diseases as communicable and non-communicable based on the corresponding presence or absence of a pathogenic microorganism required for disease transmission. Nevertheless, a crucial discrepancy in disease taxonomy was the observed inheritance of diseases or traits in offspring following mathematical ratios as per Mendel's early studies in peas (Mangino and Spector, 2012). In the 1860s, Gregor Mendel outlined the Mendelian laws of inheritance which describe how heredity factors (genes), of which an offspring attains two versions (alleles – one from each parent) can affect variation in phenotypes (Bateson and Mendel, 1902). Mendel witnessed through the crossing of pea plants how a phenotype, in his case the colour of the flower, is passed through two successive generations in a distinct manner (rather than being a blend of the colour of the parents) via certain principles of segregation. For a given gene, which of the two parental alleles an offspring receives is random, and by performing a large number of crosses, Mendel was able to infer the two alleles (genotype) of each individual plant depending on whether the phenotype displayed dominance or recessive characteristics (Figure 2.1).

While Mendel's laws could sufficiently describe the experimental discrete inheritance patterns of some traits, they did not appear to apply to the majority of traits where variation appeared to be continuous, nor to discrete traits that did not follow any noticeable patterns of Mendelian inheritance. Furthermore, Mendel's laws appeared to be unreliable with natural selection, where evolution occurs via the accumulation of minute, gradual variations. These apparent inconsistent observations were resolved in the 1930s in what became known as the modern evolutionary synthesis. Ronald Fisher and others showed that quantitative traits such

as height can be described by numerous genes, each with small additive effects acting according to Mendel's laws of inheritance (Fisher, 1930). Together, these minor independent effects, along with the environment give rise to a phenotype that estimates the normal distribution.

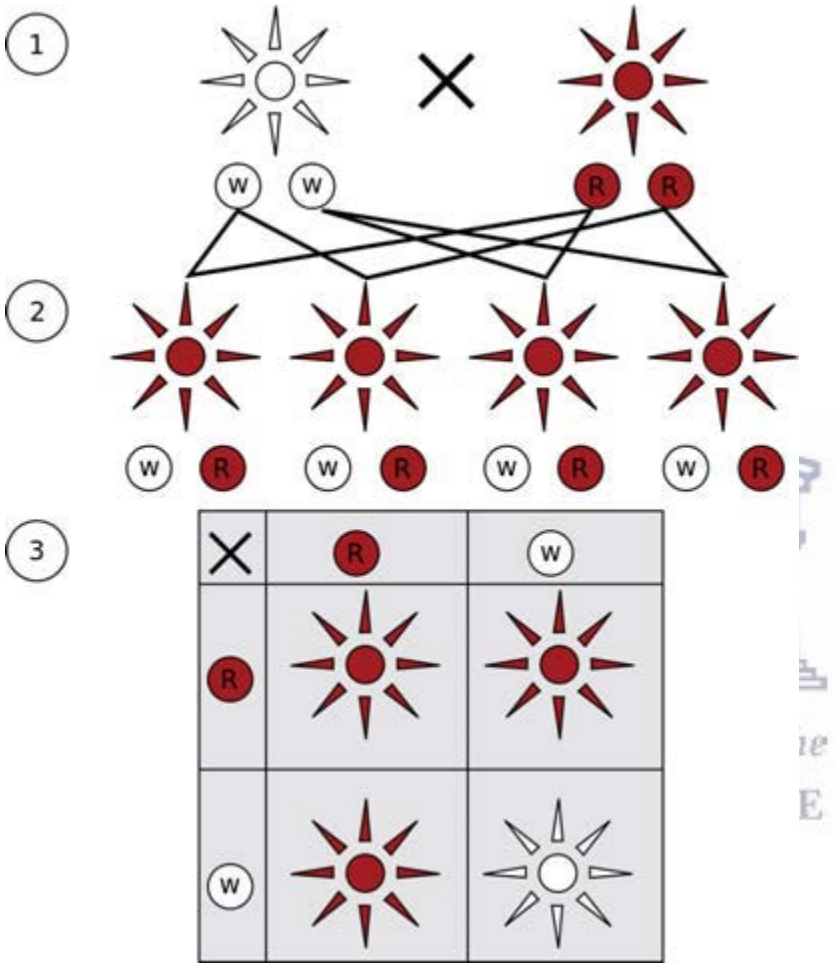

Figure 2.1: Mendel's laws of inheritance. This figure depicts two alleles: W and R which gives rise to either a white or red phenotype when both copies are present. Red is dominant and white is recessive. In (1) the parental generation, the parents are homozygous for each of the alleles. In (2) the first generation, all offspring are heterozygotes and will show the red phenotype. When heterozygotes cross, (3) the offspring will show a 3:1 red:white ratio depending upon which of the two alleles they inherit (Fisher, 1930).

Binary phenotypes such as disease status are often the result of multiple genes, each with small effects, and the environment. These complex disorders can be modelled quantitatively with a liability threshold model in a similar manner to that proposed by Fisher (Falconer and Mackay, 1996).

2.1.1 Modes of inheritance

For many single-gene/monogenic disorders, inheritance of a mutated copy or copies of a gene causes a characteristic phenotype, and inheritance of that phenotype follows a Mendelian segregation pattern. The pattern of inheritance can forecast whether the mutated gene is on an autosome (chromosomes 1-22) or is sex linked (on the X or Y chromosome) and whether the disorder is dominant (in which case a single copy of the mutated gene is adequate to cause the disorder) or recessive (both copies of a gene must be mutated to cause the disorder) (Wallace, 1999).

2.1.2 Autosomal dominant inheritance

In autosomal dominant inheritance an affected person generally has at least one affected parent, the disorder affects both sexes and can be transferred by both sexes, and an affected person has a 50% chance of passing the defect onto their offspring. Autosomal dominant conditions include Huntington disease and achondroplasia (Wallace, 1999).

2.1.3 Autosomal recessive inheritance

In autosomal recessive inheritance, affected persons are generally born to unaffected parents. Parents of affected people are ordinarily asymptomatic but carry a single copy of the mutated gene. There is an increased incidence of autosomal recessive disorders in families where parents are related. Offspring of parents who are both heterozygous for the mutated gene have a 25% chance of inheriting the disease if they receive the mutated copy from either parent; and the disease affects both sexes. Autosomal recessive conditions include cystic fibrosis and SCA (Grant, 1997; Read, 1992).

2.2 Complex diseases

Compared to monogenic diseases, where a single variant is sufficient to cause a disease phenotype, complex diseases represent a greater synergy between environmental and genetic factors. No single genetic variant is adequate to cause a complex disease phenotype but rather numerous variants of low penetrance at multiple loci contribute synergistically to enhance disease susceptibility. The leading model for complex diseases is referred to as the common disease-common variant hypothesis which envisages that several commonly occurring variants from multiple genes individually contribute a small effect on disease susceptibility but additively exert a considerable effect in the expression of complex disease (Reich and Lander, 2001). An emergent hypothesis has unified the potential contribution of rare variants of proportionately larger effect on complex disease susceptibility; nonetheless, this concept was assessed for validity in the most common complex diseases (Pritchard, 2001). In support of the heterogeneous nature of complex disease susceptibility, complex diseases are not inherited according to the models which apply to monogenic diseases. Instead, common variants are believed to contribute to the overall picture of disease susceptibility. GWAS have identified hundreds of genetic variants associated with complex human diseases and traits, and have provided valuable intuitions into their genetic architecture. Most variants identified so far confer relatively small increments in risk, and explain only a small proportion of familial clustering, leading many to question how the remaining, 'missing' heritability can be explained (Manolio *et al*., 2009).

2.3 Penetrance of a variant

Penetrance is defined as the percentage of individuals having a specific mutation or genotype who display clinical signs or phenotype of the related disease (Cooper and Krawczak, 2013). Complete penetrance designates that all individuals who have the related disease causing

mutation will develop clinical symptoms of the disease e.g. familial adenomatous polyposis, multiple endocrine neoplasia and retinoblastoma. Incomplete or reduced penetrance specifies that some individuals fail to express the trait, even though they carry the allele e.g. incomplete penetrance of the dominant mutations in the *LMNA* gene of Emery Dreifuss muscular dystrophy (Vytopil *et al*., 2002). Low penetrance designates that this allele will only sometimes produce the symptoms at a measurable level e.g. low penetrance of retinoblastoma for p. V654L mutation of *RBI* gene (Hung *et al*., 2011). Pseudo-incomplete penetrance specifies that the reflection of non-penetrance is erroneous because the clinical examination is incomplete or the symptoms have not yet appeared at the time of the examination. This is also observed in germ line mosaicism that is only seen in the first generation when the parents of numerous affected children with a dominant disease are healthy (Hung *et al*., 2011). Mutable expressivity on the other hand means the degree to which a genotype is phenotypically expressed in individuals. Individuals with a certain mutation can display variances in disease severity, even among members of the same family. Examples of diseases that display a range of phenotypes include neurofibromatosis, holoprosencephaly and genetic syndromes (Lobo, 2008; Shawky *et al*., 2013; Shawky *et al*., 2014). Healthy individuals can harbour a huge number of possibly or mildly detrimental variants and imaginably tens of potentially severe disease alleles without suffering any obvious ill effects (Xue *et al*., 2012). These variants may harm the protein in question, but the intact protein may not be necessary for the health of the carrier. The individual may be an asymptomatic carrier of a single recessive mutant allele or the mutation is dominant, but the clinical phenotype might only be mild and lie within the range of normal healthy variations or become apparent only in later decades of life (Cooper and Krawczak, 2013).

2.4 Pedigree of the family in this study

The figure below shows the pattern of inheritance for SLE for the pedigree analysed in the current study. In the first generation, individuals are the children of a cross between an unaffected mother, and two different unaffected fathers. None of the individuals in the first generation were available for the study. In the second generation, an unaffected female (SLE_01B) and an unaffected male have two daughters: one daughter (SLE03) is affected with neuro lupus and the other is unaffected (SLE02). In this same generation, an unaffected male has an affected daughter (183/14), with lupus nephritis. In the third generation, an affected female (SLE03) and a male with type 1 DM have monozygotic twin daughters. One twin (PID_017) is affected, with immune thrombocytopenia; and the other affected with SLE and type 1 DM (Figure 2.4).

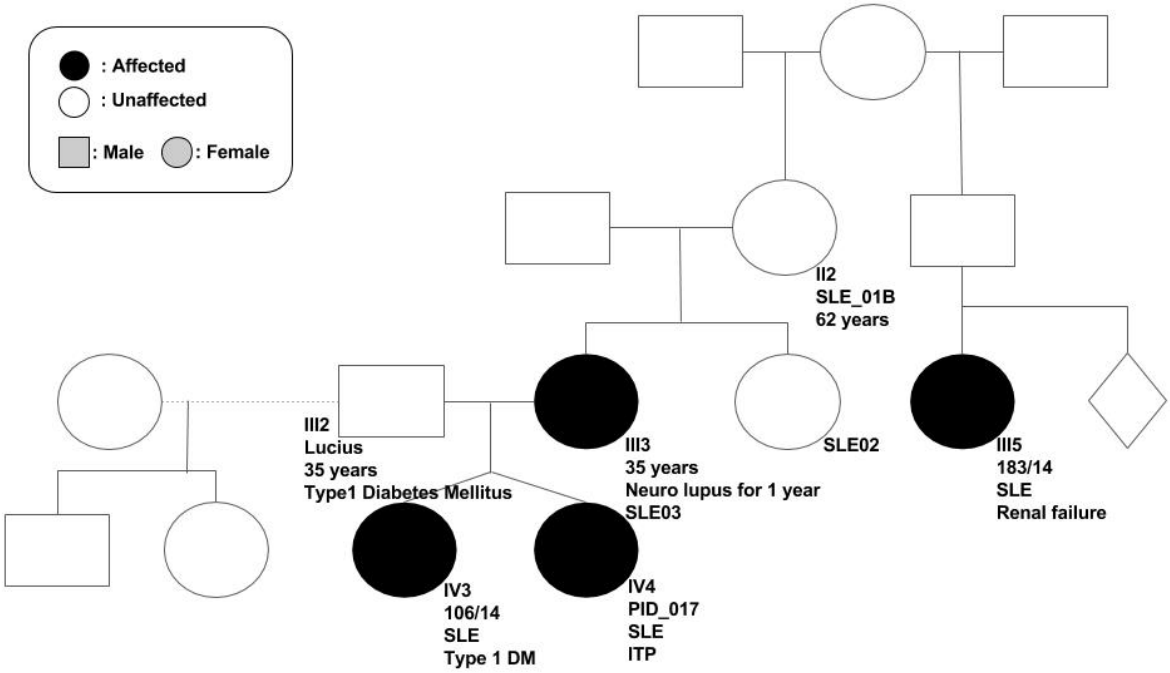

Figure 2.4: Family pedigree showing a familial history of SLE.

2.5 Individuals sequenced

Due to financial constraints, only five exomes of this SLE family were sequenced. Some samples were not available for selection, such as for the first generation. In this study, one negative control was utilized. The control was related to the affected individuals, which allowed for the exclusion of a large number of overlapping candidate variants. A pair of monozygotic twins and a distantly related affected family member also formed part of this study. Only one twin was sequenced, because the twins are identical and both are affected, so they are expected to have the same exome data apart for a couple of *de novo* mutations which are not of interest because they are unlikely to contribute to SLE. Sequencing the distantly related affected member allows the shared candidate list to be substantially reduced; because the affected cousins are expected to share the aetiological variant/s, but are likely to share only a small percentage of total variants. One can calculate this based on how much DNA two cousins are likely to share (Figure 2.5); and in this case the cousin's parents only shared a mother but had different fathers, further increasing their genetic diversity.

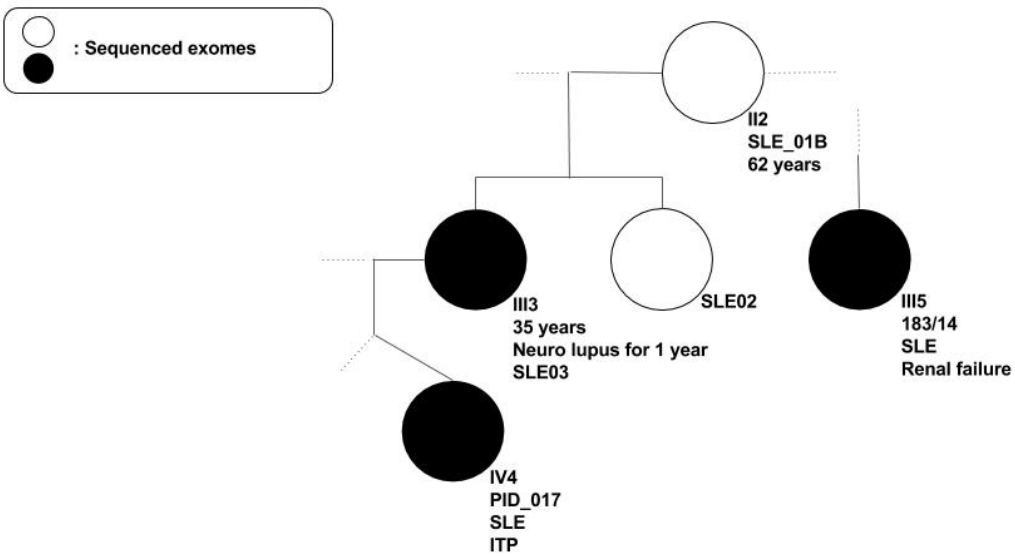

Figure 2.5: Individuals sequenced. Affected: SLE03, PID_017 and 183/14. Unaffected: SLE_01B and SLE02.

2.6 Inheritance models defined for the analysis

SLE is a complex disease known to have multiple triggers and it is not clear from the existing pedigree whether the mode of inheritance is dominant or recessive. Firstly, although it seems clear that this family has a strong inheritance pattern for SLE, the pedigree is not extended enough to establish with confidence the true frequency of disease presentation in each generation. Given the limited family information available and the paucity of data that might indicated the mode of inheritance for the SLE phenotype, several different inheritance models were explored for this study. Four inheritance models were proposed for the analysis, and these are referred to as the susceptibility model, tipping point model, protective mutation model and the asymptomatic model. These models are outlined below.

2.6.1 Susceptibility model

The premise:

This model is based on a premise that all family members directly descended from, and including, the mother of SLE_01B have a background of SLE susceptibility variants that predispose them to SLE, but that are not sufficient on their own for the disease phenotype to be expressed. For instance, SLE02 may be predisposed to SLE, i.e. susceptible, but does not have a disease phenotype (Figure 2.6.1). The model assumes that all individuals share a proportion of variants with a low minor allele frequency (MAF), as a consequence of the relatedness of these individuals; and that a proportion of these variants may be associated with a predisposition, or susceptibility to developing SLE.

The strategy:

For this susceptibility model, in order to identify variants that may predispose all the related family members to having SLE, all sequenced individuals were considered as susceptible to

SLE even if they did not present with the disease phenotype. Shared minor alleles that occurred in all sequenced individuals were identified as candidate pathogenic variants related to SLE when they are found to fall in genes that have been previously associated with SLE pathogenesis.

The output variant list:

The variant list defined for this model is a list of rare minor alleles *(MAF<1%)* that are shared by all individuals in the family, and are also associated with genes that have been implicated in SLE previously. These variants are candidates for susceptibility to SLE although they may not be sufficient for presentation of disease.

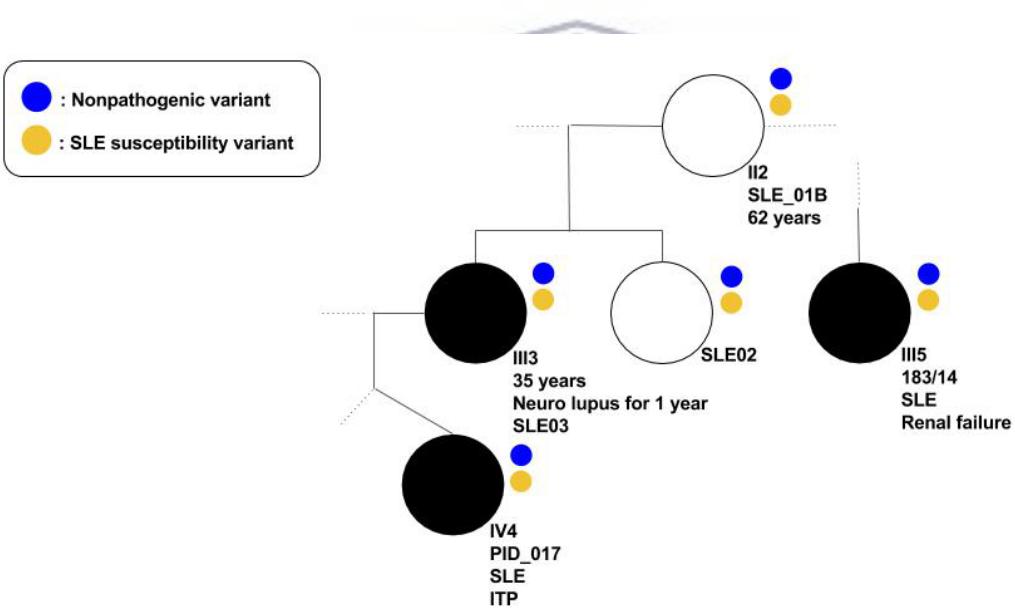

Figure 2.6.1: Susceptibility model.

2.6.2 Tipping point model

The premise:

This model is based on a premise that all affected family members from, and including, the mother of SLE_01B, have a background of SLE susceptibility that may predispose them to SLE; but only those presenting with the disease phenotype additionally have SLE tipping point variants that result in the expression of the disease, that are sufficient on their own for the disease phenotype to be expressed in the context of general susceptibility (Figure 2.6.2). The model assumes that all affected individuals share such tipping point variants with a low MAF, as a consequence of the relatedness of these affected individuals; and that these tipping point variants may be associated with the presentation of the SLE phenotype.

The strategy:

For this tipping point model, in order to identify tipping point variants that tip the scales for SLE presentation in affected members, shared minor alleles that occurred in affected sequenced individuals but not unaffected individuals were identified as candidate pathogenic variants related to SLE when they are found to fall in genes that have been previously associated with SLE pathogenesis.

The output variant list:

The variant list defined for this model is a list of rare minor alleles *(MAF<1%)* that are shared by all affected individuals in the family, and are also associated with genes that have been implicated in SLE previously. These variants are candidates for presentation of the SLE phenotype, and may be sufficient for presentation of disease.

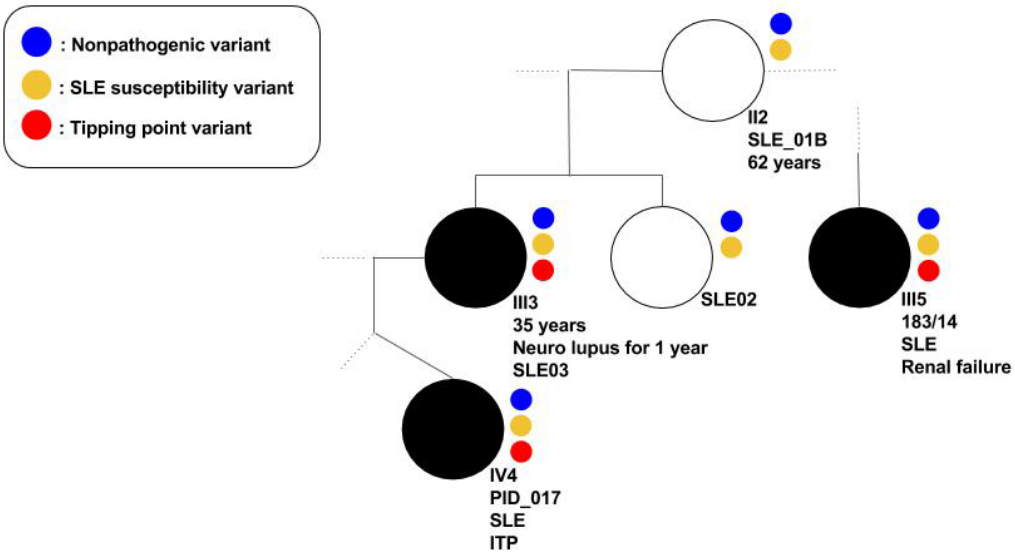

Figure 2.6.2: Tipping point model.

2.6.3 Protective mutation model

The premise:

This model is based on a premise that all family members have a background susceptibility to SLE, but that unaffected family members directly descended from the mother of SLE_01B have SLE-protective mutation variants that harbour a protective effect against presentation of the SLE disease, that are sufficient on their own for the disease phenotype not to be expressed (Figure 2.6.3). The model assumes that all unaffected individuals share a proportion of protective mutation variants with a low MAF, as a consequence of the relatedness of these unaffected individuals; and that a proportion of these protective mutation variants may prevent SLE disease presentation.

The strategy:

For this protective mutation model, in order to identify protective mutation variants that harbour a protective effect against SLE disease presentation in unaffected members, shared

minor alleles that occurred in unaffected sequenced individuals were identified as candidate disease-preventing variants related to SLE when they are found to fall in genes that have been previously associated with SLE pathogenesis.

The output variant list:

The variant list defined for this model is a list of rare minor alleles *(MAF<1%)* that are shared by all unaffected individuals in the family but not affected individuals, and are also associated with genes that have been implicated in SLE previously. These variants are candidates for resistance to SLE and may be sufficient for the disease not to manifest itself.

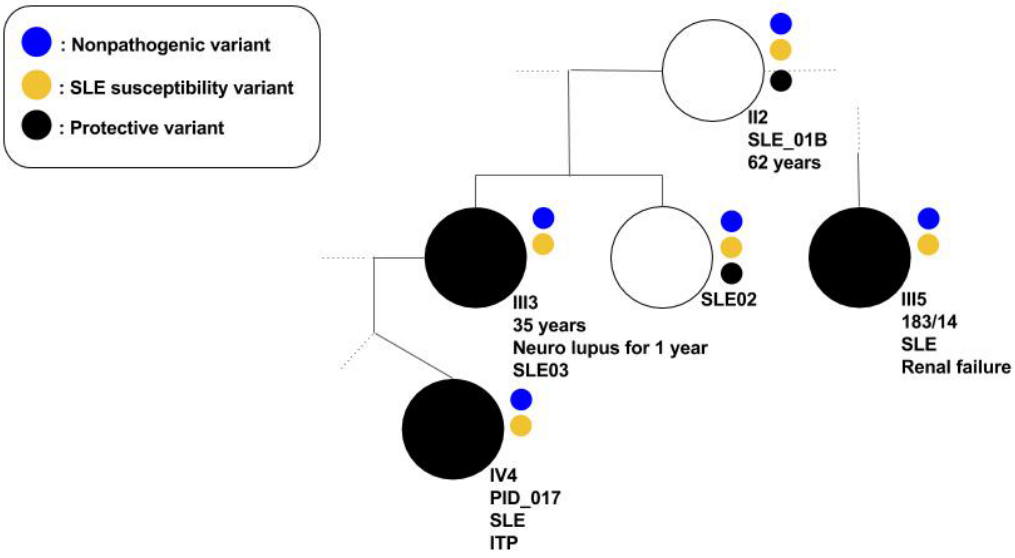

Figure 2.6.3: Protective mutation model.

2.6.4 Asymptomatic model

The premise:

This model is based on a premise that the unaffected female (SLE_01B) is an asymptomatic carrier of a single recessive mutant allele. The model assumes that this unaffected female carried a recessive mutant allele with a low MAF that causes SLE. If the variant were also

present as a recessive allele in those who married into the family, it could have been passed down to those affected as a homozygous mutation, such that the clinical phenotype for SLE susceptibility became apparent.

The strategy:

If this were an autosomal recessive inheritance, the disease variant would need to be:

- The variant must have low MAF.
- The minor allele must be heterozygous in the unaffected (Aa).
- The minor allele must be homozygous in all affected (aa).

For this asymptomatic model, in order to identify a single recessive mutant allele in the unaffected grandmother that has been passed down to the affected individuals, shared minor alleles that occurred in the unaffected mother and the affected sequenced individuals were identified as candidate variants related to SLE when they are found to fall in genes that have been previously associated with SLE pathogenesis.

The output variant list:

The variant list defined for this model is a list of rare minor alleles *(MAF<1%)* that are shared by one unaffected and all affected individuals in the family, are heterozygous in the unaffected and homozygous in the affected, and are also associated with genes that have been implicated in SLE previously. These variants are candidates for SLE susceptibility and may be sufficient for the disease to manifest itself.

Discussion

Family studies provide numerous opportunities for the investigation and interpretation of unidentified genetic variation of various types underlying complex diseases. Family studies may enable the discovery of rare and low frequency variants, and the identification of their associations with complex diseases, because predisposing variants will be present at much higher frequency in affected relatives of an index case. Family studies also allows for the investigation of parent-of-origin-specific effects. If not appropriately accounted for, such effects could mask associations and reduce the proportion of heritability. High-density SNP data in pedigrees can be utilized to localize predisposition genes. Family studies may also be useful in identifying gene-gene interactions, because affected relatives are more likely to share two nearby epistatic loci in linkage disequilibrium that would be unlinked in unrelated individuals (MacLean *et al*., 1993; Zhao *et al*., 2006).

An extreme and clearly inherited phenotype, such as the one in this family, may not represent exactly how SLE occurs in individuals without familial SLE. However, researching such unusual families can help to understand the molecular pathways underlying the disease phenotype; and this can give insights into more general mechanisms of disease presentation for SLE. Exploring the different possible ways in which genetic variants may contribute to SLE in this family has the potential to lead to wider insights about the disease mechanism, in addition to possibly assisting the family to better understand and manage their own illness. With such a limited number of exomes for the analysis, it is not possible to define the mode of inheritance with certainty. Instead, it is important to use the combinations of distant and close relationships, along with affected and unaffected status, to try and reduce the size of the list of potential candidates. Considering a variety of ways in which the inheritance pattern might manifest, and in which candidate variants might impact the disease phenotype, is important in order to prevent a premature restriction of the pool of possible aetiological

variants for this disease phenotype. Even with the limited number of exomes sequenced, there are multiple lines of evidence that can be used to restrict the candidate variant list. Sharing/not-sharing of alleles between family members is one type of evidence; and allele frequency in the general public is another whereby only rare alleles will be selected as potential candidates. Also, implication in SLE mechanisms by prior research studies will be used as a filtering step to narrow down the list of candidate variants that may be relevant in this particular family.

The proposed asymptomatic model is highly unlikely, because the chance of both branches of the family (the two cousins) both getting the second alleles from members who married into the family is highly unlikely. We do not have enough information on the extended family to determine the probability of SLE presenting in each generation. Furthermore, this model was considered because we did not know anything about the family's social, cultural or demographic background, and this kind of inheritance might be possible – however small the chance - in a very small, closed community or a cultural environment that encourages some level of consanguineous marriage.

Particularly if we consider that there might be an underlying susceptibility to SLE throughout the whole family, as proposed in the susceptibility model, such a recessive variant may not be pathogenic in the general population when homozygous, but becomes pathogenic when homozygous in this particular family because of the already existing genetic susceptibility to SLE. So it may not need to be a super-rare or new mutation to cause SLE if it occurs on the background of an existing susceptibility to SLE – and if this is the case it might explain how the second allele has been brought in to the family through multiple individuals who have married family members.

Chapter 3: Materials and Methods

3.1 Case Studies

In order to pinpoint shared causative genetic variants allegedly associated with SLE, we implemented WES in five SLE patients upon obtaining ethical approval from the University of Cape Town (Ethical Clearance number: HREC REF: 092/2014). Cases were related and had a familial history of SLE. Three patients were affected with SLE. One monozygotic twin affected with immune thrombocytopenia (Sample ID: PID_017), their mother affected with neuro lupus (Sample ID: SLE03) and the mother's cousin affected with lupus nephritis (Sample ID: 183/14). The other two individuals were unaffected. The grandmother of the twins (Sample ID: SLE_01B) and the twins' mother's sister (Sample ID: SLE02). Blood DNAs were harvested from all five patients using a PaxGENE DNA protocol. These samples were then shipped to Otogenetics Corporation (Atlanta, USA; www.otogenetics.com) for WES.

3.1.1 Whole-Exome Capture and Sequencing

The acquired DNA (2µg), from each sample was sheared, 1µg of that was used for library prep of which 800-1200 nM of the library was then exposed to WES. WES was carried out using the Agilent AV5 exome (51Mb) capture library kit according to the manufacturer's instructions (Otogenetics: http://www.otogenetics.com/). Paired-end exome sequencing was done using the Illumina HiSeq2500 platform at 50x read coverage. The sequencing run generated FASTQ files consisting of between 40 000 000 and 55 000 000 short reads per sample with an average sequence length of 106 bp. These files were then sent to the South African National Bioinformatics Institute for the analysis described further below.

3.1.2 Sequence read quality control analysis

It is essential to perform quality control checks on raw sequence data coming from high throughput sequencing pipelines before doing any further analysis. FastQC is a tool which affords a modular set of analyses thus giving a rapid idea on whether or not the data has any problems (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). The main function of FastQC is to inspect reads for base quality and distribution, duplicates, length and over-represented k-mers (http://chipster.csc.fi/manual/fastqc.html).

FastQC v 0.11.2 was used to examine the reads. All datasets had high Phred quality scores above 30. This meant that the probability of incorrect base calling was 1 in 1000 with base call accuracy being 99.9%. No read trimming was required (Figure 3.1.2).

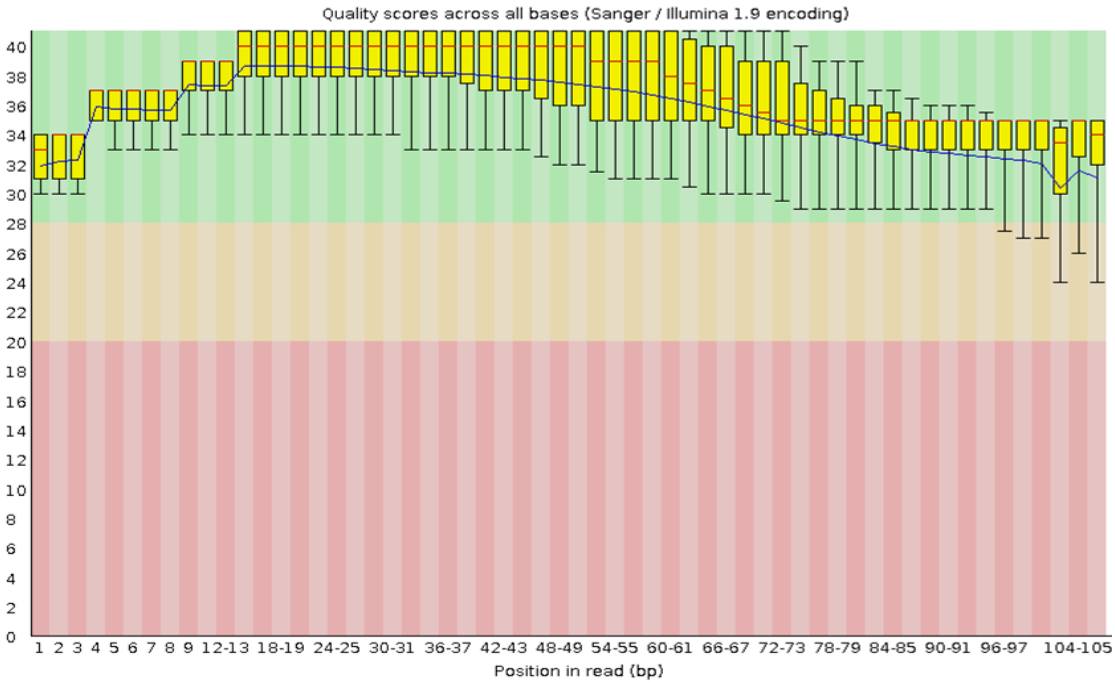

Figure 3.1.2: Quality control results for one of the sequenced samples (SLE02). The quality scores show per base quality for each sequenced base position across all reads in this sample. As expected the quality of the bases was high in the middle positions of the reads and decreased towards the end as well as the beginning of the reads. Overall, per- base quality was high indicating good quality data.

3.1.3 Sequence Alignment

The high-quality sequence reads per sample were grouped for further downstream analysis with GATK then mapped to the human reference genome hg19.nix, using NovoAlign v 3.01.00. NovoAlign is a short read sequence aligner which maps sequence reads across an indexed set of reference sequences and outputs a SAM file. This tool uses a repetitious search algorithm to find the best alignment and any other alignments with related scores (http://www.novocraft.com/userfiles/file/Novocraft.pdf). In total more than 99% of sequence per sample was uniquely mapped to the genome with 1% unmapped (Table 3.1.3).

Table 3.1.3: Summary of mapping statistics for exome sequenced samples. The mapped data is the sum of read bases that aligned to the reference genome.

Sample	Total reads	Mapped	Percent (%)
183/14	48940226	48840053	99.76
PID_017	39986248	39912595	99.77
SLE02	48973760	48867371	99.74
SLE_01B	52101350	51970556	99.71
SLE_03	54144212	54030855	99.75

A typical workflow for exome sequencing data analysis and variant calling with GATK is shown in figure 3.1.3.

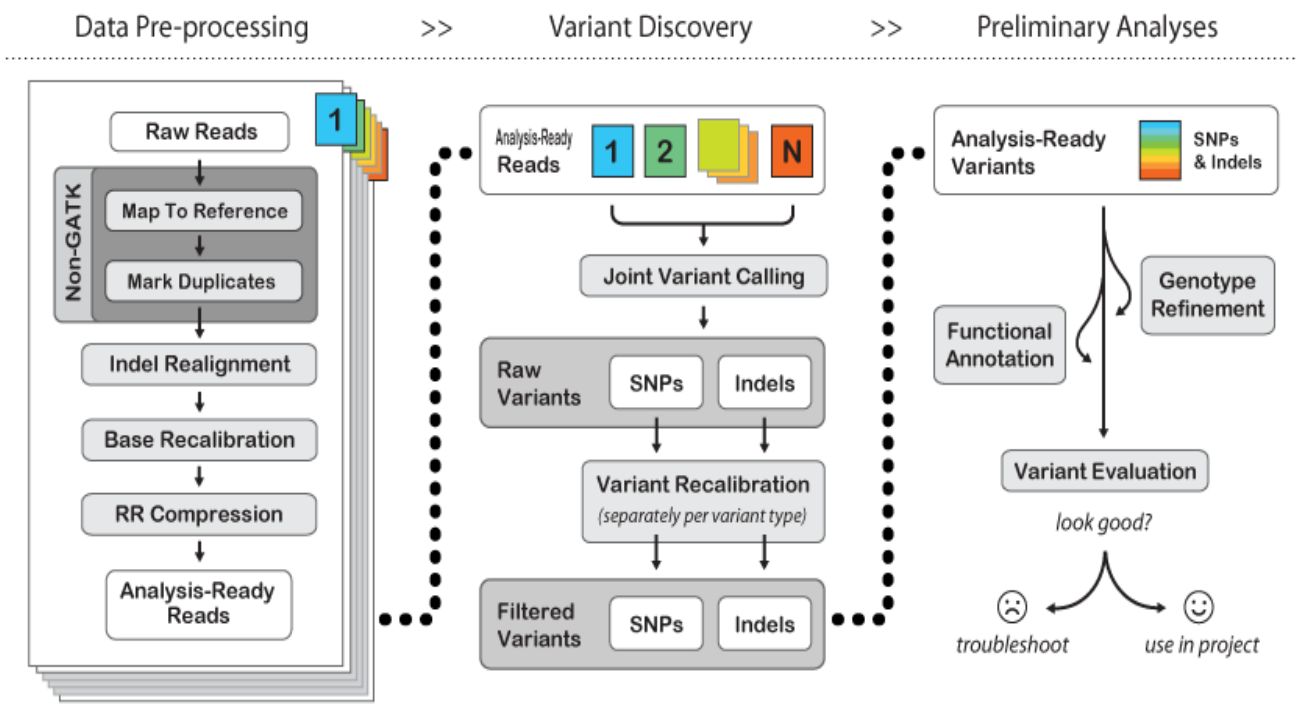

Figure 3.1.3: Workflow for WES data analysis and variant calling with GATK (http://weallseqtoseq.blogspot.co.za/2013/10/gatk-best-practices-workshop-data-pre.html).

3.1.4 Conversion of SAM to BAM format and Marking of PCR duplicate reads

The output SAM files in (3.1.3) were converted to sorted BAM files using Picard-Tools v 1.115. This tool was also utilized to mark PCR duplicates producing marked.bam files. PCR duplicates arise during DNA prep methods. Marking them is essential because duplicates cause biases which might skew variant calling results

(https://www.broadinstitute.org/gatk/events/slides/1212/GATKwh0-BP-1-Map_and_Dedup.pdf).

3.1.5 Local realignment around known indels

The marked.bam files were realigned around known indels using GATK's (v 3.2.2) RealignerTargetCreator tool, producing marked.realigned.bam files. GATK is a software package for analysis of high-throughput sequencing data

(https://www.broadinstitute.org/gatk/). The local realignment process is designed to consume one or more BAM files and to locally realign reads such that the number of mismatching bases is minimized across all the reads. This step improves the original alignment and it also uncovers hidden indels in reads thus eliminating potential false positive SNP's

(https://www.broadinstitute.org/gatk/events/slides/1212/GATKwh0-BP-2-Realignment.pdf).

3.1.6 Base Quality Score Recalibration

The marked.realigned.bam files were then subjected to recalibration using GATK's BaseRecalibrator tool, producing marked.realigned.recalibrated.bam files. This step assigns accurate quality scores to each sequenced base because the quality scores issued by the sequencer are sometimes inaccurate and biased. If not corrected then this will result in bad calls

(https://www.broadinstitute.org/gatk/events/slides/1503/GATKwh6-BP-3-Base_recalibration.pdf).

3.1.7 Variant Calling

All marked.realigned.recalibrated.bam files for each sample were pulled together and variants were called using GATK's HaplotypeCaller which outputs a rawSNP.vcf (variant call format) file. HaplotypeCaller calls SNP's and indels simultaneously via local re-assembly of haplotypes in an active region. The program determines which regions of the genome it needs to operate on, based on the presence of significant evidence for variation

(https://www.broadinstitute.org/gatk/guide/tooldocs/org_broadinstitute_gatk_tools_walkers_haplotypecaller_HaplotypeCaller.php).

3.1.8 Variant Quality Score Recalibration (VQSR) for SNP's

Recalibration parameters for SNP's were prepared by setting the mode in the command line to SNP. The rawSNP.vcf file was recalibrated against resource call sets (HapMap, 1000Genomes, and Single Nucleotide Polymorphism Database (dbSNP) in order to build a SNP recalibration model, using GATK's VariantRecalibrator tool. This step produced two output files namely: outputSNP.recal and OutputSNP.tranches. The purpose of Variant Quality Score Recalibration (VQSR) is to calculate a new quality score that is apparently super well calibrated known as the variant quality score log-odds (VQSLOD). This new score allows SNP's to be filtered in a way where a balance between sensitivity (trying to discover all the real variants) and specificity (trying to limit the false positives that creep in when filters get too lenient) are maintained

(http://gatkforums.broadinstitute.org/gatk/discussion/39/variant-quality-score-recalibration-vqsr).

The output files from VQSR (outputSNP.recal and OutputSNP.tranches), along with the rawSNP.vcf file served as an input for GATK's ApplyRecalibration tool. The output from this step was a recalibrated SNP file (output.recalibratedSNP.vcf).

The ApplyRecalibration tool implements the second pass in a two-stage process known as VQSR. The first pass is carried out by the VariantRecalibrator tool. In short, the first pass constructs a Gaussian mixture model by looking at the distribution of annotation values over a high-quality subset of the input call set and then scoring all input variants according to the

model. On the other hand, the second pass entails the filtering of variants based on score cutoffs identified in the first pass.

Utilizing the tranche and recalibration files generated during the first pass, the ApplyRecalibration tool inspects each variant's VQSLOD value and determines which tranche it falls in. Variants in tranches that fall beneath the specified truth sensitivity filter level will have their FILTER field annotated with the analogous tranche level. This will lead to a call set that is filtered to the chosen level but preserves the information needed to increase sensitivity if necessary

(https://www.broadinstitute.org/gatk/guide/tooldocs/org_broadinstitute_gatk_tools_walkers_variantrecalibration_ApplyRecalibration.php).

3.1.9 Variant Quality Score Recalibration (VQSR) for Indels

Recalibration parameters for Indels were prepared by setting the mode in the command line to Indel. The output.recalibratedSNP.vcf file in (3.1.8), served as input for Indel recalibration. The output.recalibratedSNP.vcf was recalibrated against resource call sets (Mills, 1000Genomes, and dbSNP) in order to build an Indel recalibration model, using GATK's VariantRecalibrator tool. GATK's ApplyRecalibration tool was then used to recalibrate the Indels to the desired level in the call sets. The output from this step was a recalibrated Indel file (output.recalibratedINDEL.vcf)

(https://www.broadinstitute.org/gatk/guide/article?id=2805).

3.1.10 Variant filtration and the selection of high-quality variants

Parameters were set to filter out low-quality variants thus retaining high-quality variants. GATK's VariantFiltration tool was used to filter output.recalibratedINDEL.vcf, producing a filtered_SNPs.vcf file. The VariantFiltration tool is devised for hard-filtering variant calls

based on certain criteria. Records are hard-filtered by alternating the value in the FILTER field to something other than PASS

(https://www.broadinstitute.org/gatk/guide/tooldocs/org_broadinstitute_gatk_tools_walkers_filters_VariantFiltration.php).

The filtered_SNPs.vcf file served as input for the selection of high-quality variants. GATK's SelectVariants tool was used to extract high-quality variants that passed all filters, generating a Combined.FCKD.HC.snps.indels.recalDone.filteredDone.passed.vcf. In total, 91 467 variants were filtered and of these, 85 689 variants were of high quality. These 85 689 variants were first annotated using Ensembl's Variant Effect Predictor (VEP) tool to see the coding consequences. As a result, VEP showed a total of 70.8 % non-synonymous and 29.1 % synonymous variants were obtained. Also, 4.7 % frameshift variants were identified, 2.1 % of them stop gain, 2.3 % stop loss, 0.4 % inframe deletions and 0.2 % inframe insertions were identified (Figure 3.1.10).

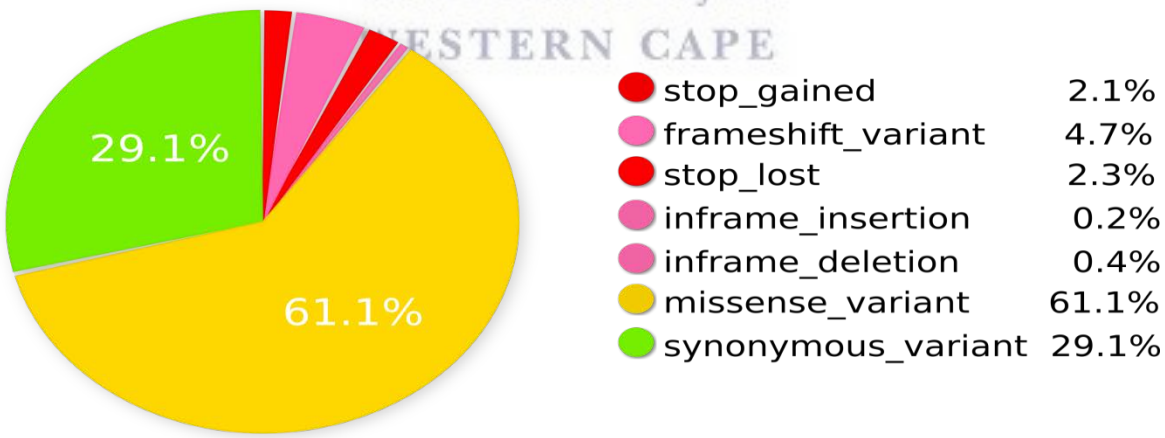

Figure 3.1.10: Coding consequences of variants. Variants were categorised using Ensembl's VEP tool into "stop-gained" a sequence variant whereby at least one base of a codon is changed, resulting in a premature stop codon, leading to a shortened transcript; "frameshift variant" a sequence variant which causes a disruption of the translational reading frame, because the number of nucleotides inserted or deleted is not a multiple of three; "stop-lost" a codon variant that changes at least one base of the canonical start codon; "inframe insertion" an inframe non synonymous variant that inserts bases into the coding sequence; "inframe deletion" an inframe non synonymous variant that deletes bases from the coding sequence; "missense variant" a sequence variant, that changes one or more bases, resulting in a different amino acid sequence but where the length of the protein is preserved; "synonymous variant" a sequence variant where there is no resulting change to the encoded amino acid.

3.1.11 Variant Functional Annotation

Further variant annotation was done using the Variant Annotation Analysis and Search Tool (VAAST) v 1.0.4. VAAST is a software suite for disease gene discovery from NGS data (Figure 3.1.11). The fundamental component of VAAST is a probabilistic disease-gene finder that links amino acid substitution and allele frequency information to search for and prioritize genes harboring damaging alleles. VAAST incorporates robust models of cross-species sequence conservation, which further improve the accuracy in differentiating between benign and disease-causing variation (Yandell *et al*., 2011). The basic inputs to VAAST consist of: (1) a set of target (case) variant files in either VCF or Genome Variation Format (GVF) format; (2) a set of background (control) variant files in either VCF or GVF format; (3) a set of features to be scored, usually genes - this file should be in GFF3 format, and (4) a multi fasta file of the reference genome (Yandell *et al*., 2011). An initial step in any VAAST analysis is to annotate variant files, followed by the selection and analysis of variants.

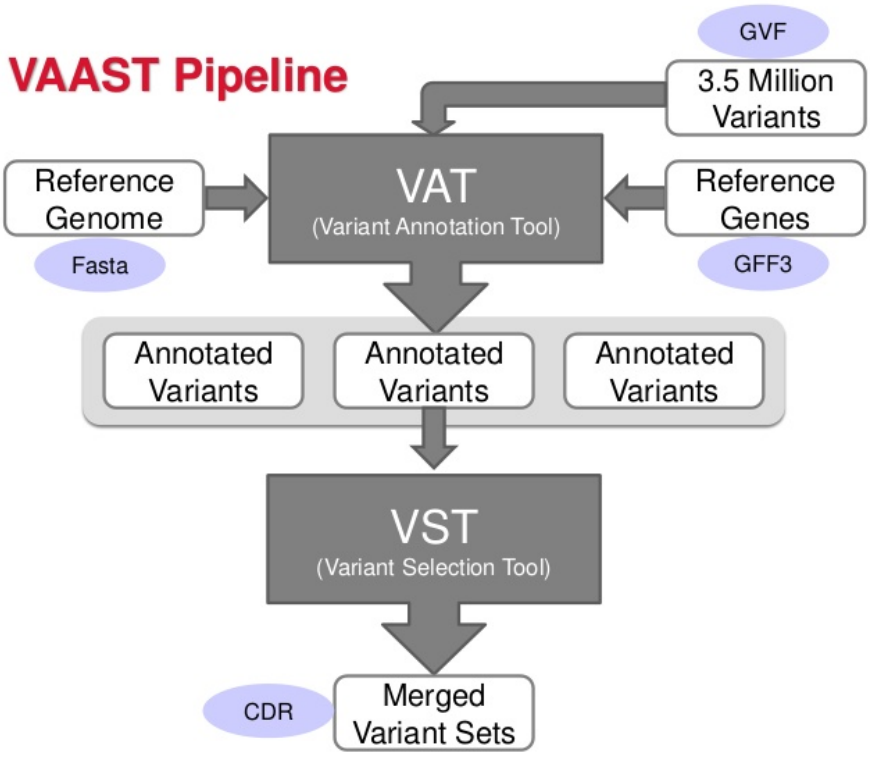

Figure 3.1.11: A typical VAAST workflow (Moore, 2011). VAT requires three files as input: (1) GVF - a file containing the sequence alterations that will be annotated; (2) GFF3 - a file containing sequence features (gene models and possibly other features); (3) FASTA - a file containing the genome's sequence. VAT annotates the functional impact of each sequence alteration (variant) in the input GVF file when it overlaps a sequence feature in the provided GFF3 file. These annotations are given with a Variant_effect attribute added to the final column of the GVF line. The annotated GVF line with the added Variant_effect attribute provides details of the impact of the sequence alteration on sequence feature. VST condenses variant files from multiple individuals into a single condenser (CDR) file. VST applies set operations across the individual input GVF files. These operations include union (U), intersection (I), left relative complement (C), and symmetric difference (D). VST also includes a shared (S) operation that specifies a cut-off to an intersection style operator to exclude or include variants based upon the number of individuals that share them.

The Combined.FCKD.HC.snps.indels.recalDone.filteredDone.passed.vcf was converted to GVF using vaast_converter tool, producing a gvf_out file. The gvf_out file was then indexed using vaast_indexer tool, outputting 183/14.gvf, PID_017.gvf, SLE03.gvf, SLE02.gvf and SLE_01B.gvf. Each gvf, was sorted using vaast_sort tool and annotated with vaast_VAT (Variant Annotation Tool), outputting 183/14.vat.gvf, PID_017.vat.gvf, SLE03.vat.gvf, SLE02.vat.gvf and SLE_01B.vat.gvf. VAAST's VAT tool annotates variants for individual genomes for the effects they cause on genomic features.

3.1.12 Variant Selection

All annotated variants (vat.gvf), served as inputs for variant selection using vaast_VST (Variant Selection Tool). The variant selection was done based on the four inheritance models that were proposed in section 2.6.

3.1.12.1 Susceptibility model

Variants shared by three affected and two unaffected members were selected. The command line codes in tables 3.1.12.1 and 3.1.12.1.1 illustrates how this was done.

Table 3.1.12.1: Linux code for selection of variants shared by three affected members.

```
VST \
-o 'I(0,1,2)' \
-b hg19 \
$DATA/183/14.vat.gvf \
$DATA/PID_017.vat.gvf \
$DATA/SLE03.vat.gvf \
> $DATA/Shared_By_3Affected_Only_variants.cdr
```

Table 3.1.12.1.1: Linux code for selection of variants shared by two unaffected members.

```
VST \
-o 'I(0,1)' \
-b hg19 \
$DATA/SLE02.vat.gvf \
$DATA/SLE_01B.vat.gvf \
> $DATA/Shared_By_2Unaffected_members.cdr
```

A VAAST analysis was run on both Shared_By_3Affected_Only_variants.cdr and Shared_By_2Unaffected_members.cdr. There were 169 genes found to be statistically significant with p-values < 0.05 amongst the three affected members. Nine hundred and fifty genes were found to be statistically significant with p-values < 0.05 amongst the two unaffected members. VAAST uses permutations to calculate the statistical significance of the results. The number of permutations specified for all analysis was set to 1e5 (100 000), for multiple sample testing. A comparative analysis between these two gene lists, showed a gene overlap of a hundred and forty eight genes. These hundred and forty eight genes were

compared to 1277 known SLE genes curated from UniProt, ClinVar, Orphanet, GWAS as well as Catalog and Comparative Toxicogenomics Databases. All the above mentioned databases are stored in a database of gene-disease associations (DisGeNET). Variants were found in the *CD177*, *CD24*, *CD72*, *EDN1*, *HP*, *OR2T5*, *PDCD1LG2* and *PIK3CA* genes. The 148 genes were also compared to 733 differentially expressed (DE) SLE genes curated from a microarray meta-analysis study and the gene overlap was seven. Variants were found in the *CD177*, *GYPE*, *HP*, *HSD11B2*, *PDLIM1*, *PRSS1* and *SYNE2* genes.

3.1.12.2 Tipping point model

Variants shared by three affected members were selected. The command line code in table 3.1.12.2 illustrates how this was done.

Table 3.1.12.2: Linux code for selection of variants shared by three affected members.

```
VST \
-o 'C(I(0,1,2),U(3,4))' \
-b hg19 \
$DATA/183/14.vat.gvf \
$DATA/PID_017.vat.gvf \
$DATA/SLE03.vat.gvf \
$DATA/SLE02.vat.gvf \
$DATA/SLE_01B.vat.gvf \
> $DATA/Variants_Shared_By_3Affected_members.cdr
```

In the -o operation above the ‘C’ operation is left complement of cases (183/14.vat.gvf, PID_017.vat.gvf, SLE03.vat.gvf) relative to controls (SLE02.vat.gvf, SLE_01B.vat.gvf), the ‘I' operation is the intersection of all cases and the U operation is the union of all controls. A VAAST analysis was run on the Variants_Shared_By_3Affected_members.cdr. Sixteen genes were found to be statistically significant with p-values < 0.05. These sixteen genes were compared to 1277 DisGeNET gene list and the gene overlap was found to be one. A variant was found in the *STAT4* gene. The sixteen genes were also compared to 733 DE SLE genes and the gene overlap was zero.

3.1.12.3 Protective mutation model

Variants shared by two unaffected members were selected. The command line code in table 3.1.12.3 illustrates how this was done.

Table 3.1.12.3: Linux code for selection of variants shared by two unaffected members.

```
VST \
-o 'C(I(0,1),U(2,3,4))' \
-b hg19 \
$DATA/SLE02.vat.gvf \
$DATA/SLE_01B.vat.gvf \
$DATA/183/14.vat.gvf \
$DATA/PID_017.vat.gvf \
$DATA/SLE03.vat.gvf \
> $DATA/Variants_Shared_By_2Unaffected_members.cdr
```

In the -o operation above the 'C' operation is left complement of cases (SLE02.vat.gvf, SLE_01B.vat.gvf) relative to controls (SLE03.vat.gvf, 183/14.vat.gvf, PID_017.vat.gvf), the 'I' operation is the intersection of all cases and the U operation is the union of all controls. A VAAST analysis was run on the Variants_Shared_By_2Unaffected_members.cdr. A hundred and ninety-four genes were found to be statistically significant with p-values < 0.05. These hundred and ninety-four genes were compared to 1277 DisGeNET gene list and the gene overlap was six. Variants were found in the *C3*, *DLAT*, *ISG15*, *LILRA2*, *PIK3CD* and *TBC1D9* genes. The hundred and ninety-four genes were also compared to 733 DE SLE genes and the gene overlap was ten. Variants were found in the *AOC3*, *DTX3*, *ISG15*, *MORC1*, *MYH8*, *NYX*, *SERPINB13*, *SLC9A2*, *TBX6* and *TF* genes.

3.1.12.4 Asymptomatic model

Variants shared by the affected mother, twin and cousin, and unaffected grandmother were selected. The command line codes in tables 3.1.12.4 and 3.1.12.4.1 illustrates how this was done.

Table 3.1.12.4: Linux code for selection of variants of unaffected granny.

```
VST \
-o 'I(0)' \
-b hg19 \
$DATA/SLE_01B.vat.gvf \
> $DATA/Granny.cdr
```

Table 3.1.12.4.1: Linux code for selection of variants shared by affected mom, twin and cousin.

```
VST \
-o 'I(0,1,2)' \
-b hg19 \
$DATA/183/14.vat.gvf \
$DATA/PID_017.vat.gvf \
$DATA/SLE03.vat.gvf \
> $DATA/Mom_twin_cousin.cdr
```

A VAAST analysis was run on both Granny.cdr and Mom_twin_cousin.cdr. The granny had 930 genes that were found to be statistically significant having p-values < 0.05. The mom, twin, and cousin had 343 genes that were found to be statistically significant with p-values < 0.05. A comparative analysis between these two gene lists showed a gene overlap of 118 genes. These 118 genes were compared to 1277 DisGeNET genes. The resulting gene overlap was eight with variants found in the *CD177*, *CD24*, *EDN1*, *HP*, *MERTK*, *OR2T5*, *PDCD1LG2* and *TNFRSF10B* genes. The 118 genes were also compared to 733 DE SLE genes and the gene overlap was eleven with variants found in the *CD177*, *CYP4B1*, *GYPC*, *GYPE*, *HP*, *HSD11B2*, *IL36G*, *LRRC1*, *PDLIM1*, *PRSS1* and *RLN1* genes. A comparative analysis was also done between the 1277 DisGeNET and the 733 DE SLE gene lists and yielded a gene overlap of 60 genes. These 60 genes were set-aside for further analysis.

Chapter 4: Variant Prioritization

4.1 Background

The aim of variant prioritization is to create a well-organized ranking of observed genetic variation (James *et al*., 2016).

An individual human genome consists of 3-4 million variants, or locations that vary from the human reference genome. Due to the large number of variants, it is important to identify, filter and prioritise those with some kind of association with the researchers' target phenotype and to decrease these variants down to a manageable number. This can be done by means of various heuristics such as allele frequencies and their effects on protein functions (Cooper and Shendure, 2011; Goldstein *et al*., 2013). There is a need in the biomedical research community for a tool that is capable of filtering these millions of variants based on the most up-to-date annotations and utilizes the growing arsenal of genome analysis methods. The number of bioinformatics pipelines for analyzing NGS data is fast increasing. Nonetheless, a majority of these tools focus on processing raw sequence data to detect high confidence genomic variants rather than focusing on downstream analyses such as annotation-based variant filtering and statistical analysis (Lam *et al*., 2012; McKenna *et al*., 2010; Pabinger *et al*., 2014).

4.1.1 A seven-level filtration framework for variant prioritization using a Tool for Automated selection and Prioritization for Efficient Retrieval (TAPER) of sequence variants

The discovery of a single, credible disease-causing mutation for a particular disease is proving to be as problematic as finding a needle in a haystack. Furthermore, it has been well documented that every individual or pedigree will carry numerous so-called private mutations

that do not cause explicit disease (Majewski *et al*., 2011) but may not be documented in existing databases of human genomic variation. To overcome the challenge of finding pathogenic variants, a novel tool named Tool for Automated selection and Prioritization for Efficient Retrieval (TAPER) of sequence variants, was developed for prioritization of sequence variants from WES data. TAPER is composed of seven steps to filter and prioritize candidate variants across individual patients that have been subjected to WES (Glanzmann *et al*., 2016). These seven steps are illustrated as follows. (1) The submission of the processed VCF files to an online variant caller such as wANNOVAR, which allows for the annotation of functional consequences of genetic variants from high-throughput sequencing data. This procedure is performed through an independent submenu in TAPER. (2) Elimination of all synonymous variants and variants that do not cause frameshifts – synonymous variations are defined as codon substitutions that do not change the amino acid and are not likely to be the underlying cause for disease – for this reason these variants, along with those that do not cause frameshifts, are removed from the list of prioritized variants. (3) Elimination of variants in the 1000 Genomes Project (1KGP) that are found at a frequency of greater than 1% - any variant that is found in the 1KGP database at a frequency of 1% or less is regarded as being rare. It is hypothesized that rare variants are likely to cause disease and for this reason, variants with very low or no available frequency data are prioritized. (4) Elimination of variants in the Exome Sequencing Project 6500 with a frequency of greater than 1% - this second frequency based step is based on that of the 1KGP data in (3) above. Rare, possible disease-causing variants are likely to be at an extremely low frequency in this database and any variant with a frequency that is higher than 1% is removed from the list of interest. (5) Elimination of variants with negative Genomic Evolutionary Rate Profiling (GERP) scores - GERP scores are the conservation scores from the database for nonsynonymous SNPs functional predictions (Liu *et al*., 2011), higher scores are indicative of greater conservation,

scores > 0 are considered to be conserved. Highly conserved genes are believed to be more likely to have deleterious phenotype effects if they are dysregulated or structurally altered (Kumar *et al*., 2011), (6) Elimination of all variants with positive functional analysis through hidden Markov Models (FATHMM) scores - FATHMM scores are utilized to regulate species specific weightings for the predictions of functional effects of protein missense variants (Shihab *et al*., 2013). The utilization of FATHMM scores, have been proven to outperform the conventional prediction methods such as Sorting Intolerant from Tolerant (SIFT), Poly-Phen2 and MutationTaster (Shihab *et al*., 2013; Rackham *et al*., 2015). Positive FATHMM scores predict a tolerance to the variation while negative FATHMM scores predict intolerance to the variation, and is consequently regarded to be pathogenic. Following proof of concept analysis it was determined that the best possible cut-off value for the FATHMM score is 1.0. (7) Identification of related disorders for prioritized genes – the final step of TAPER concludes whether the genes of interest have been associated to any other disorders using database such as the DISEASES database. If any of the disorders are related or similar to the disease of interest, then that gene and variant will become the leading candidate for further analysis.

4.1.2 Strengths and weaknesses of the TAPER method

The TAPER method has its strengths and weaknesses (Table 4.1.2).

Table 4.1.2: Strengths and weaknesses of the TAPER method (Glanzmann *et al*., 2016).

Strengths	Weakness
TAPER varies significantly from other variant prioritization tools such as PhenIX and Exomiser, it does not require a hypothesis driven approach, meaning that phenotypic information about the disease of interest, inheritance patterns and knowledge of pathways are not required to aid in variant prioritization.	Given the common approach to WES data analysis through TAPER, one of the major drawbacks is the fact that TAPER can only be utilized once a VCF file has already been generated, meaning that a biomedical researcher is still largely dependent on bioinformatics capacity in order to perform quality control and sequence alignment on samples that have been sequenced.
When TAPER outputs are compared to other software packages, TAPER produces a list of variants that is more convenient due to minute numbers and thereby making variant selection for further analysis easier.	One more limitation of the TAPER program is the fact that cloud based storage is not yet possible. TAPER is unable to upload filtered results to data servers such as Dropbox or Google Drive.
TAPER is able to read pre-annotated variation data from several file formats and permits users to search, sort and sift through larger data sets, by using each of TAPER's seven functions independently.	Lastly, TAPER can only be used on a Windows operating system, which is restrictive to individuals who may use Linux or Macintosh (iOS) based systems.
Due to its known high discriminative power, the efficiency and accuracy of TAPER was tested using sets of data which contained known pathogenic mutations – this endorsed for the discovery of less stringent cut-off criteria which otherwise would have allowed for the loss of possible disease-causing variants.	–
Overall, TAPER enables groups of researchers, particularly those in resource constrained laboratories with partial bioinformatics capabilities and resources, to interpret and analyse sequence variation data, thereby making NGS technologies such as WES more feasible in these laboratory setups.	–

Methodology

4.2 Exome data analysis for variants

Variants shared by all affected, unaffected and those shared amongst all family members, that passed the VAAST filters were prioritized. These variants were used as query seeds for the identification of rare and novel variants associated with SLE. Characteristics of rare and novel exome variants are: Rare – nonsynonymous missense variants, MAF < 1%, amino acid change and deleterious. Novel – variants with no rsID and not registered in any of the known variant databases. Variant coordinates were manually entered and searched against three

databases. The databases used were Exome Variant Server (EVS), Exome Aggregation Consortium (ExAC) and dbSNP. A brief description of each database is given below.

The main objective of EVS is to identify novel genes and mechanisms contributing to disease by launching the application of next-generation sequencing of the protein-coding regions of the human genome across distinct, richly-phenotyped populations and to share these datasets and findings with the scientific community to broaden and enhance the diagnosis, management and treatment of disease (http://evs.gs.washington.edu/EVS/).

ExAC is an alliance of investigators intending to aggregate and integrate exome sequencing data from a wide range of large-scale sequencing projects, and to make summary data available for the wider scientific community. Datasets housed in ExAC stretch across 60,706 unrelated individuals sequenced as part of diverse disease-specific and population genetic studies (http://exac.broadinstitute.org/).

The dbSNP database was established by the National Center for Biotechnology Information as a Single Nucleotide Polymorphism database (Sherry *et al*., 1999). Since its initiation in September 1998, dbSNP served as a central, public repository for genetic variation data (Smigielski *et al*., 2000).

Variants reported in dbSNP, EVS and ExAC as rare missense variants with MAF < 1% were retained, as these variants were most likely to cause the disease phenotype, based on their known coding mutations. In order to help identify causal variants, each variant's protein prediction scores were calculated using PolyPhen2 v2.1.0 and SIFT v1.0.3 web servers. Further searches in dbSNP, EVS and ExAC showed that ten variants were novel as these were not registered in any of the known variant databases. Protein prediction scores were calculated for these novel variants using PROVEAN v1.1.3 web server.

Results

We identified ten novel variants one each in (*MYH8*, *NYX*, *SERPINB13*, *CD177*, *CD24*, *HSD11B2*, *MERTK* and *IL36G*), and two in *PRSS1*. These variants were not registered in any of the known variant databases such as dbSNP, EVS or amongst the 60,706 sequenced exomes housed in ExAC. The variants found in *NYX*, *SERPINB13* and *IL36G* were predicted to be deleterious. All cases in this family shared variants in the *CD177*, *HSD11B2* and *PRSS1* genes. The HGNC (HUGO Gene Nomenclature Committee) symbol of each gene carrying these variants was queried against 733 DE and 1277 DisGeNET SLE genes (Table 4.2).

Table 4.2: Genes carrying variants not registered in any of the known variant databases.

Gene	Variants	PROVEAN prediction	733 DE SLE gene list	1277 DisGeNET known SLE gene list
MYH8	(chr17:10302089;G->A; T->I)	tolerated	Yes	No
NYX	(chrX:41332869;C->T; R->W)	deleterious	Yes	No
SERPINB13	(chr18:61259660;C->G; L->V)	deleterious	Yes	No
CD177	(chr19:43865333;G->A; A->T)	not found	Yes	Yes
CD24	(chrX:2115466;T->A; T->S)	not found	No	Yes
HSD11B2	(chr16:67470633;G->A; M->I)	tolerated	Yes	No
MERTK	(chr2:112722780; T->C; M->T)	not found	No	Yes
IL36G	(chr2:113742568;T->G; L->R)	deleterious	Yes	No
PRSS1	(chr7:142460764;G->A; V->I)	tolerated	Yes	No
	(chr7:142460752;C->G; Q->E)	tolerated	Yes	No

Yes: gene in 733 DE SLE gene list or in 1277 DisGeNET known SLE gene list.

No: gene not in 733 DE SLE gene list or in 1277 DisGeNET known SLE gene list.

Our study also reported nineteen rare missense variants (Table 4.2.1), with MAF < 1% one each in (*STAT4*, *C3*, *ISG15*, *PIK3CD*, *TBC1D9*, *AOC3*, *DTX3*, *MORC1*, *SLC9A2*, *TBX6*, *TF*, *HP*, *SYNE2*, *MERTK*, *TNFRSF10B*, *GYPC* and *LRRC1*), and two in *LILRA2*. One of the

variants found in *LILRA2*, located on chr19:55098805; G->A had no SNP ID but had a reported allele frequency. The variants found in *STAT4*, *ISG15*, *TBX6* and *GYPC* were all predicted to be deleterious by both SIFT and PolyPhen2 algorithms. The three affected members shared a variant in the *STAT4* gene and the two unaffected members shared a variant in the *ISG15* gene.

Table 4.2.1: Variant summary data showing nineteen variants with MAF < 1%.

Gene	SNP rsID	p-value	Variant	MAF	VAAST variant score	PolyPhen2 pred.	SIFT pred.	Effect	Susceptibility model	Tipping point model	Protective mutation model	Asymptomatic model
STAT4	rs3024839	0.0003	chr2:191940982; T->C; I->V	0.0776%	17.47	damaging 0.925	deleterious 0.006	Missense, non-coding transcript exon	Not selected	Selected	Not selected	Not selected
C3	rs554587967	0.0013	chr19:6718386; C->T; R->H	0.0008%	24.52	Not found	Not found	Missense, non-coding transcript exon	Not selected	Not selected	Selected	Not selected
DLAT	rs11553595	0.045	chr11:111899635; A->G; Q->R	1.9%	14.45	benign 0.04	tolerated 0.552	Missense, 3' UTR, intron	Not selected	Not selected	Selected	Not selected
ISG15	rs139516378	0.0003	chr1:949511; G->A; G->S	0.005%	22.94	damaging 0.98	deleterious 0.008	Missense	Not selected	Not selected	Selected	Not selected
LILRA2	Not found in dbSNP/No rsID	0.03	chr19:55098805; G->A; G->R	0.0016%	11.15			Missense, non-coding transcript exon, intron	Not selected	Not selected	Selected	Not selected
	rs75028967	0.03	chr19:55087490; T->C; V->A	0.63%	8.49	benign 0.001	tolerated 0.762	Missense, intron	Not selected	Not selected	Selected	Not selected
	rs7249811	0.03	chr19:55087313; T->G; V->G	10.2%	1.62	benign 0.0	tolerated 1	Missense, intron	Not selected	Not selected	Selected	Not selected
PIK3CD	rs28730673	0.0007	chr1:9780045; C->T; R->C	0.0157%	20.70	benign 0.001	tolerated 0.299	Missense	Not selected	Not selected	Selected	Not selected

TBC1D9	rs368519500	0.003	chr4:141543870; C->T; V->M	0.0086%	24.52	benign 0.001	tolerated 0.11	Missense	Not selected	Not selected	Selected	Not selected
AOC3	rs201143286	0.000006	chr17:41004913; C->T; T->M	0.005%	22.54		deleterious 0.022	Missense	Not selected	Not selected	Selected	Not selected
	rs477207	0.03	chr17:41008373; G->A; G->S	1.4%	6.15	benign 0.034	tolerated 0.922	Missense	Not selected	Not selected	Selected	Not selected
	rs2228470	0.01	chr17:41003859; C->T; H->Y	1.44%	5.54	benign 0.001	tolerated 1	Missense	Not selected	Not selected	Selected	Not selected
DTX3	rs201505564	0.00003	chr12:58001023; G->T; R->L	0.001121%	20.06	Not found	Not found	Missense	Not selected	Not selected	Selected	Not selected
MORC1	rs17225637	0.00003	chr3:108780836; T->A; K->M	0.5%	7.92	benign 0.437	deleterious 0.029	Missense	Not selected	Not selected	Selected	Not selected
MYH8	No SNP found	0.00005	chr17:10302089; G->A; T->I		20.06				Not selected	Not selected	Selected	Not selected
NYX	No SNP found	0.00007	chrX:41332869; C->T; R->W		20.06				Not selected	Not selected	Selected	Not selected
SERPINB13	No SNP found	0.01	chr18:61259660; C->G; L->V		20.06				Not selected	Not selected	Selected	Not selected
SLC9A2	rs143663218	4.34e-09	chr2:103318873; G->A; R->H	0.0042%	16.25	benign 0.018	tolerated 0.351	Missense, non-coding transcript exon	Not selected	Not selected	Selected	Not selected
TBX6	rs56098093	1.61e-06	chr16:30100401; C->T; G->S	0.46%	12.45	damaging 1.0	deleterious 0	Missense	Not selected	Not selected	Selected	Not selected
TF	rs41295774	0.01	chr3:133476698; A->G; H->R	0.3%	8.67	benign 0.07	tolerated 0.414	Missense, 3' UTR	Not selected	Not selected	Selected	Not selected
CD177	No SNP found	0.002	chr19:43865333; G->A; A->T		41.30				Selected	Not selected	Not selected	Selected
CD24	No SNP found	4.34e-09	chrX:2115466; T->A; T->S		36.50				Selected	Not selected	Not selected	Selected
CD72	rs34791102	4.34e-09	chr9:35612978; G->A; P->L	5.3%	4.52	damaging 0.87	deleterious 0	Missense, non-coding transcript exon	Selected	Not selected	Not selected	Not selected

EDN1	rs5370	0.01	chr6:12296255; G->T; K->N	23.45%	8.82	damaging 0.454	deleterious 0.025	Missense	Selected	Not selected	Not selected	Selected
HP	rs470428	0.0007	chr16:72094682; A->G; T->A	0.04%	28.62		Not found	Missense, 3' UTR, non-coding transcript exon	Selected	Not selected	Not selected	Selected
OR2T5	rs1770043	4.34e-09	chr1:248651927; A->G; K->R	55.63%	48.43		Not found	Missense	Selected	Not selected	Not selected	Selected
PDCD1LG2	rs7854413	4.34e-09	chr9:5557708; T->C; I->T	15.49%	3.15	damaging 0.816	deleterious 0.016	Missense	Selected	Not selected	Not selected	Selected
PIK3CA	rs3729680	0.03	chr3:178927410; A->G; I->M	6.5%	9.32	benign 0.011	Not found	Missense	Selected	Not selected	Not selected	Not selected
GYPE	rs28721877	4.34e-09	chr4:144801662; C->T; G->E	43.29%	35.34		Not found	Missense	Selected	Not selected	Not selected	Selected
HSD11B2	No SNP found	4.34e-09	chr16:67470633; G->A; M->I		35.35				Selected	Not selected	Not selected	Selected
PDLIM1	rs2296961	0.006	chr10:97023630; T->C; N->S	25.7%	4.44	benign 0.001	tolerated 0.868	Missense, non-coding transcript exon	Selected	Not selected	Not selected	Selected
PRSS1	No SNP found	0.006	chr7:142460764; G->A; V->I		13.60				Selected	Not selected	Not selected	Selected
	No SNP found	4.34e-09	chr7:142460752; C->G; Q->E		10.82				Selected	Not selected	Not selected	Selected
SYNE2	rs375789929	4.34e-09	chr14:64469904; T->C; V->A	0.00083%	24.52	benign 0.0	tolerated 0.967	Missense, 5' UTR	Selected	Not selected	Not selected	Not selected
	rs11847087	4.34e-09	chr14:64520020; A->G; N->S	4.2%	3.86	benign 0.021	tolerated 0.212	Missense, 5' UTR	Selected	Not selected	Not selected	Not selected
MERTK	No SNP found	0.01	chr2:112722780; T->C; M->T		12.41				Not selected	Not selected	Not selected	Selected
	rs150870104	0.01	chr2:112686742; C->T; P->L	0.03%	7.92	benign 0.001	tolerated 0.244	Missense, intron	Not selected	Not selected	Not selected	Selected
TNFRSF10B	rs148653452	0.01	chr8:228800269; A->T; L->H	0.0041%	8.15	benign 0.1	tolerated 0.626	Missense, 3' UTR	Not selected	Not selected	Not selected	Selected

CYP4B1	rs4646491	0.002	chr1:47280884; C->T; R->C	15.28%	6.86	damaging 0.525	deleterious 0.025	Missense, synonymous, 3' UTR, non-coding transcript exon	Not selected	Not selected	Not selected	Selected
	rs2297809	4.34e-09	chr1:47282772; C->T; R->C	15.55%	6.56	damaging 1.0	deleterious 0.004	Missense, 3' UTR, non-coding transcript exon	Not selected	Not selected	Not selected	Selected
	rs2297810	4.34e-09	chr1:47280859; G->A; M->I	23.7121%	4.70	benign 0.153	deleterious 0.042	Missense, stop gained, 3' UTR, non-coding transcript exon	Not selected	Not selected	Not selected	Selected
GYPC	rs115201071	4.34e-09	chr2:127453543; T->C; I->T	0.23%	19.17	damaging 0.893	deleterious 0.027	Missense, non-coding transcript exon	Not selected	Not selected	Not selected	Selected
IL36G	No SNP found	4.34e-09	chr2:113742568; T->G; L->R		12.64				Not selected	Not selected	Not selected	Selected
LRRC1	rs200754582	4.34e-09	chr6:53787481; A->T; M->L	0.0033%	8.83		tolerated 0.289	Missense	Not selected	Not selected	Not selected	Selected
RLN1	rs113678308	4.34e-09	chr9:5339730; A->G; L->S	Variant not found in ExAC	12.64		tolerated 0.122	Missense	Not selected	Not selected	Not selected	Selected

Discussion

In the filtration and prioritization procedure of exome sequence variants, it is acceptable to allow a reasonably larger false positive rate at this step and diminish the chance of missing true causative variants because there are additional criteria subsequently employed to exclude the false positive variants.

So the above analysis may not sufficiently illustrate that causal mutation(s) for a complex disease can be easily detected by processing the sequenced data of only a small sample size. Nevertheless, these results suggest that the filtration and prioritization procedure utilized above can help dramatically downsize the number of candidate variants to a very small subset that is human-manageable; and is unlikely to filter out good candidates.

Our study reported four rare missense variants with MAF < 1% and predicted to be deleterious by both algorithms. However, there remains an uncertainty as to whether these variants are in actual fact deleterious, because some algorithms use common source data to make predictions and a poor call by one algorithm may then be replicated by other algorithms accordingly.

In future, more stringent MAF thresholds (say, 0.005 or even 0.0) could be incorporated into filtration and prioritization methods for complex disorders in order to exclude additional common variants or rare benign sequence variants; and this could further improve the fidelity of the results.

Chapter 5: Pathway Analysis of Prioritized Variants

5.1 Pathway analysis

In pathway analysis gene sets linked to biological pathways are tested for significant relationships with a phenotype. Primary data for pathway analysis is frequently sourced from genotyping or gene expression arrays, though in theory any data elements that could be mapped to genes or gene products could be used. Importantly, analysing genomic data through functionally-derived gene sets can disclose larger effects related to cellular processes and mechanisms that are otherwise concealed from gene- or SNP-based analysis. For instance, high-profile studies in breast cancer (Menashe *et al*., 2010), Crohn's disease (Wang *et al*., 2010), and type 2 diabetes (Zhong *et al*., 2010) revealed that functionally-related genes can communally influence disease susceptibility, even if individual loci do not display genome-wide significant association. As such, pathway analysis represents a potentially powerful and biologically-oriented bridge between genotypes and phenotypes.

Information derived from pathway analysis includes in depth and contextualized findings to help understand the mechanisms of the disease in question, discovery of genes and proteins associated with the etiology of a specific disease, prediction of drug targets, as well as understanding how to intervene therapeutically in disease processes and conduct targeted literature searches. Furthermore, data integration such as integrating diverse biological information from the scientific literature, knowledge databases, genome sequences, protein sequences, motifs and structures. Functional discovery by assigning functions to genes can also be retrieved from pathway analysis

(http://bioinformatics.mdanderson.org/MicroarrayCourse/Lectures09/Pathway%20Analysis.pdf).

5.1.1 General principles for gene set enrichment analysis

The analysis of functional genomics data from high-throughput experiments frequently comprises the calculation of potential functional relations between a gene or protein set of interest, e.g. DE genes in a microarray study and known gene/protein sets representing cellular processes and pathways. To identify and prioritize these alleged relations, a wide range of enrichment analysis tools have been developed in recent years. One such tool is EnrichNet, and is described here as an example that illustrates the principles and steps of pathway analysis clearly. EnrichNet generates a novel graph-based statistic, developed to manipulate information from molecular network structures linking two gene/protein sets, with a new interactive visualization of network sub-structures. This united network analysis and visualization allows direct molecular interpretation of how a user-defined set of genes/proteins is related to a gene/protein set of known function. Based on previous work that entails the merging of network and pathway analysis methods (Glaab *et al.*, 2010*a*, *b*) the integrated data sources (molecular interaction data, cellular pathway data and tissue-specific gene expression data) and analysis techniques (graph-based statistical analysis and force-directed layout generation for sub- networks) are designed to build on each other to provide a clearer and more detailed understanding of gene/protein set functional relations. A general workflow for EnrichNet involves the following: (1) Input - a list of 10 or more human gene or protein identifiers and the selection of a database of interest (Kyoto Encyclopedia of Genes and Genomes (Kanehisa *et al.*, 2006), BioCarta (Nishimura, 2001), WikiPathways (Pico *et al.*, 2008), Reactome (Joshi-Tope *et al.*, 2005), Pathway Interaction Database (Schaefer *et al.*, 2009), InterPro (Apweiler *et al.*, 2001) or Gene Ontology (GO) (Ashburner *et al*., 2000), from which reference gene/protein sets will be extracted, (2) Processing - after aligning the target and reference datasets onto a genome-scale molecular interaction network (two default networks are available, alternatively a user-defined network can be provided, but the

availability of sufficient interaction data for the alignment of the target and reference datasets has to be confirmed. Next, a network analysis procedure is applied, this procedure entails two basic steps: a procedure to score the distances between the mapped target gene set and reference datasets in the network using a random walk with restart algorithm and the comparison of these scores against a background model and (3) Output - a ranking table of the reference datasets (e.g. cellular pathways, processes and complexes) is produced, comprising their network-based association scores and tissue-specific association scores across 60 human tissues. For every pathway, a hyperlink allows the user to create an interactive graph-based visualization of the sub-network representing the analysed datasets in the molecular interaction network. The user can explore this network by zooming into it, searching and highlighting specific genes/proteins and retrieving additional annotations and topological information by clicking on a node of interest.

5.1.2 Ingenuity Pathway Analysis (IPA)

Ingenuity Pathway Analysis (IPA) is an all-in-one, web-based application that allows the user to analyze, integrate, and understand data resulting from gene expression, microRNA, SNP microarrays, metabolomics, proteomics experiments and small-scale experiments that produces gene lists. IPA searches for targeted information on genes, proteins, chemicals, and drugs, and builds interactive models of users experimental systems. IPA's data analysis and search abilities helps the user to understand the significance of the data, specific target, or candidate biomarker in the context of larger biological or chemical systems, supported by the Ingenuity knowledge base of highly structured, detail-rich biological and chemical findings (http://repository.countway.harvard.edu/xmlui/bitstream/handle/10473/4740/Ingenuity%20Pathwahttp://lsl.sinica.edu.tw/Services/Class/files/20111228.pdfys.pdf?sequence=1).

IPA curators extract knowledge from the literature for this database, which currently contains human, mouse and rat genetic information. The IPA database houses 13,600 human genes, 11,000 mouse genes and 6,600 rat genes (http://www.ingenuity.com/).

IPA utilizes the following databases as reference sets for data curation. GO, Entrez Gene and Pfam for synonyms, protein family and domains, Genomic Institute of the Novartis Research Foundation and Plasma Proteome for tissue and biofluid expression and location, Biomolecular Interaction Network Database, *Database* of Interacting Proteins, Munich Information Center for Protein Sequences, IntAct - molecular interaction database, Biological General Repository for Interaction Datasets, Molecular INTeraction database and Cognia for molecular interactions, TarBase, TargetScan and miRecords for miRNA/mRNA target databases and Online Mendelian Inheritance in Man and GWAS databases for gene to disease associations

(http://lsl.sinica.edu.tw/Services/Class/files/20111228.pdf).

5.1.2.2 Advantages and disadvantages of IPA

IPA is a commercial system that has been highly curated and is in extensive use across both commercial and academic research enterprises. A significant advantage of using IPA for pathway analysis is the highly curated evidences that are used to build gene interaction networks. Whilst IPA software can build indirect associations between genes, these are often more tenuous or remote indirect interactions that can be more difficult to define clearly. This disadvantage has been overcome in this study by modelling only the direct gene-gene interactions defined by the software.

5.2 Transfer data files into IPA

Three datasets (rsID's rare variants, HGNC novel variants and 733 DE 1277 DisGeNET known SLE genes overlap 60), were uploaded for analysis via QIAGEN's Ingenuity® Pathway Analysis (IPA®, QIAGEN Redwood City, www.qiagen.com/ingenuity). Flexible Format was utilized as the default format for data upload and 'Gene symbol – human HGNC and dbSNP' as the identifier types for the gene names and rsIDs. IPA's core analysis was run on all uploaded dataset files.

5.2.1 Gene Interaction Networks

Networks for all rsID's rare variants were selected, and under molecules, in-network the network with the largest number of genes was chosen and viewed. Under the build tool, trim was selected and under interactions (Default: Indirect), was highlighted for the removal of all indirect interactions.

5.2.2 Expansion of selected networks using the Grow Tool

The grow tool under the build tab was used to expand the rsID's rare variants network. Molecules which emerged to be nodes were identified. These nodes, along with molecules shown in white (molecules added in by IPA), were expanded. Parameters for the grow tool were set to include both direct and indirect interactions, and adding 10 molecules at a time. The dataset for expansion was fixed to my specified HGNC novel variants dataset. All genes in the pathway that added no value to the network were deleted. The resulting network was further expanded. Parameters were set to include direct interactions only, and adding all molecules. The dataset for this expansion was fixed to my specified 733 DE 1277 DisGeNET known SLE genes overlap 60 dataset.

5.2.3 Identification of Top Regulated Genes

From the cross-analysis of all pooled lists, six candidate genes carrying rare variants were found to interact with known and DE SLE-associated genes (Table 5.2.3).

Table 5.2.3: The six genes containing variants identified by my study were found to interact with known and DE SLE-associated genes.

Gene name	Description	Cellular process	Disease	Reference
phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit delta (*PIK3CD*)	*PIK3CD* is involved in the immune response. This gene encodes a class I PI3K found predominantly in leukocytes. Like other class I PI3Ks (p110-alpha p110-beta, and p110-gamma), the encoded protein binds p85 adapter proteins and GTP-bound RAS. class I PI3K proteins phosphorylates itself, and not p85 protein.	proliferation of B lymphocytes, T lymphocytes, embryonic cell lines, epithelial cell lines, kidney cell lines, pancreatic cancer cell lines, breast cancer cell lines, mast cells, CD4+ T-lymphocytes, fibroblast cell lines, prostate cancer cell lines, thymocytes	grade 1 and 2 follicular lymphoma, hypersensitive reaction, locally advanced HER2 negative breast cancer, lymphocytosis, metastasis, primary B-cell immunodeficiency, primary sclerosing cholangitis, psoriasiform dermatitis	Rommel *et al*., 2007; Okkenhaug *et al*., 2002
spectrin repeat-containing nuclear envelope protein 2 (*SYNE2*)	*SYNE2* encodes a nuclear outer membrane protein that binds cytoplasmic F-actin. This binding restrains the nucleus to the cytoskeleton and assists in the maintenance of the structural integrity of the nucleus. Numerous transcript variants encoding various isoforms have been found for this gene.	migration of the nucleus	epithelial cancer, adenocarcinoma, acute myeloid leukemia	Luxton *et al*., 2010; Aoki *et al*., 2007
ubiquitin-like modifier (*ISG15*)	*ISG15* encodes a ubiquitin-like protein that is conjugated to intracellular target proteins upon activation by interferon-alpha and interferon-beta. Multiple functions have been attributed to the encoded protein, including chemotactic activity towards neutrophils, the direction of ligated target proteins to intermediate filaments, cell-to-cell signaling, and antiviral activity during viral infections. Conjugates of this protein have been found to be noncovalently attached to intermediate filaments.	ubiquitination in embryonic cell lines, epithelial cell lines, kidney cell lines, hepatoma cell lines	prostate cancer, relapsing-remitting MS, SLE	Okumura *et al*., 2006; Werneke *et al*., 2011
Transferrin (*TF*)	*TF* encodes a glycoprotein with a proximate molecular weight of 76.5 kDa. It is believed to have evolved as a result of an ancient gene duplication event that led to the generation of homologous C and N-terminal domains each of which binds one ion of ferric iron. This protein functions as a transporter of iron from the intestine, reticuloendothelial system, and liver parenchymal cells to all proliferating cells in the body. This protein may also have a physiologic role as granulocyte/pollen-binding protein (GPBP) involved in the removal of certain organic matter and allergens from serum.	apoptosis of ovarian cancer cell lines, ovarian cancer cells	non-insulin-dependent DM, organismal death, osteoarthritis, ovarian cancer, panhypopituitarism - X-linked, plasma cell neoplasm, polycystic ovary syndrome, psoriasis, renal impairment, ulcerative colitis	Fassl *et al*., 2003; Kvasnicka *et al*., 2000

tumor necrosis factor receptor superfamily member 10b (*TNFRSF10B*)	The protein encoded by *TNFRSF10B* is a member of the TNF-receptor superfamily and contains an intracellular death domain. This receptor can be activated by tumor necrosis factor-related apoptosis inducing ligand (TNFSF10/TRAIL/APO-2L), and transduces an apoptosis signal. Studies with FADD-deficient mice suggested that FADD, a death domain containing adaptor protein, is required for the apoptosis mediated by this protein. Two transcript variants encoding various isoforms and one non-coding transcript have been found for this gene.	apoptosis of colorectal cancer cell lines, leukemia cell lines, breast cancer cell lines	alveolar rhabdomyosarcoma, arthritis, chronic hepatitis B, experimental autoimmune encephalomyelitis, head, and neck squamous cell cancer	Shiraishi *et al*., 2005; Lee *et al*., 2001
signal transducer and activator of transcription 4 (*STAT4*)	The protein encoded by *STAT4* is a member of the STAT family of transcription factors. In response to cytokines and growth factors, STAT family members are phosphorylated by the receptor associated kinases, and then form homo- or heterodimers that translocate to the cell nucleus where they act as transcription activators. This protein is required for mediating responses to IL12 in lymphocytes, and regulating the differentiation of T helper cells. Mutations in this gene may be associated with systemic lupus erythematosus and rheumatoid arthritis. Alternate splicing results in multiple transcript variants that encode the same protein.	production in T lymphocytes, Th1 cells, lymphocytes, mononuclear leukocytes, natural killer cells	collagen-induced arthritis, colitis, infection, experimental autoimmune encephalomyelitis, insulin-dependent DM, melanoma, melanoma cancer, SLE	Moser and Murphy, 2000; Adamson *et al*., 2009

5.2.4 Path Designer Mode for color coding genes

IPA's Path Designer tool was used to differentiate between all candidate genes, known and DE SLE genes. The fill color was adjusted to green for genes with rare variants, red for genes with novel variants, blue for genes with rare and novel variants and yellow for known and DE SLE genes (Figure 5.2.4).

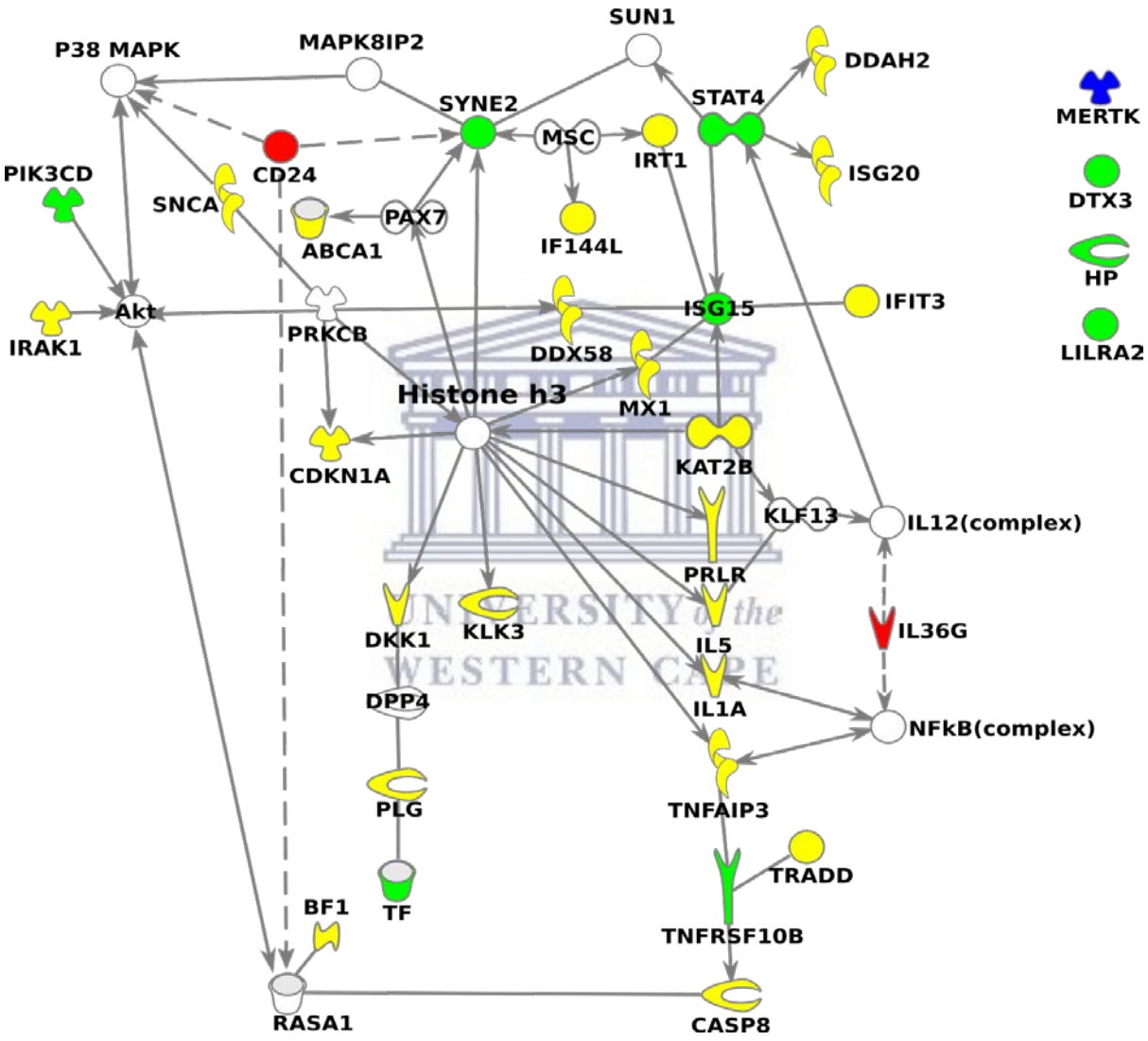

Figure 5.2.4: Ingenuity gene network analysis of the top six regulator genes (in green) interacting with known and DE SLE genes (in yellow). The genes in green contain rare variants, the genes in red contain novel variants and the blue genes contain rare and novel variants identified in my study.

5.2.5 Summary of IPA Core Analysis

Top Canonical Pathways		
Name	**p-value**	**Overlap**
Acute Phase Response Signaling	3.19E-04	1.8 % 3/169
Tec Kinase Signaling	3.25E-04	1.8 % 3/170
JAK/Stat Signaling	2.01E-03	2.4 % 2/83
FLT3 Signaling in Hematopoietor Progenitor Cells	2.11E-03	2.4 % 2/85
TR/RXR Activiation	2.79E-03	2.0 % 2/98

Upstream Regulator	p-value of overlap	Target Molecules
IL6	6.68E-06	HP, MERTK, STAT4
IFNA10	4.57E-05	ISG15, STAT4
IFNA21	4.57E-05	ISG15, STAT4
IFNA5	4.57E-05	ISG15, STAT4
IFNA7	4.57E-05	ISG15, STAT4
IFNA14	4.57E-05	ISG15, STAT4
IFNA6	4.57E-05	ISG15, STAT4
IFNA8	5.27E-05	ISG15, STAT4
IFNA16	5.27E-05	ISG15, STAT4
IFNA4	1.38E-04	ISG15, STAT4
4-methylnitrosoamino-1-3 -pyridyl-1-butanone	2.63E-04	HP, TF
SP600125	4.33E-04	ISG15, PIK3CD, TNFRSF10B
STAT3	4.66E-04	DTX3, HP, ISG15, TNFRSF10B
IFNA2	5.18E-04	ISG15, STAT4, TNFRSF10B
IFNA1/IFNA3	7.06E-04	ISG15, STAT4
EIF4B	7.32E-04	ISG15
TLR2/3/4	7.32E-04	STAT4
KPNB1	7.32E-04	TNFRSF10B
dihydrotanshinone I	7.32E-04	TNFRSF10B
SLC25A5	7.32E-04	TNFRSF10B

Figure 5.2.5: Summary of IPA Core Analysis. The core analysis details multiple parameters. The top five canonical pathways and top upstream regulators are shown here.

Discussion

IPA revealed direct interactions between six of our candidate genes with known and DE SLE genes (Figure 5.2.4). Here, we also report that some of our top six candidates mapped to known SLE pathways (Figure 5.2.5).

The *PIK3CD* gene mapped to the Acute Phase Response Signaling, Tec Kinase Signaling, Janus kinase/signal transducers and activators of transcription (JAK/STAT) Signaling, FLT3 Signaling in Hematopoietic Progenitor Cells pathways. *PIK3CD* mapped to the TR/RXR Activation pathway as well. The *STAT4* gene mapped to the Tec Kinase Signaling, JAK/STAT Signaling, FLT3 Signaling in Hematopoietic Progenitor Cells pathways; and the *TF* gene mapped to the Acute Phase Response Signaling pathway. The *TNFRSF10B* gene mapped to the Tec Kinase Signaling pathway. These pathways have been implicated in SLE, AD and inflammatory response. A brief description of each pathway is given below.

Acute Phase Response Signaling is a coordinated response to tissue injury, infection or inflammation. A remarkable feature of this response is the selection of acute phase proteins, which are involved in the restoration of homeostasis (Moshage, 1997).

Tec Kinase Signaling or Tyrosine kinases have roles in the control of cell survival, activation, and differentiation. They are also involved in signaling processes in cells known to be important in the pathogenesis of AD. Studies from animal models and SLE patients have reported abnormalities of tyrosine kinases in T cells, B cells, plasma cells and other immune cells (Shao and Cohen, 2014).

The JAK/STAT pathway is one of a few pleiotropic cascades required to transduce a myriad of signals for development and homeostasis in animals, from humans to flies. In mammals, the JAK/STAT pathway is the principal signaling mechanism for a vast array of cytokines

and growth factors. JAK activation stimulates cell proliferation, differentiation, cell migration and apoptosis (Rawlings *et al*., 2004). Recent research has found aberrant JAK/STAT signaling in inflammatory conditions and AD such as SLE (Goropevšek *et al*., 2016).

FLT3 signaling plays a central role in regulating the survival and differentiation of lymphoid progenitors into B cell precursors in bone marrow (Dolence *et al*., 2014). FLT3 signaling induces the phosphorylation of *STAT3* and *STAT5A*, both of which are associated with JAK/STAT signaling (Zhang, 2000; Laouar, 2003; Singh, 2012). Flt3L has also been found to increase during inflammatory conditions, such as RA disease (Guermonprez, 2013; Ramos, 2013; Tobon, 2010).

The retinoid X receptor (RXR) plays a key role in nuclear signalling pathways along with its dimeric partners such as the retinoic acid receptor, the peroxisome proliferator-activated receptors, the thyroid hormone receptor (TR), the Vitamin D receptor, the pregnane X receptor, the liver X receptor and the constitutive androstane receptor. RXR is regarded as a master regulator of numerous biological pathways, especially in the liver where many of its partners are expressed and are active (Ouamrane *et al*., 2003). Activated TR/RXR pathways may also contribute to tissue injury in the kidney (a target end-organ of SLE) (Frangou *et al*., 2016).

As shown in figure 5.2.5, some gene candidates (*ISG15*, *STAT4*, *TNFRSF10B*), are regulated by *STAT3* and *IL6*. Both *STAT3* and *IL6* have been implicated in a wide variety of inflammation-associated disease states. Mutations in *STAT3* are associated with infantile-onset multisystem AD and hyper-immunoglobulin E syndrome (Aziz *et al*., 2007). *IL6* has been linked to DM, RA and other AD's (Nakagawa *et al*., 2004).

The candidate genes are selected on the basis of their previous association with SLE, and it is therefore to be expected that they may map to known SLE pathways. The pathway analysis

confirms that the selected genes do have roles to play in relevant cellular mechanisms and regulatory processes that may underlie autoimmune diseases, inflammation and SLE.

Chapter 6: Discussion

6.1 Overview

This project investigated the exome sequence of five individuals from a South African family with multi-generational, inherited SLE.

- because the available family information is limited and the pedigree is not extensive, it is not possible to define the mode of inheritance of SLE, and in this study there were four different models of possible genetic inheritance of SLE that were explored.

- because SLE is a complex disease, it is likely that there are multiple aetiological genes of individually low penetrance, that together combine to cause the disease phenotype.

- the models explored genes that act to confer susceptibility to SLE, SLE phenotype and protection from SLE.

- rare alleles with very low frequency in the general population were identified. Some were novel, some were previously described.

- variant prioritisation was undertaken using information about genes that have previously been associated with SLE.

6.2 Identification of candidate aetiological variants

Novel variants: In our search for coding variants contributing to SLE in five family members from Cape Town, we identified ten novel variants (Table 4.2), one each in (*MYH8*, *NYX*, *SERPINB13*, *CD177*, *CD24*, *HSD11B2*, *MERTK* and *IL36G*), and two in *PRSS1*. These variants were not registered in any of the known variant databases such as dbSNP, EVS or amongst the 60,706 sequenced exomes housed in ExAC. Seven variants were found in DE SLE genes (*MYH8*, *NYX*, *SERPINB13*, *HSD11B2*, *PRSS1* and *IL36G*). Two variants were found in known SLE genes (*CD24* and *MERTK*). One variant was found in a known and DE SLE gene (*CD177*). The variants characterised in this bioinformatics analysis must next be confirmed by independent genotyping methods in the laboratory to confirm that they are not artefacts of the WES method used to detect them. Further validation studies such as disease modelling, would need to be conducted in order to elucidate the exact role, if any, of these novel variants in the pathogenesis of SLE.

The density of SNPs needed to map complex diseases is likely to vary across populations with diverse demographic histories (Chakravarti, 2001; Reich and Lander, 2001; Gabriel *et al*., 2002). The strategy for most efficiently using SNPs to map complex disease genes depends on various parameters that are, at present, not fully known. For example, the CD/CV hypothesis states that common genetic diseases are affected by common disease susceptibility alleles at a few loci that are at high frequency across ethnically diverse populations (Reich and Lander, 2001; Weiss and Clark, 2002; Pritchard, 2001; Chakravarti, 1999; Wright *et al*., 1999). There is ample to be learnt from the genetics of sub-Saharan African populations regarding the origin and nature of human complex disease. Presently, we have little understanding of the genetic structure of sub-Saharan populations and the genetic basis of complex disease in African populations because very few studies have been conducted in African ethnic groups. Our lack of knowledge about the involvement of genetics to disease in

African populations extends to single-locus diseases, which are rarely reported in African populations (Tishkoff and Williams, 2002).

6.3 Rare alleles found in genes known to be associated with SLE

Our study also reported nineteen rare missense variants in DE and known SLE genes (Table 4.2.1), with MAF < 1%, one each in (*STAT4*, *C3*, *ISG15*, *PIK3CD*, *TBC1D9*, *AOC3*, *DTX3*, *MORC1*, *SLC9A2*, *TBX6*, *TF*, *HP*, *SYNE2*, *MERTK*, *TNFRSF10B*, *GYPC* and *LRRC1*), and two in *LILRA2*. One of the variants found in *LILRA2*, located on chr19:55098805; G->A had no SNP ID but had an allele frequency reported. The variants found in *STAT4*, *ISG15*, *TBX6* and *GYPC* were all predicted to be deleterious by both SIFT and PolyPhen2 algorithms. However, the finding of a missense mutation on its own cannot be assumed to be causal. Therefore, the frequency data along with the SIFT and PolyPhen2 prediction scores of these variants provided further evidence for a possible role in SLE – if an allele is very rare in the general population this might possibly be because it has deleterious effects and is selected against. The variant selection tool used in the analysis incorporated robust models of cross-species sequence conservation, which further improve the accuracy in differentiating between benign and disease-causing variation. Therefore, if a variant causes changes that are not good for the species then they are more likely to be selected against.

Furthermore, despite the successful findings of four rare deleterious variants in *STAT4*, *ISG15*, *TBX6* and *GYPC* in this SLE family, WES also has the benefit of identifying novel variants in genes not currently known to be involved in SLE pathogenesis. Without more detailed insights regarding the involved genes as well as the characteristics of their variants, an efficient and sensitive genetic screening test for SLE is currently unattainable. Ultimately, however, establishing a low-cost and regular genetic screening test could save thousands of patients and families from invasive studies and unexpected SLE episodes. Another concern

that requires attention is that of multiple protein isoforms. Many genes such as *STAT4* have more than one transcript, and the protein prediction scores vary by transcript. Therefore, additional basic research needs to be performed on the role each transcript plays and the effect a variant has on each of them.

Whilst the Mendelian inheritance model was also considered (Asymptomatic Model) in this analysis, it is clear that there is insufficient background information and too few family members in order to conduct a meaningful investigation of Mendelian inheritance in the family. Undertaking the consideration of the fourth model highlighted the fact that complex diseases require different and more flexible approaches to analyse genetic contributors; but did not provide any convincing insights into the inheritance of SLE in this family.

There are numerous exclusive advantages and disadvantages to using WES to identify candidate disease variants, and these are discussed below.

6.4 Benefits of using WES to identify candidate aetiological variants

The major advantages are the breadth and depth of the data generated. For a relatively low price, all known protein-coding regions of the genome are included, and each region is sequenced in often 50-100 fold redundancy. Hence, if a patient with SLE does not have a *STAT4*, *ISG15*, *TBX6* and *GYPC* mutation, a search for a variant in other genes could be embarked upon. Considering the fact that an average individual has approximately 10,000 protein-coding non-reference variants (Ng *et al*., 2010), more methods must be used to significantly downsize the target region for study, which could include standard linkage analysis or more complex inherited-by-descent sharing analysis (Browning and Browning, 2011).

6.5 Limitations of using WES to identify candidate aetiological variants

Despite the depth of the data, a drawback of this method is that more complex structural variation, such as CNV (deletions, duplications), large indels (insertion or deletion of multiple bases), or epigenetic variations could be undetected. If these were existent, their discovery would depend upon other approaches, such as array comparative genomic hybridization for CNVs or bisulfite sequencing for epigenetic variations.

Based on the results, three of the nineteen rare variants found, one in *STAT4*, one in *C3* and one in *ISG15*, best supported two of the four proposed inheritance models in section 2.6. These models are the tipping point and protective mutation models. The three affected members of the proposed tipping point model shared a variant found within a known SLE susceptibility gene *STAT4*.

STAT4 is a Th1 transcription factor that has been reported to mediate Th1 T-cell response, Th1 cytokines, IL-12 and IL-23, (Remmers *et al*., 2007; Watford *et al*., 2004), and IFNγ signaling (Farrar and Murphy, 2000; Morinobu *et al.*, 2002). *STAT4* also induces a Th1 T-cell response, increasing IFNγ discharge. This influx of IFNγ would target organs such as the kidneys, propagating further IFNγ release and chronic inflammation (Li *et al*., 2011).

On the other hand, the two unaffected members of the proposed protective mutation model shared variants within the *C3* and *ISG15* genes.

The *C3* gene confers a protective effect against AD. *C3* is responsible for "waste removal" by promoting phagocytosis which helps to eliminate dying/apoptotic cells, thus decreasing the exposure to self-antigens (Markiewski and Lambris, 2007).

ISG15 is an interferon (IFN)-α/β-inducer. It was found that *ISG15* also negatively regulates IFN-α/β responses via the stabilization of the *USP18* gene, thereby preventing auto-

inflammatory consequences of uncontrolled IFN-α/β amplification, a signature seen in SLE (Zhang *et al*., 2015).

However, unlike *STAT4* and *ISG15* that formed part of the top six candidate genes (in green) that interacted with known and DE SLE genes (in yellow) (Figure 5.2.4), *C3* was not part of this network.

Furthermore, given the fact that *STAT4* and *ISG15* are also regulated by certain SLE upstream regulator genes such as *STAT3*, it could be possible that the expression levels of *STAT4* might be up-regulated thus leading to the manifestation of the disease phenotype. On the contrary, *ISG15* expression levels might also be up-regulated thus leading to the inhibition of the disease phenotype given the dual roles of this gene as a promoter and inhibitor of SLE.

Therefore, the above mentioned variants could contribute to a plausible explanation for disease susceptibility amongst the affected members and a lack thereof amongst unaffected members of this family. It is also plausible that the susceptibility variants proposed from the analysis of the Susceptibility Model, the Tipping Point variants identified in the Tipping Point model, and the Protective variants identified in the Protective mutation model could work together in a dynamic process that drives the presentation of the disease phenotype or suppresses it.

Figure 6 is a proposed model for how these variants might work individually, or together, at different levels in the continuum from unaffected phenotype to susceptible phenotype, and from susceptible phenotype to disease phenotype. In many complex diseases there is also a strong role for environmental factors that may also affect the progression to the disease state. Figure 6 also shows the interplay between variants that play different roles in the progression from unaffected to disease-susceptible and affected phenotypes; showing a possible role for

some of the candidate SLE variants identified in this study in the progression to Familial SLE.

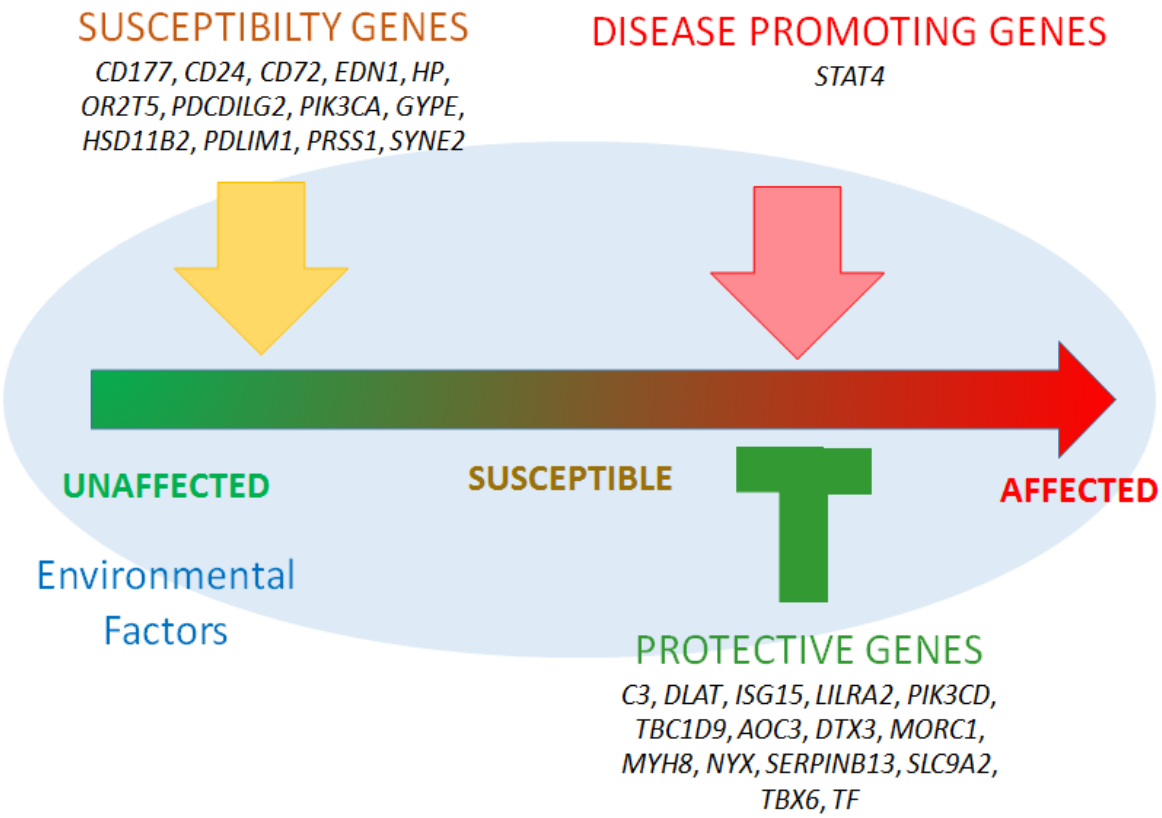

Figure 6: A proposed model showing how variants play different roles in the progression from unaffected to disease-susceptible and affected phenotypes.

Conclusion

Approximately 85% of disease-related mutations are found in the exome. WES is an efficient way to detect novel disease-causing genes.

WES has mainly been used to identify single gene variants that underlie Mendelian diseases, however there is a pressing need to also understand genetic contributions to complex diseases. Nonetheless, a study conducted by Ellyard and colleagues (Ellyard *et al*., 2014) was the first to demonstrate that WES can be used to identify rare or novel deleterious variants as genetic causes of SLE.

Analysing genetic contributors to complex disease requires more flexible models of disease inheritance, and consideration of the interplay between multiple genes (as well as environmental factors in many cases). This is why pathway analyses are important, to understand the potential effects on cellular processes by groups of genes rather than single genes.

Although the current study has analysed a rare family with inherited SLE, insights and findings from single rare or 'extreme' cases of any disease can often offer insights into the cellular and biological mechanisms underlying the disease phenotype, and these are often generaliseable or informative for less rare or idiopathic versions of the same disease. In this way, identifying genes for rare diseases increases our general understanding of genetic dysfunction and its effects in the disease state.

As the body of knowledge around genetics underlying complex diseases grows, this will assist with future studies into complex disease genetics. This is fairly new territory in comparison to all the work that has already been successfully undertaken in identifying single gene variants that underlie Mendelian diseases.

Future directions

A few directions may be taken in the future for the identification of coding variants associated with familial systemic lupus erythematosus through whole exome sequencing. One limitation to my study was a relatively small sample size, therefore, a larger sample size and performing additional exome sequencing on other families would increase the power to detect additional variants in the same gene(s). Furthermore, a more valuable approach would be to integrate data from multiple profiling techniques (e.g. genomic, epigenomic, transcriptomic) for specimens from the same individuals, thus allowing for a more comprehensive assessment of potential heritable factors in disease susceptibility.

References

Abbas, A.K., Murphy, K.M., Sher, A., 1996. Functional diversity of helper T lymphocytes. Nature 383, 787–793.

Abelson, A.-K., Delgado-Vega, A.M., Kozyrev, S.V., Sánchez, E., Velázquez-Cruz, R., Eriksson, N., Wojcik, J., Reddy, P.L., Lima, G., D'Alfonso, S., 2008. STAT4 associates with SLE through two independent effects that correlate with gene expression and act additively with IRF5 to increase risk. Ann. Rheum. Dis.

Achiron, A., Feldman, A., Mandel, M., Gurevich, M., 2007. Impaired Expression of Peripheral Blood Apoptotic-Related Gene Transcripts in Acute Multiple Sclerosis Relapse. Ann. N. Y. Acad. Sci. 1107, 155–167.

Adamson, A.S., Collins, K., Laurence, A., O'Shea, J.J., 2009. The Current STATus of lymphocyte signaling: new roles for old players. Curr. Opin. Immunol. 21, 161–166.

Adelson, M.E., Feola, M., Trama, J., Tilton, R.C., Mordechai, E., 2005. Simultaneous detection of herpes simplex virus types 1 and 2 by real-time PCR and Pyrosequencing. J. Clin. Virol. 33, 25–34.

Ahmadian, A., Gharizadeh, B., Gustafsson, A.C., Sterky, F., Nyrén, P., Uhlén, M., Lundeberg, J., 2000a. Single-nucleotide polymorphism analysis by pyrosequencing. Anal. Biochem. 280, 103–110.

Ahmadian, A., Lundeberg, J., Nyrén, P., Uhlén, M., Ronaghi, M., 2000b. Analysis of the p53 tumor suppressor gene by pyrosequencing. Biotechniques 28, 140–147.

Akahoshi, M., Nakashima, H., Tanaka, Y., Kohsaka, T., Nagano, S., Ohgami, E., Arinobu, Y., Yamaoka, K., Niiro, H., Shinozaki, M., 1999. Th1/Th2 balance of peripheral T helper cells in systemic lupus erythematosus. Arthritis Rheum. 42, 1644–1648.

Alamanos, Y., Voulgari, P.V., Siozos, C., Katsimpri, P., Tsintzos, S., Dimou, G., Politi, E.N., Rapti, A., Laina, G., Drosos, A.A., 2003. Epidemiology of systemic lupus erythematosus in northwest Greece 1982-2001. J. Rheumatol. 30, 731–735.

Alarcón, G.S., McGwin, G., Sanchez, M.L., Bastian, H.M., Fessler, B.J., Friedman, A.W., Baethge, B.A., Roseman, J., Reveille, J.D., 2004. Systemic lupus erythematosus in three ethnic groups. XIV. Poverty, wealth, and their influence on disease activity. Arthritis Care Res. 51, 73–77.

Al-Arfaj, A.S., Al-Balla, S.R., Al-Dalaan, A.N., Al-Saleh, S.S., Bahabri, S.A., Mousa, M.M., Sekeit, M.A., 2002. Prevalence of systemic lupus erythematosus in central Saudi Arabia. Saudi Med. J. 23, 87–89.

Alcorta, D.A., Barnes, D.A., Dooley, M.A., Sullivan, P., Jonas, B., Liu, Y., Lionaki, S., Reddy, C.B., Chin, H., Dempsey, A.A., 2007. Leukocyte gene expression signatures in antineutrophil cytoplasmic autoantibody and lupus glomerulonephritis. Kidney Int. 72, 853–864.

Al-Janadi, M., Al-Balla, S., Al-Dalaan, A., Raziuddin, S., 1993. Cytokine profile in systemic lupus erythematosus, rheumatoid arthritis, and other rheumatic diseases. J. Clin. Immunol. 13, 58–67.

Alkan, C., Coe, B.P., Eichler, E.E., 2011. Genome structural variation discovery and genotyping. Nat. Rev. Genet. 12, 363–376.

Allantaz, F., Chaussabel, D., Stichweh, D., Bennett, L., Allman, W., Mejias, A., Ardura, M., Chung, W., Smith, E., Wise, C., 2007. Blood leukocyte microarrays to diagnose systemic onset juvenile idiopathic arthritis and follow the response to IL-1 blockade. J. Exp. Med. 204, 2131–2144.

Altobelli, E., Chiarelli, F., Valenti, M., Verrotti, A., Blasetti, A., Di Orio, F., 1998. Family history and risk of insulin-dependent diabetes mellitus: a population-based case-control study. Acta Diabetol. 35, 57–60.

Andrews, B.S., Eisenberg, R.A., Theofilopoulos, A.N., Izui, S., Wilson, C.B., McConahey, P.J., Murphy, E.D., Roths, J.B., Dixon, F.J., 1978. Spontaneous murine lupus-like syndromes. Clinical and immunopathological manifestations in several strains. J. Exp. Med. 148, 1198–1215.

Ansorge, W.J., 2009. Next-generation DNA sequencing techniques. New Biotechnol. 25, 195–203.

Anstey, N.M., Bastian, I., Dunckley, H., Currie, B.J., 1993. Systemic lupus erythematosus in Australian aborigines: high prevalence, morbidity and mortality. Aust. N. Z. J. Med. 23, 646–651.

Aoki, H., Moro, O., Tagami, H., Kishimoto, J., 2007. Gene expression profiling analysis of solar lentigo in relation to immunohistochemical characteristics. Br. J. Dermatol. 156, 1214–1223.

Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M.D.R. and Durbin, R., 2001. The InterPro database, an integrated documentation resource for protein families, domains and functional sites. Nucleic acids research, 29(1), 37-40.

Armstrong, D.L., Reiff, A., Myones, B.L., Quismorio, F.P., Klein-Gitelman, M., McCurdy, D., Wagner-Weiner, L., Silverman, E., Ojwang, J.O., Kaufman, K.M., 2009. Identification of new SLE-associated genes with a two-step Bayesian study design. Genes Immun. 10, 446–456.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T. and Harris, M.A., 2000. Gene Ontology: tool for the unification of biology. Nature genetics, 25(1), 25-29.

Aslanidis, S., Pyrpasopoulou, A., Kontotasios, K., Doumas, S., Zamboulis, C., 2008. Parvovirus B19 infection and systemic lupus erythematosus: activation of an aberrant pathway? Eur. J. Intern. Med. 19, 314–318.

Aziz, M.H., Manoharan, H.T., Church, D.R., Dreckschmidt, N.E., Zhong, W., Oberley, T.D., Wilding, G., Verma, A.K., 2007. Protein kinase Cε interacts with signal transducers and activators of transcription 3 (Stat3), phosphorylates Stat3Ser727, and regulates its constitutive activation in prostate cancer. Cancer Res. 67, 8828–8838.

Badano, J.L., Katsanis, N., 2002. Beyond Mendel: an evolving view of human genetic disease transmission. Nat. Rev. Genet. 3, 779–789.

Baechler, E.C., Batliwalla, F.M., Karypis, G., Gaffney, P.M., Ortmann, W.A., Espe, K.J., Shark, K.B., Grande, W.J., Hughes, K.M., Kapur, V., 2003. Interferon-inducible gene expression signature in peripheral blood cells of patients with severe lupus. Proc. Natl. Acad. Sci. 100, 2610–2615.

Baechler, E.C., Bauer, J.W., Slattery, C.A., Ortmann, W.A., Espe, K.J., Novitzke, J., Ytterberg, S.R., Gregersen, P.K., Behrens, T.W., Reed, A.M., 2007. An interferon signature in the peripheral blood of dermatomyositis patients is associated with disease activity. Mol. Med. 13, 59.

Bains, W., Smith, G.C., 1988. A novel method for nucleic acid sequence determination. J. Theor. Biol. 135, 303–307.

Ballou, S.P., Khan, M.A., Kushner, I., 1982. Clinical features of systemic lupus erythematosus. Arthritis Rheum. 25, 55–60.

Barnes, M.G., Grom, A.A., Thompson, S.D., Griffin, T.A., Pavlidis, P., Itert, L., Fall, N., Sowders, D.P., Hinze, C.H., Aronow, B.J., 2009. Subtype-specific peripheral blood gene expression profiles in recent-onset juvenile idiopathic arthritis. Arthritis Rheum. 60, 2102–2112.

Barrett, J.C., Cardon, L.R., 2006. Evaluating coverage of genome-wide association studies. Nat. Genet. 38, 659–662.

Bartels, C.M., Muller, D., 2011. Systemic lupus erythematosus (SLE). N. Y. NY US Medscape Ref.

Bates, G.P., 2005. History of genetic disease: the molecular genetics of Huntington disease—a history. Nat. Rev. Genet. 6, 766–773.

Bateson, W. and Mendel, G., 1902. *Mendel's Principles of Heredity: A Defence, with a* Translation of Mendel's Original Papers on Hybridisation. the University Press.

Batliwalla, F.M., Baechler, E.C., Xiao, X., Li, W., Balasubramanian, S., Khalili, H., Damle, A., Ortmann, W.A., Perrone, A., Kantor, A.B., 2005. Peripheral blood gene expression profiling in rheumatoid arthritis. Genes Immun. 6, 388–397.

Battiwalla, F.M., Li, W., Ritchlin, C.T., Xiao, X., Brenner, M., Laragione, T., Shao, T., Durham, R., Kemshetti, S., Schwarz, E., 2005. Microarray analyses of peripheral blood cells identifies unique gene expression signature in psoriatic arthritis. Mol. Med. 11, 21–29.

Bell, J., 1951. Part I. On brachydactyly and symphalangism. Treas. Hum. Inherit. 5.

Benkovic, S.J., Cameron, C.E., 1995. [20] Kinetic analysis of nucleotide incorporation and misincorporation by klenow fragment of Escherichia coli DNA polymerase I. Methods Enzymol. 262, 257–269.

Ben-Menachem, E., 2010. Systemic lupus erythematosus: a review for anesthesiologists. Anesth. Analg. 111, 665–676.

Bennett, L., Palucka, A.K., Arce, E., Cantrell, V., Borvak, J., Banchereau, J., Pascual, V., 2003. Interferon and granulopoiesis signatures in systemic lupus erythematosus blood. J. Exp. Med. 197, 711–723.

Berglund, E.C., Kiialainen, A., Syvänen, A.-C., 2011. Next-generation sequencing technologies and applications for human genetic history and forensics. Investig. Genet. 2, 1.

Bertsias, G., Cervera, R., Boumpas, D.T., 2012. Systemic lupus erythematosus: pathogenesis and clinical features. EULAR Textb. Rheum. Dis. Geneva Switz. Eur. Leag. Rheum. 476–505.

Bias, W.B., Reveille, J.D., Beaty, T.H., Meyers, D.A., Arnett, F.C., 1986. Evidence that autoimmunity in man is a Mendelian dominant trait. Am. J. Hum. Genet. 39, 584.

Blotzer, J.W., 1983. Systemic lupus erythematosus I: historical aspects. Md. State Med. J. 32, 439–441.

Bossingham, D., 2003. Systemic lupus erythematosus in the far north of Queensland. Lupus 12, 327–331.

Botstein, D., White, R.L., Skolnick, M., Davis, R.W., 1980. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am. J. Hum. Genet. 32, 314.

Boyer, G.S., Templin, D.W., Lanier, A.P., 1991. Rheumatic diseases in Alaskan Indians of the southeast coast: high prevalence of rheumatoid arthritis and systemic lupus erythematosus. J. Rheumatol. 18, 1477–1484.

Branton, D., Deamer, D.W., Marziali, A., Bayley, H., Benner, S.A., Butler, T., Di Ventra, M., Garaj, S., Hibbs, A., Huang, X., 2008. The potential and challenges of nanopore sequencing. Nat. Biotechnol. 26, 1146–1153.

Brownstein, C.A., Beggs, A.H., Homer, N., Merriman, B., Timothy, W.Y., Flannery, K.C., DeChene, E.T., Towne, M.C., Savage, S.K., Price, E.N., 2014. An international effort towards developing standards for best practices in analysis, interpretation and reporting of clinical genome sequencing results in the CLARITY Challenge. Genome Biol. 15, 1.

Burczynski, M.E., Peterson, R.L., Twine, N.C., Zuberek, K.A., Brodeur, B.J., Casciotti, L., Maganti, V., Reddy, P.S., Strahs, A., Immermann, F., 2006. Molecular classification of Crohn's disease and ulcerative colitis patients using transcriptional profiles in peripheral blood mononuclear cells. J. Mol. Diagn. 8, 51–61.

Burton, P.R., Clayton, D.G., Cardon, L.R., Craddock, N., Deloukas, P., Duncanson, A., Kwiatkowski, D.P., McCarthy, M.I., Ouwehand, W.H., Samani, N.J., 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447, 661–678.

Carreno, L., Lopez-Longo, F.J., Monteagudo, I., Rodriguez-Mahou, M., Bascones, M., Gonzalez, C.M., Saint-Cyr, C., Lapointe, N., 1999. Immunological and clinical differences between juvenile and adult onset of systemic lupus erythematosus. Lupus 8, 287–292.

Cederholm, J., Wibell, L., 1991. Familial influence on type 1 (insulin-dependent) diabetes mellitus by relatives with either insulin-treated or type 2 (non-insulin-dependent) diabetes mellitus. Diabetes Res. Edinb. Scotl. 18, 109–113.

Center, C., 1995. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat Genet 11, 241–247.

Cervera, R., Espinosa, G., D'Cruz, D., 2009. Systemic lupus erythematosus: pathogenesis, clinical manifestations and diagnosis. EULAR Compend. Rheum. Dis. BMJ Publ. Group Eur. Leag. Rheum. 257–68.

Chakravarty, E.F., Bush, T.M., Manzi, S., Clarke, A.E., Ward, M.M., 2007. Prevalence of adult systemic lupus erythematosus in California and Pennsylvania in 2000: estimates obtained using hospitalization data. Arthritis Rheum. 56, 2092–2094.

Chambers, S., Raine, R., Rahman, A., Hagley, K., De Ceulaer, K., Isenberg, D., 2008. Factors influencing adherence to medications in a group of patients with systemic lupus erythematosus in Jamaica. Lupus 17, 761–769.

Chan, R.-Y., Lai, F.-M., Li, E.-M., Tam, L.-S., Chow, K.-M., Li, P.-T., Szeto, C.-C., 2006. Imbalance of Th1/Th2 transcription factors in patients with lupus nephritis. Rheumatology 45, 951–957.

Chang, D.M., Su, W.L., Chu, S.J., 2002. The expression and significance of intracellular T helper cytokines in systemic lupus erythematosus. Immunol. Invest. 31, 1–12.

Charlton, B., Lafferty, K.J., 1995. The Th1/Th2 balance in autoimmunity. Curr. Opin. Immunol. 7, 793–798.

Chu, J.-L., Drappa, J., Parnassa, A., Elkon, K.B., 1993. The defect in Fas mRNA expression in MRL/lpr mice is associated with insertion of the retrotransposon, ETn. J. Exp. Med. 178, 723–730.

Cock, P.J., Fields, C.J., Goto, N., Heuer, M.L., Rice, P.M., 2010. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res. 38, 1767–1771.

Cohen, P.L., Eisenberg, R.A., 1991. Lpr and gld: single gene models of systemic autoimmunity and lymphoproliferative disease. Annu. Rev. Immunol. 9, 243–269.

Consortium, 1000 Genomes Project, 2010. A map of human genome variation from population-scale sequencing. Nature 467, 1061–1073.

Consortium, I.H., 2005. A haplotype map of the human genome. Nature 437, 1299–1320.

Cooper, D.N., Krawczak, M., Polychronakos, C., Tyler-Smith, C. and Kehrer-Sawatzki, H., 2013. Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. Human genetics, 132(10), 1077-1130.

Cooper, G.M. and Shendure, J., 2011. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nature Reviews Genetics, 12(9), 628-640.

Crow, M.K., Wohlgemuth, J., 2003. Microarray analysis of gene expression in lupus. Arthritis Res Ther 5, 1.

Dahlquist, G., Blom, L., Tuvemo, T., Nyström, L., Sandström, A., Wall, S., 1989. The Swedish childhood diabetes study—results from a nine year case register and a one year case-referent study indicating that type 1 (insulin-dependent) diabetes mellitus is associated with both type 2 (non-insulin-dependent) diabetes mellitus and autoimmune disorders. Diabetologia 32, 2–6.

Dalca, A.V., Brudno, M., 2010. Genome variation discovery with high-throughput sequencing data. Brief. Bioinform. 11, 3–14.

Del Junco, D.J., Luthra, H.S., Annegers, J.F., Worthington, J.W., Kurland, L.T., 1984. The familial aggregation of rheumatoid arthritis and its relationship to the HLA-DR4 association. Am. J. Epidemiol. 119, 813–829.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., Del Angel, G., Rivas, M.A., Hanna, M., 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43, 491–498.

Dolence, J.J., Gwin, K.A., Shapiro, M.B., Medina, K.L., 2014. Flt3 signaling regulates the proliferation, survival, and maintenance of multipotent hematopoietic progenitors that generate B cell precursors. Exp. Hematol. 42, 380–393. e3.

Dramanac, R., Labat, I., Brukner, I., Crkvenjakov, R., 1989. Sequencing of megabase plus DNA by hybridization: theory of the method. Genomics 4, 114–128.

Edwards, C.J., Feldman, J.L., Beech, J., Shields, K.M., Stover, J.A., Trepicchio, W.L., Larsen, G., Foxwell, B.M., Brennan, F.M., Feldmann, M., 2007. Molecular profile of peripheral blood mononuclear cells from patients with rheumatoid arthritis. Mol. Med.-Camb. MA THEN N. Y.- 13, 40.

Edworthy, S.M., 2001. Clinical manifestations of systemic lupus erythematosus. Kelleys Textb. Rheumatol. 7, 1201–47.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., 2009. Real-time DNA sequencing from single polymerase molecules. Science 323, 133–138.

Ellyard, J.I., Jerjen, R., Martin, J.L., Lee, A., Field, M.A., Jiang, S.H., Cappello, J., Naumann, S.K., Andrews, T.D., Scott, H.S., 2014. Brief Report: Identification of a Pathogenic Variant in TREX1 in Early-Onset Cerebral Systemic Lupus Erythematosus by Whole-Exome Sequencing. Arthritis Rheumatol. 66, 3382–3386.

Elston, R.C., Stewart, J., 1971. A general model for the genetic analysis of pedigree data. Hum. Hered. 21, 523–542.

Emamian, E.S., Leon, J.M., Lessard, C.J., Grandits, M., Baechler, E.C., Gaffney, P.M., Segal, B., Rhodus, N.L., Moser, K.L., 2009. Peripheral blood gene expression profiling in Sjögren's syndrome. Genes Immun. 10, 285–296.

Espinosa, V., Kettlun, A.M., Zanocco, A., Cardemil, E., Valenzuela, M.A., 2003. Differences in nucleotide-binding site of isoapyrases deduced from tryptophan fluorescence. Phytochemistry 63, 7–14.

Falconer, D.S. and Mackay, T.F., 1996. Introduction to quantitative genetics.

Fall, N., Barnes, M., Thornton, S., Luyrink, L., Olson, J., Ilowite, N.T., Gottlieb, B.S., Griffin, T., Sherry, D.D., Thompson, S., 2007. Gene expression profiling of peripheral blood from patients with untreated new-onset systemic juvenile idiopathic arthritis reveals molecular heterogeneity that may predict macrophage activation syndrome. Arthritis Rheum. 56, 3793–3804.

Farrar, J.D., Murphy, K.M., 2000. Type I interferons and T helper development. Immunol. Today 21, 484–489.

Fassl, S., Leisser, C., Huettenbrenner, S., Maier, S., Rosenberger, G., Strasser, S., Grusch, M., Fuhrmann, G., Leuhuber, K., Polgar, D., 2003. Transferrin ensures survival of ovarian carcinoma cells when apoptosis is induced by TNFα, FasL, TRAIL, or Myc. Oncogene 22, 8343–8355.

Fisher, R.A., 1930. The genetical theory of natural selection: a complete variorum edition. Oxford University Press.

Flicek, P., Birney, E., 2009. Sense from sequence reads: methods for alignment and assembly. Nat. Methods 6, S6–S12.

Font, J., Cervera, R., Espinosa, G., Pallarés, L., Ramos-Casals, M., Jiménez, S., García-Carrasco, M., Seisdedos, L., Ingelmo, M., 1998. Systemic lupus erythematosus (SLE) in childhood: analysis of clinical and immunological findings in 34 patients and comparison with SLE characteristics in adults. Ann. Rheum. Dis. 57, 456–459.

Foster, H., Fay, A., Kelly, C., Charles, P., Walker, D., Griffiths, I., 1993. Thyroid disease and other autoimmune phenomena in a family study of primary Sjögren's syndrome. Rheumatology 32, 36–40.

Frangou, E., Grigoriou, M., Banos, A., Bertsias, G., Dermitzakis, E., Boumpas, D., 2016. SP129comparitive Transcriptome Profiling of the spleen and kidneys by next generation sequencing in a lupus murine model reveals novel molecular pathways. Nephrol. Dial. Transplant. 31, i129–i129.

Franke, A., McGovern, D.P., Barrett, J.C., Wang, K., Radford-Smith, G.L., Ahmad, T., Lees, C.W., Balschun, T., Lee, J., Roberts, R., 2010. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat. Genet. 42, 1118–1125.

Friou, G.J., 1957. Clinical application of lupus serum-nucleoprotein reaction using the fluorescent antibody technique, in: Journal of Clinical Investigation. Rockefeller Univ Press 1114 First Ave, 4th Fl, New York, NY 10021, pp. 890–890.

Gateva, V., Sandling, J.K., Hom, G., Taylor, K.E., Chung, S.A., Sun, X., Ortmann, W., Kosoy, R., Ferreira, R.C., Nordmark, G., 2009. A large-scale replication study identifies TNIP1, PRDM1, JAZF1, UHRF1BP1 and IL10 as risk loci for systemic lupus erythematosus. Nat. Genet. 41, 1228–1233.

Gharizadeh, B., Ghaderi, M., Donnelly, D., Amini, B., Wallin, K.-L., Nyrén, P., 2003. Multiple-primer DNA sequencing method. Electrophoresis 24, 1145–1151.

Gill, J.M., Quisel, A.M., Rocca, P.V., Walters, D.T., 2003. Diagnosis of systemic lupus erythematosus. Am. Fam. Physician 68, 2179–2186.

Ginn, L.R., Lin, J.-P., Plotz, P.H., Bale, S.J., Wilder, R.L., Mbauya, A., Miller, F.W., 1998. Familial autoimmunity in pedigrees of idiopathic inflammatory myopathy patients suggests common genetic risk factors for many autoimmune diseases. Arthritis Rheum. 41, 400–405.

Glaab, E., Baudot, A., Krasnogor, N. and Valencia, A., 2010. Extending pathways and processes using molecular interaction networks to analyse cancer genome data. BMC bioinformatics, 11(1), 1.

Glaab, E., Baudot, A., Krasnogor, N. and Valencia, A., 2010. TopoGSA: network topological gene set analysis. Bioinformatics, 26(9), 1271-1272.

Gladman, D.D., Urowitz, M.B., Esdaile, J.M., Hahn, B.H., Klippel, J., Lahita, R., Liang, M.H., Schur, P., Petri, M., Wallace, D., 1999. Guidelines for referral and management of systemic lupus erythematosus in adults. Arthritis Rheum. 42, 1785–1796.

Glanzmann, B., Herbst, H., Kinnear, C.J., Möller, M., Gamieldien, J. and Bardien, S., 2016. A new tool for prioritization of sequence variants from whole exome sequencing data. Source Code for Biology and Medicine, 11(1), 10.

Glessner, J.T., Wang, K., Cai, G., Korvatska, O., Kim, C.E., Wood, S., Zhang, H., Estes, A., Brune, C.W., Bradfield, J.P., 2009. Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. Nature 459, 569–573.

Goldgar, D.E., Easton, D.F., Byrnes, G.B., Spurdle, A.B., Iversen, E.S., Greenblatt, M.S., 2008. Genetic evidence and integration of various data sources for classifying uncertain variants into a single model. Hum. Mutat. 29, 1265–1272.

Goldstein, D.B., Allen, A., Keebler, J., Margulies, E.H., Petrou, S., Petrovski, S. and Sunyaev, S., 2013. Sequencing studies in human genetics: design and interpretation. Nature Reviews Genetics, 14(7), 460-470.

Gonzaga-Jauregui, C., Lupski, J.R., Gibbs, R.A., 2012. Human genome sequencing in health and disease. Annu. Rev. Med. 63, 35.

Goropevšek, A., Holcar, M., Avčin, T., 2016. The Role of STAT Signaling Pathways in the Pathogenesis of Systemic Lupus Erythematosus. Clin. Rev. Allergy Immunol. 1–18.

Gourley, I.S., Patterson, C.C., Bell, A.L., 1997. The prevalence of systemic lupus erythematosus in Northern Ireland. Lupus 6, 399–403.

Graham, D.S.C., Graham, R.R., Manku, H., Wong, A.K., Whittaker, J.C., Gaffney, P.M., Moser, K.L., Rioux, J.D., Altshuler, D., Behrens, T.W., 2008. Polymorphism at the TNF superfamily gene TNFSF4 confers susceptibility to systemic lupus erythematosus. Nat. Genet. 40, 83–89.

Graham, R.R., Kyogoku, C., Sigurdsson, S., Vlasova, I.A., Davies, L.R., Baechler, E.C., Plenge, R.M., Koeuth, T., Ortmann, W.A., Hom, G., 2007. Three functional variants of IFN regulatory factor 5 (IRF5) define risk and protective haplotypes for human lupus. Proc. Natl. Acad. Sci. 104, 6758–6763.

Grant, M., 1997. Globins, Genes and Globinopathies. Biological Sciences Review, 9, 2-5.

Greenberg, S.A., Pinkus, J.L., Pinkus, G.S., Burleson, T., Sanoudou, D., Tawil, R., Barohn, R.J., Saperstein, D.S., Briemberg, H.R., Ericsson, M., 2005. Interferon-α/β–mediated innate immune mechanisms in dermatomyositis. Ann. Neurol. 57, 664–678.

Grumet, F.C., Coukell, A., Bodmer, J.G., Bodmer, W.F., McDevitt, H.O., 1971. Histocompatibility (HL-A) antigens associated with systemic lupus erythematosus: a possible genetic predisposition to disease. N. Engl. J. Med. 285, 193–196.

Gudmundsson, S., Steinsson, K., 1990. Systemic lupus erythematosus in Iceland 1975 through 1984. A nationwide epidemiological study in an unselected population. J. Rheumatol. 17, 1162–1167.

Guermonprez, P., Helft, J., Claser, C., Deroubaix, S., Karanje, H., Gazumyan, A., Darasse-Jèze, G., Telerman, S.B., Breton, G., Schreiber, H.A., 2013. Inflammatory Flt3l is essential to mobilize dendritic cells and for T cell responses during Plasmodium infection. Nat. Med. 19, 730–738.

Han, G.M., Chen, S.L., Shen, N., Ye, S., Bao, C.D., Gu, Y.Y., 2003. Analysis of gene expression profiles in human systemic lupus erythematosus using oligonucleotide microarray. Genes Immun. 4, 177–186.

Hargraves, M.M., Richmond, H., Morton, R., 1948. Presentation of two bone marrow elements; the tart cell and the LE cell., in: Proceedings of the Staff Meetings. Mayo Clinic. p. 25.

Hart, H.H., Grigor, R.R., Caughey, D.E., 1983. Ethnic difference in the prevalence of systemic lupus erythematosus. Ann. Rheum. Dis. 42, 529–532.

Hinrichs, A.L., Suarez, B.K., 2011. Incorporating linkage information into a common disease/rare variant framework. Genet. Epidemiol. 35, S74–S79.

Hochberg, M.C., 1987. Prevalence of systemic lupus erythematosus in England and Wales, 1981-2. Ann. Rheum. Dis. 46, 664–666.

Hochberg, M.C., 1997. Updating the American College of Rheumatology revised criteria for the classification of systemic lupus erythematosus. Arthritis Rheum. 40, 1725–1725.

Hom, G., Graham, R.R., Modrek, B., Taylor, K.E., Ortmann, W., Garnier, S., Lee, A.T., Chung, S.A., Ferreira, R.C., Pant, P.K., 2008. Association of systemic lupus erythematosus with C8orf13–BLK and ITGAM–ITGAX. N. Engl. J. Med. 358, 900–909.

Hopkinson, N.D., Doherty, M., Powell, R.J., 1993. The prevalence and incidence of systemic lupus erythematosus in Nottingham, UK, 1989–1990. Rheumatology 32, 110–115.

Hopkinson, N.D., Doherty, M., Powell, R.J., 1994. Clinical features and race-specific incidence/prevalence rates of systemic lupus erythematosus in a geographically complete cohort of patients. Ann. Rheum. Dis. 53, 675–680.

Hung, C.C., Lin, S.Y., Lee, C.N., Chen, C.P., Lin, S.P., Chao, M.C., Chiou, S.S. and Su, Y.N., 2011. Low penetrance of retinoblastoma for p. V654L mutation of the RB1 gene. BMC medical genetics, 12(1), 1.

Husby, I.-M.G., 1999. Application of the 1982 revised criteria for the classification of systemic lupus erythematosus on a cohort of 346 Norwegian patients with connective tissue disease. Scand. J. Rheumatol. 28, 81–87.

Ingram, V.M., 1957. Gene mutations in human haemoglobin: the chemical difference between normal and sickle cell haemoglobin. Nature 180, 326–328.

Izui, S., Iwamoto, M., FOSSATI, U., Merino, R., Takahashi, S., IBNOU-ZEKRI, N., 1995. The Yaa gene model of systemic lupus erythematosus. Immunol. Rev. 144, 137–156.

Jacob, C.O., Zhu, J., Armstrong, D.L., Yan, M., Han, J., Zhou, X.J., Thomas, J.A., Reiff, A., Myones, B.L., Ojwang, J.O., 2009. Identification of IRAK1 as a risk gene with critical role in the pathogenesis of systemic lupus erythematosus. Proc. Natl. Acad. Sci. 106, 6256–6261.

Jadassohn, J., 1904. Lupus erythematodes. Handbach Hautkrakheiten Wien Alfred Hold. 298–404.

James, R.A., Campbell, I.M., Chen, E.S., Boone, P.M., Rao, M.A., Bainbridge, M.N., Lupski, J.R., Yang, Y., Eng, C.M., Posey, J.E. and Shaw, C.A., 2016. A visual and curatorial approach to clinical variant prioritization and disease gene discovery in genome-wide diagnostics. Genome Medicine, 8(13).

James, J.A., Kaufman, K.M., Farris, A.D., Taylor-Albert, E., Lehman, T.J., Harley, J.B., 1997. An increased prevalence of Epstein-Barr virus infection in young patients suggests a possible etiology for systemic lupus erythematosus. J. Clin. Invest. 100, 3019.

Janković, S.M., Radosavlijević, V.R., Marinković, J.M., 1997. Risk factors for Graves' disease. Eur. J. Epidemiol. 13, 15–18.

Johnson, A.E., Gordon, C., Palmer, R.G., Bacon, P.A., 1995. The prevalence and incidence of systemic lupus erythematosus in Birmingham, England. Arthritis Rheum. 38, 551–558.

Jones, M.A., Silman, A.J., Whiting, S., Barrett, E.M., Symmons, D.P., 1996. Occurrence of rheumatoid arthritis is not increased in the first degree relatives of a population based inception cohort of inflammatory polyarthritis. Ann. Rheum. Dis. 55, 89–93.

Jonsson, H., Nived, O., Sturfelt, G., Silman, A., 1990. Estimating the incidence of systemic lupus erythematosus in a defined population using multiple sources of retrieval. Rheumatology 29, 185–188.

Jonsson, H., Nived, O.L.A., Sturfelt, G., 1989. Outcome in systemic lupus erythematosus: a prospective study of patients from a defined population. Medicine (Baltimore) 68, 141–150.

Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., Jassal, B., Gopinath, G.R., Wu, G.R., Matthews, L. and Lewis, S., 2005. Reactome: a knowledgebase of biological pathways. Nucleic acids research, 33(suppl 1), D428-D432.

Ju, J., Kim, D.H., Bi, L., Meng, Q., Bai, X., Li, Z., Li, X., Marma, M.S., Shi, S., Wu, J., 2006. Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. Proc. Natl. Acad. Sci. 103, 19635–19640.

Kaizer, E.C., Glaser, C.L., Chaussabel, D., Banchereau, J., Pascual, V., White, P.C., 2007. Gene expression in peripheral blood mononuclear cells from children with diabetes. J. Clin. Endocrinol. Metab. 92, 3705–3711.

Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K.F., Itoh, M., Kawashima, S., Katayama, T., Araki, M. and Hirakawa, M., 2006. From genomics to chemical genomics: new developments in KEGG. Nucleic acids research, 34(suppl 1), D354-D357.

Kang, I., Quan, T., Nolasco, H., Park, S.-H., Hong, M.S., Crouch, J., Pamer, E.G., Howe, J.G., Craft, J., 2004. Defective control of latent Epstein-Barr virus infection in systemic lupus erythematosus. J. Immunol. 172, 1287–1294.

Kaposi, M., 1872. Neue Beitrage zur Kenntniss des Lupus erythematosus. Arch Dermat Syph 4, 36–78.

Kariuki, S.N., Moore, J.G., Kirou, K.A., Crow, M.K., Utset, T.O., Niewold, T.B., 2009. Age- and gender-specific modulation of serum osteopontin and interferon-α by osteopontin genotype in systemic lupus erythematosus. Genes Immun. 10, 487–494.

Keeler, C.E., 1953. The Caribe Cuna moon-child and its heredity. J. Hered. 44, 163–172.

Koumantaki, Y., Giziaki, E., Linos, A., Kontomerkos, A., Kaklamanis, P., Vaiopoulos, G., Mandas, J., Kaklamani, E., 1997. Family history as a risk factor for rheumatoid arthritis: a case-control study. J. Rheumatol. 24, 1522–1526.

Kumar, S., Dudley, J.T., Filipski, A. and Liu, L., 2011. Phylomedicine: an evolutionary telescope to explore and diagnose the universe of disease mutations. Trends in Genetics, 27(9), 377-386.

Kunman, D.M., Steiberg, A.D., 1995. Inquiry into murine and human lupus. Immunol. Rev. 144, 157–193.

Kvasnička, J., Marek, J., Kvasnička, T., Weiss, V., Markova, M., Štiěpán, J., Umlaufova, A., 2000. Increase of adhesion molecules, fibrinogen, type-1 plasminogen activator inhibitor and orosomucoid in growth hormone (GH) deficient adults and their modulation by recombinant human GH replacement. Clin. Endocrinol. (Oxf.) 52, 543–548.

Lam, H.Y., Pan, C., Clark, M.J., Lacroute, P., Chen, R., Haraksingh, R., O'Huallachain, M., Gerstein, M.B., Kidd, J.M., Bustamante, C.D. and Snyder, M., 2012. Detecting and annotating genetic variations using the HugeSeq pipeline. Nature biotechnology, 30(3), 226-229.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., 2001. International human genome sequencing consortium. Nature 409, 860–921.

Laouar, Y., Welte, T., Fu, X.-Y., Flavell, R.A., 2003. STAT3 is required for Flt3L-dependent dendritic cell differentiation. Immunity 19, 903–912.

Lawrence, J.S., Martins, C.L., Drake, G.L., 1987. A family survey of lupus erythematosus. 1. Heritability. J. Rheumatol. 14, 913–921.

Lee, S.H., Shin, M.S., Kim, H.S., Lee, H.K., Park, W.S., Kim, S.Y., Lee, J.H., Han, S.Y., Park, J.Y., Oh, R.R., 2001. Somatic mutations of TRAIL-receptor 1 and TRAIL-receptor 2 genes in non-Hodgkin's lymphoma. Oncogene 20, 399–403.

Lee-Kirsch, M.A., Gong, M., Chowdhury, D., Senenko, L., Engel, K., Lee, Y.-A., de Silva, U., Bailey, S.L., Witte, T., Vyse, T.J., 2007. Mutations in the gene encoding the 3′-5′ DNA exonuclease TREX1 are associated with systemic lupus erythematosus. Nat. Genet. 39, 1065–1067.

Lehman, T.J., Schechter, S.J., Sundel, R.P., Oliveira, S.K., Huttenlocher, A., Onel, K.B., 2004. Thalidomide for severe systemic onset juvenile rheumatoid arthritis: a multicenter study. J. Pediatr. 145, 856–857.

Lequerré, T., Gauthier-Jauneau, A.-C., Bansard, C., Derambure, C., Hiron, M., Vittecoq, O., Daveau, M., Mejjad, O., Daragon, A., Tron, F., 2006. Gene profiling in white blood cells predicts infliximab responsiveness in rheumatoid arthritis. Arthritis Res. Ther. 8, 1.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25, 2078–2079.

Li, L., Mohan, C., 2007. Genetic basis of murine lupus nephritis, in: Seminars in Nephrology. Elsevier, pp. 12–21.

Li, N., Grivennikov, S.I., Karin, M., 2011. The unholy trinity: inflammation, cytokines, and STAT3 shape the cancer microenvironment. Cancer Cell 19, 429–431.

Lin, J.-P., Cash, J.M., Doyle, S.Z., Peden, S., Kanik, K., Amos, C.I., Bale, S.J., Wilder, R.L., 1998. Familial clustering of rheumatoid arthritis with other autoimmune diseases. Hum. Genet. 103, 475–482.

Liu, X., Jian, X. and Boerwinkle, E., 2011. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. Human mutation, 32(8), 894-899.

Lobo, I., 2008. Same genetic mutation, different genetic disease phenotype. Nature Education, 1(1), 64.

Lopez, P., Mozo, L., Gutierrez, C., Suarez, A., 2003. Epidemiology of systemic lupus erythematosus in a northern Spanish population: gender and age influence on immunological features. Lupus 12, 860–865.

Lovmar, L., Syvänen, A.-C., 2006. Multiple displacement amplification to create a long-lasting source of DNA for genetic studies. Hum. Mutat. 27, 603–614.

Luxton, G.G., Gomes, E.R., Folker, E.S., Vintinner, E., Gundersen, G.G., 2010. Linear arrays of nuclear envelope proteins harness retrograde actin flow for nuclear movement. Science 329, 956–959.

Lynch, D.H., Watson, M.L., Alderson, M.R., Baum, P.R., Miller, R.E., Tough, T., Gibson, M., Davis-Smith, T., Smiths, C.A., Hunter, K., 1994. The mouse Fas-ligand gene is mutated in gld mice and is part of a TNF family gene cluster. Immunity 1, 131–136.

Maas, K., Chan, S., Parker, J., Slater, A., Moore, J., Olsen, N., Aune, T.M., 2002. Cutting edge: molecular portrait of human autoimmune disease. J. Immunol. 169, 5–9.

Majewski, J., Schwartzentruber, J., Lalonde, E., Montpetit, A. and Jabado, N., 2011. What can exome sequencing do for you?. Journal of medical genetics, pp.jmedgenet-2011.

Mangino, M. and Spector, T., 2012. Understanding coronary artery disease using twin studies. Heart, pp.heartjnl-2012.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., 2009. Finding the missing heritability of complex diseases. Nature 461, 747–753.

Manzi, S., Meilahn, E.N., Rairie, J.E., Conte, C.G., Medsger, T.A., Jansen-McWilliams, L., D'agostino, R.B., Kuller, L.H., 1997. Age-specific incidence rates of myocardial infarction and angina in women with systemic lupus erythematosus: comparison with the Framingham Study. Am. J. Epidemiol. 145, 408–415.

Mardis, E.R., 2011. A decade/'s perspective on DNA sequencing technology. Nature 470, 198–203.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.-J., Chen, Z., 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437, 376–380.

Marini, R., Costallat, L.T., 1999. Young age at onset, renal involvement, and arterial hypertension are of adverse prognostic significance in juvenile systemic lupus erythematosus. Rev. Rhum. Engl. Ed 66, 303–309.

Markiewski, M.M., Lambris, J.D., 2007. The role of complement in inflammatory diseases from behind the scenes into the spotlight. Am. J. Pathol. 171, 715–727.

Maskarinec, G., Katz, A.R., 1995. Prevalence of systemic lupus erythematosus in Hawaii: is there a difference between ethnic groups? Hawaii Med. J. 54, 406–409.

Masutani, K., Akahoshi, M., Tsuruya, K., Tokumoto, M., Ninomiya, T., Kohsaka, T., Fukuda, K., Kanai, H., Nakashima, H., Otsuka, T., 2001. Predominance of Th1 immune response in diffuse proliferative lupus nephritis. Arthritis Rheum. 44, 2097–2106.

Maxam, A.M., Gilbert, W., 1977. A new method for sequencing DNA. Proc. Natl. Acad. Sci. 74, 560–564.

Mccarty, D.J., Manzi, S., Medsger, T.A., Ramsey-Goldman, R., Laporte, R.E., Kwoh, C.K., 1995. Incidence of systemic lupus erythematosus race and gender differences. Arthritis Rheum. 38, 1260–1270.

McClain, M.T., Heinlen, L.D., Dennis, G.J., Roebuck, J., Harley, J.B., James, J.A., 2005. Early events in lupus humoral autoimmunity suggest initiation through molecular mimicry. Nat. Med. 11, 85–89.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. and DePristo, M.A., 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research, 20(9), 1297-1303.

McKernan, K.J., Peckham, H.E., Costa, G.L., McLaughlin, S.F., Fu, Y., Tsung, E.F., Clouser, C.R., Duncan, C., Ichikawa, J.K., Lee, C.C., 2009. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. Genome Res. 19, 1527–1541.

Meaburn, E., Schulz, R., 2012. Next generation sequencing in epigenetics: insights and challenges, in: Seminars in Cell & Developmental Biology. Elsevier, pp. 192–199.

Men, A.E., Wilson, P., Siemering, K., Forrest, S., 2008. Sanger DNA sequencing. -Gener. Genome Seq. Pers. Med. 1–11.

Menashe, I., Maeder, D., Garcia-Closas, M., Figueroa, J.D., Bhattacharjee, S., Rotunno, M., Kraft, P., Hunter, D.J., Chanock, S.J., Rosenberg, P.S. and Chatterjee, N., 2010. Pathway analysis of breast cancer genome-wide association study highlights three pathways and one canonical signaling cascade. Cancer research, 70(11), 4453-4459.

Merino, R., Fossati, L., Lacour, M., Lemoine, R., Higaki, M., Izui, S., 1992. H-2-linked control of the Yaa gene-induced acceleration of lupus-like autoimmune disease in BXSB mice. Eur. J. Immunol. 22, 295–299.

Metzker, M.L., 2010. Sequencing technologies—the next generation. Nat. Rev. Genet. 11, 31–46.

Midgard, R., Grønning, M., Riise, T., Kvåle, G., Nyland, H., 1996. Multiple sclerosis and chronic inflammatory diseases A case-control study. Acta Neurol. Scand. 93, 322–328.

Mok, C.C., Lau, C.S., 2003. Pathogenesis of systemic lupus erythematosus. J. Clin. Pathol. 56, 481–490.

Molokhia, M., McKeigue, P., 2000. Risk for rheumatic disease in relation to ethnicity and admixture. Arthritis Res. Ther. 2, 1.

Moore, J.E., Shulman, L.E., Scott, J.T., 1957. The Natural History of Systemic Lupus Erythematosus:—An Approach to its Study through Chronic Biologic False Positive Reactors: Interim Report. Trans. Am. Clin. Climatol. Assoc. 68, 59.

Morel, L., 2010. Genetics of SLE: evidence from mouse models. Nat. Rev. Rheumatol. 6, 348–357.

Morinobu, A., Gadina, M., Strober, W., Visconti, R., Fornace, A., Montagna, C., Feldman, G.M., Nishikomori, R., O'Shea, J.J., 2002. STAT4 serine phosphorylation is critical for IL-12-induced IFN-γ production but not for cell proliferation. Proc. Natl. Acad. Sci. 99, 12281–12286.

Mosca, M., Ruiz-Irastorza, G., Khamashta, M.A., Hughes, G.R., 2001. Treatment of systemic lupus erythematosus. Int. Immunopharmacol. 1, 1065–1075.

Moser, K.L., Kelly, J.A., Lessard, C.J., Harley, J.B., 2009. Recent insights into the genetic basis of systemic lupus erythematosus. Genes Immun. 10, 373–379.

Moser, K.L., Neas, B.R., Salmon, J.E., Yu, H., Gray-McGuire, C., Asundi, N., Bruner, G.R., Fox, J., Kelly, J., Henshall, S., 1998. Genome scan of human systemic lupus erythematosus: evidence for linkage on chromosome 1q in African-American pedigrees. Proc. Natl. Acad. Sci. 95, 14869–14874.

Moser, K.L., Rioux, J.D., Altshuler, D., Behrens, T.W., 2008. Polymorphism at the TNF superfamily gene TNFSF4 confers susceptibility to systemic lupus erythematosus. Nat. Genet. 40, 83–89.

Moser, M., Murphy, K.M., 2000. Dendritic cell regulation of TH1-TH2 development. Nat. Immunol. 1, 199–205.

Mosmann, T.R., Cherwinski, H., Bond, M.W., Giedlin, M.A., Coffman, R.L., 1986. Two types of murine helper T cell clone. I. Definition according to profiles of lymphokine activities and secreted proteins. J. Immunol. 136, 2348–2357.

Mosmann, T.R., Coffman, R.L., 1989. TH1 and TH2 cells: different patterns of lymphokine secretion lead to different functional properties. Annu. Rev. Immunol. 7, 145–173.

Nagata, C., Fujita, S., Iwata, H., Kurosawa, Y., Kobayashi, K., Kobayashi, M., Motegi, K., Omura, T., Yamamoto, M., Nose, T., 1995. Systemic lupus erythematosus: A case-control epidemiologic study in Japan. Int. J. Dermatol. 34, 333–337.

Nakagawa, H., Liyanarachchi, S., Davuluri, R.V., Auer, H., Martin, E.W., de la Chapelle, A., Frankel, W.L., 2004. Role of cancer-associated stromal fibroblasts in metastatic colon cancer to the liver and their expression profiles. Oncogene 23, 7366–7377.

Naleway, A.L., Davis, M.E., Greenlee, R.T., Wilson, D.A., McCarty, D.J., 2005. Epidemiology of systemic lupus erythematosus in rural Wisconsin. Lupus 14, 862–866.

Neale, B.M., Kou, Y., Liu, L., Ma'Ayan, A., Samocha, K.E., Sabo, A., Lin, C.-F., Stevens, C., Wang, L.-S., Makarov, V., 2012. Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature 485, 242–245.

Nguyen, C., Limaye, N., Wakeland, E.K., 2002. Susceptibility genes in the pathogenesis of murine lupus. Arthritis Res. Ther. 4, 1.

Nielsen, R., Paul, J.S., Albrechtsen, A., Song, Y.S., 2011. Genotype and SNP calling from next-generation sequencing data. Nat. Rev. Genet. 12, 443–451.

Nightingale, A.L., Farmer, R.D., de Vries, C.S., 2006. Incidence of clinically diagnosed systemic lupus erythematosus 1992–1998 using the UK General Practice Research Database. Pharmacoepidemiol. Drug Saf. 15, 656–661.

Nishimura, D., 2001. BioCarta. Biotech Software & Internet Report: The Computer Software Journal for Scient, 2(3), 117-120.

Nossent, H.C., 2001. Systemic lupus erythematosus in the Arctic region of Norway. J. Rheumatol. 28, 539–546.

Nossent, J.C., 1992. Systemic lupus erythematosus on the Caribbean island of Curacao: an epidemiological investigation. Ann. Rheum. Dis. 51, 1197–1201.

Nurieva, R.I., Chung, Y., Martinez, G.J., Yang, X.O., Tanaka, S., Matskevitch, T.D., Wang, Y.-H., Dong, C., 2009. Bcl6 mediates the development of T follicular helper cells. Science 325, 1001–1005.

O'Roak, B.J., Vives, L., Girirajan, S., Karakoc, E., Krumm, N., Coe, B.P., Levy, R., Ko, A., Lee, C., Smith, J.D., 2012. Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. Nature 485, 246–250.

Ogawa, N., Itoh, M., Goto, Y., 1992. Abnormal production of B cell growth factor in patients with systemic lupus erythematosus. Clin. Exp. Immunol. 89, 26–31.

Ogilvie, E.M., Khan, A., Hubank, M., Kellam, P., Woo, P., 2007. Specific gene expression profiles in systemic juvenile idiopathic arthritis. Arthritis Rheum. 56, 1954–1965.

Okkenhaug, K., Bilancio, A., Farjot, G., Priddle, H., Sancho, S., Peskett, E., Pearce, W., Meek, S.E., Salpekar, A., Waterfield, M.D., 2002. Impaired B and T cell antigen receptor signaling in p110δ PI 3-kinase mutant mice. Science 297, 1031–1034.

Okumura, A., Lu, G., Pitha-Rowe, I., Pitha, P.M., 2006. Innate antiviral response targets HIV-1 release by the induction of ubiquitin-like protein ISG15. Proc. Natl. Acad. Sci. 103, 1440–1445.

Oslek, W., 1904. On the visceral manifestations of the eeythema geoup of skin diseases:[Third paper.]. Am. J. Med. Sci. 127, 1–23.

Ouamrane, L., Larrieu, G., Gauthier, B., Pineau, T., 2003. RXR activators molecular signalling: involvement of a PPARα-dependent pathway in the liver and kidney, evidence for an alternative pathway in the heart. Br. J. Pharmacol. 138, 845–854.

Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M., Krabichler, B., Speicher, M.R., Zschocke, J. and Trajanoski, Z., 2014. A survey of tools for variant analysis of next-generation genome sequencing data. Briefings in bioinformatics, 15(2), 256-278.

Pareek, C.S., Smoczynski, R., Tretyn, A., 2011. Sequencing technologies and genome sequencing. J. Appl. Genet. 52, 413–435.

Pascual, V., Allantaz, F., Arce, E., Punaro, M., Banchereau, J., 2005. Role of interleukin-1 (IL-1) in the pathogenesis of systemic onset juvenile idiopathic arthritis and clinical response to IL-1 blockade. J. Exp. Med. 201, 1479–1486.

Pauling, L., Itano, H.A., Singer, S.J., Wells, I.C., 1949. Sickle cell anemia. Science 110, 543–8.

Pepke, S., Wold, B., Mortazavi, A., 2009. Computation for ChIP-seq and RNA-seq studies. Nat. Methods 6, S22–S32.

Peschken, C.A., Esdaile, J.M., 2000. Systemic lupus erythematosus in North American Indians: a population based study. J. Rheumatol. 27, 1884–1891.

Pico, A.R., Kelder, T., Van Iersel, M.P., Hanspers, K., Conklin, B.R. and Evelo, C., 2008. WikiPathways: pathway editing for the people. PLoS Biol, 6(7), e184.

Pinto, D., Pagnamenta, A.T., Klei, L., Anney, R., Merico, D., Regan, R., Conroy, J., Magalhaes, T.R., Correia, C., Abrahams, B.S., 2010. Functional impact of global rare copy number variation in autism spectrum disorders. Nature 466, 368–372.

Pisitkun, P., Deane, J.A., Difilippantonio, M.J., Tarasenko, T., Satterthwaite, A.B., Bolland, S., 2006. Autoreactive B cell responses to RNA-related antigens due to TLR7 gene duplication. Science 312, 1669–1672.

Potti, A., Bild, A., Dressman, H.K., Lewis, D.A., Nevins, J.R., Ortel, T.L., 2006. Gene-expression patterns predict phenotypes of immune-mediated thrombosis. Blood 107, 1391–1396.

Pritchard, J.K., 2001. Are rare variants responsible for susceptibility to complex diseases?. The American Journal of Human Genetics, 69(1), 124-137.

Pushkarev, D., Neff, N.F., Quake, S.R., 2009. Single-molecule sequencing of an individual human genome. Nat. Biotechnol. 27, 847–850.

Quintáns, B., Ordóñez-Ugalde, A., Cacheiro, P., Carracedo, A., Sobrido, M.J., 2014. Medical genomics: the intricate path from genetic variant identification to clinical interpretation. Appl. Transl. Genomics 3, 60–67.

Rackham, O.J., Shihab, H.A., Johnson, M.R. and Petretto, E., 2015. EvoTol: a protein-sequence based evolutionary intolerance framework for disease-gene prioritization. Nucleic acids research, 43(5), e33-e33.

Rahman, P., Urowitz, M.B., Gladman, D.D., Bruce, I.N., Genest Jr, J., 1999. Contribution of traditional risk factors to coronary artery disease in patients with systemic lupus erythematosus. J. Rheumatol. 26, 2363–2368.

Ramos, M.I., Perez, S.G., Aarrass, S., Helder, B., Broekstra, P., Gerlag, D.M., Reedquist, K.A., Tak, P.P., Lebre, M.C., 2013. FMS-related tyrosine kinase 3 ligand (Flt3L)/CD135 axis in rheumatoid arthritis. Arthritis Res. Ther. 15, 1.

Ramos-Casals, M., Cuadrado, M.J., Alba, P., Sanna, G., Brito-Zerón, P., Bertolaccini, L., Babini, A., Moreno, A., D'Cruz, D., Khamashta, M.A., 2008. Acute viral infections in patients with systemic lupus erythematosus: description of 23 cases and review of the literature. Medicine (Baltimore) 87, 311–318.

Rawlings, J.S., Rosler, K.M., Harrison, D.A., 2004. The JAK/STAT signaling pathway. J. Cell Sci. 117, 1281–1283.

Read, A., 1992. Cystic fibrosis. Population genetics in action. Biol Sci Rev, 4(4), 18-20.

Reddy, M.P.L., Iatan, I., Weissglas-Volkov, D., Nikkola, E., Haas, B.E., Juvonen, M., Ruel, I., Sinsheimer, J.S., Genest, J., Pajukanta, P., 2012. Exome sequencing identifies two rare variants for low HDL-C in an extended family. Circ. Cardiovasc. Genet. CIRCGENETICS. 112.963264.

Reich, D.E. and Lander, E.S., 2001. On the allelic spectrum of human disease. *TRENDS in* Genetics, 17(9), 502-510.

Remmers, E.F., Plenge, R.M., Lee, A.T., Graham, R.R., Hom, G., Behrens, T.W., De Bakker, P.I., Le, J.M., Lee, H.-S., Batliwalla, F., 2007. STAT4 and the risk of rheumatoid arthritis and systemic lupus erythematosus. N. Engl. J. Med. 357, 977–986.

Rimoin, D.L., Connor, J.M., Pyeritz, R.E., Korf, B.R., 2007. Emery and Rimoin's principles and practice of medical genetics. Churchill Livingstone Elsevier.

Risch, N., Merikangas, K., 1996. The future of genetic studies of complex human diseases. Science 273, 1516–1517.

Robertson, N.P., Fraser, M., Deans, J., Clayton, D., Walker, N., Compston, D.A.S., 1996. Age–adjusted recurrence risks for relatives of patients with multiple sclerosis. Brain 119, 449–455.

Rommel, C., Camps, M., Ji, H., 2007. PI3Kδ and PI3Kγ: partners in crime in inflammation in rheumatoid arthritis and beyond? Nat. Rev. Immunol. 7, 191–201.

Rood, R. ten C., van Suijlekom-Smit, L., den Ouden, E., Ouwerkerk, F., Breedveld, F., Huizinga, T., 1999. Childhood-onset systemic lupus erythematosus: clinical presentation and prognosis in 31 patients. Scand. J. Rheumatol. 28, 222–226.

Rus, V., Atamas, S.P., Shustova, V., Luzina, I.G., Selaru, F., Magder, L.S., Via, C.S., 2002. Expression of cytokine-and chemokine-related genes in peripheral blood mononuclear cells from lupus patients by cDNA array. Clin. Immunol. 102, 283–290.

Sadovnick, A.D., Baird, P.A., Ward, R.H., Optiz, J.M., Reynolds, J.F., 1988. Multiple sclerosis. Updated risks for relatives. Am. J. Med. Genet. 29, 533–541.

Sakkas, L.I., Moore, D.F., Akritidis, N.C., 1995. Cancer in families with systemic sclerosis. Am. J. Med. Sci. 310, 223–225.

Samanta, A., Feehally, J., Roy, S., Nichol, F.E., Sheldon, P.J., Walls, J., 1991. High prevalence of systemic disease and mortality in Asian subjects with systemic lupus erythematosus. Ann. Rheum. Dis. 50, 490–492.

Sanders, S.J., Murtha, M.T., Gupta, A.R., Murdoch, J.D., Raubeson, M.J., Willsey, A.J., Ercan-Sencicek, A.G., DiLullo, N.M., Parikshak, N.N., Stein, J.L., 2012. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature 485, 237–241.

Sanger, F., Nicklen, S., Coulson, A.R., 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. 74, 5463–5467.

Santiago-Raber, M.-L., Laporte, C., Reininger, L., Izui, S., 2004. Genetic basis of murine lupus. Autoimmun. Rev. 3, 33–39.

Sawalha, A.H., Namjou, B., Nath, S.K., Kilpatrick, J., Germundson, A., Kelly, J.A., Hutchings, D., James, J., Harley, J., 2002. Genetic linkage of systemic lupus erythematosus with chromosome 11q14 (SLEH1) in African-American families stratified by a nucleolar antinuclear antibody pattern. Genes Immun. 3, S31–S34.

Schadt, E.E., Turner, S., Kasarskis, A., 2010. A window into third-generation sequencing. Hum. Mol. Genet. 19, R227–R240.

Schaefer, C.F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T. and Buetow, K.H., 2009. PID: the pathway interaction database. Nucleic acids research, 37(suppl 1), D674-D679.

Schbath, S., Martin, V., Zytnicki, M., Fayolle, J., Loux, V., Gibrat, J.-F., 2012. Mapping reads on a genomic sequence: an algorithmic overview and a practical comparative analysis. J. Comput. Biol. 19, 796–813.

Schloss, J.A., 2008. How to get genomes at one ten-thousandth the cost. Nat. Biotechnol. 26, 1113.

Schur, P.H., 1995. Review: Genetics of systemic lupus erythematosus. Lupus 4, 425–437.

Schur, P.H., 2003. General symptomatology and diagnosis of systemic lupus erythematosus in adults. Copiado El 20.

Segasothy, M., Phillips, P.A., 2001. Systemic lupus erythematosus in Aborigines and Caucasians in central Australia: a comparative study. Lupus 10, 439–444.

Sestak, A.L., Nath, S.K., Sawalha, A.H., Harley, J.B., 2007. Current status of lupus genetics. Arthritis Res. Ther. 9, 1.

Shao, W.-H., Cohen, P.L., 2014. The role of tyrosine kinases in systemic lupus erythematosus and their potential as therapeutic targets. Expert Rev. Clin. Immunol. 10, 573–582.

Shawky, R.M., Abd-Elkhalek, H.S. and Gad, S., 2014. Intrafamilial variability in Simpson–Golabi–Behmel syndrome with bilateral posterior ear lobule creases. Egyptian Journal of Medical Human Genetics, 15(1), 87-90.

Shawky, R.M., Abd-Elkhalek, H.S., Gad, S., Mohammad, S.A. and Seifeldin, N.S., 2013. Cornelia-de Lange syndrome in an Egyptian infant with unusual bone deformities. *Egyptian* Journal of Medical Human Genetics, 14(1), 109-112.

Shen, R., Fan, J.-B., Campbell, D., Chang, W., Chen, J., Doucet, D., Yeakley, J., Bibikova, M., Garcia, E.W., McBride, C., 2005. High-throughput SNP genotyping on universal bead arrays. Mutat. Res. Mol. Mech. Mutagen. 573, 70–82.

Shendure, J., Porreca, G.J., Reppas, N.B., Lin, X., McCutcheon, J.P., Rosenbaum, A.M., Wang, M.D., Zhang, K., Mitra, R.D., Church, G.M., 2005. Accurate multiplex polony sequencing of an evolved bacterial genome. Science 309, 1728–1732.

Sherry, S.T., Ward, M., Sirotkin, K., 1999. dbSNP—database for single nucleotide polymorphisms and other classes of minor genetic variation. Genome Res. 9, 677–679.

Shihab, H.A., Gough, J., Cooper, D.N., Stenson, P.D., Barker, G.L., Edwards, K.J., Day, I.N. and Gaunt, T.R., 2013. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. Human mutation, 34(1), 57-65.

Shiraishi, T., Yoshida, T., Nakata, S., Horinaka, M., Wakada, M., Mizutani, Y., Miki, T., Sakai, T., 2005. Tunicamycin enhances tumor necrosis factor–related apoptosis-inducing ligand–induced apoptosis in human prostate cancer cells. Cancer Res. 65, 6364–6370.

Silverman, E., Eddy, A., n.d. Systemic Lupus Erythematosus. Textbook of Pediatric Rheumatology. Edited by: Cassidy JT, Petty RE, Laxer RM, Lindsley CB. 2011, Philadeplhia: Saunders. Elsevier.

Singh, M.K., Scott, T.F., LaFramboise, W.A., Hu, F.Z., Post, J.C., Ehrlich, G.D., 2007. Gene expression changes in peripheral blood mononuclear cells from multiple sclerosis patients undergoing β-interferon therapy. J. Neurol. Sci. 258, 52–59.

Singh, P., Hoggatt, J., Hu, P., Speth, J.M., Fukuda, S., Breyer, R.M., Pelus, L.M., 2012. Blockade of prostaglandin E2 signaling through EP1 and EP3 receptors attenuates Flt3L-dependent dendritic cell development from hematopoietic progenitor cells. Blood 119, 1671–1682.

Singh, S., Kumar, L., Khetarpal, R., Aggarwal, P., Marwaha, R.K., Minz, R.W., Sehgal, S., 1997. Clinical and immunological profile of SLE: some unusual features. Indian Pediatr. 34, 979–986.

Smigielski, E.M., Sirotkin, K., Ward, M., Sherry, S.T., 2000. dbSNP: a database of single nucleotide polymorphisms. Nucleic Acids Res. 28, 352–355.

Smith, C.D., Cyr, M., 1988. The history of lupus erythematosus. From Hippocrates to Osler. Rheum. Dis. Clin. North Am. 14, 1–14.

Smith, D.A., Germolec, D.R., 1999. Introduction to immunology and autoimmunity. Environ. Health Perspect. 107, 661.

Somers, E.C., Thomas, S.L., Smeeth, L., Schoonen, W.M., Hall, A.J., 2007. Incidence of systemic lupus erythematosus in the United Kingdom, 1990–1999. Arthritis Care Res. 57, 612–618.

Ståhl-Hallengren, C., Jönsen, A., Nived, O., Sturfelt, G., 2000. Incidence studies of systemic lupus erythematosus in Southern Sweden: increasing age, decreasing frequency of renal manifestations and good prognosis. J. Rheumatol. 27, 685–691.

Stenson, P.D., Ball, E.V., Howells, K., Phillips, A.D., Mort, M., Cooper, D.N., 2009. The Human Gene Mutation Database: providing a comprehensive central mutation database for molecular diagnostics and personalised genomics. Hum. Genomics 4, 1.

Stoeckman, A.K., Baechler, E.C., Ortmann, W.A., Behrens, T.W., Michet, C.J., Peterson, E.J., 2006. A distinct inflammatory gene expression profile in patients with psoriatic arthritis. Genes Immun. 7, 583–591.

Strom, B.L., Reidenberg, M.M., West, S., Snyder, E.S., Freundlich, B., Stolley, P.D., 1994. Shingles, allergies, family medical history, oral contraceptives, and other potential risk factors for systemic lupus erythematosus. Am. J. Epidemiol. 140, 632–642.

Subramanian, S., Tus, K., Li, Q.-Z., Wang, A., Tian, X.-H., Zhou, J., Liang, C., Bartov, G., McDaniel, L.D., Zhou, X.J., 2006. A Tlr7 translocation accelerates systemic autoimmunity in murine lupus. Proc. Natl. Acad. Sci. 103, 9970–9975.

Takahashi, T., Tanaka, M., Brannan, C.I., Jenkins, N.A., Copeland, N.G., Suda, T., Nagata, S., 1994. Generalized lymphoproliferative disease in mice, caused by a point mutation in the Fas ligand. Cell 76, 969–976.

Takamura, T., Honda, M., Sakai, Y., Ando, H., Shimizu, A., Ota, T., Sakurai, M., Misu, H., Kurita, S., Matsuzawa-Nagata, N., 2007. Gene expression profiles in peripheral blood mononuclear cells reflect the pathophysiology of type 2 diabetes. Biochem. Biophys. Res. Commun. 361, 379–384.

Tan, E.M., Cohen, A.S., Fries, J.F., Masi, A.T., Mcshane, D.J., Rothfield, N.F., Schaller, J.G., Talal, N., Winchester, R.J., 1982. The 1982 revised criteria for the classification of systemic lupus erythematosus. Arthritis Rheum. 25, 1271–1277.

Tan, F.K., Zhou, X., Mayes, M.D., Gourh, P., Guo, X., Marcum, C., Jin, L., Arnett, F.C., 2006. Signatures of differentially regulated interferon gene expression and vasculotrophism in the peripheral blood cells of systemic sclerosis patients. Rheumatology 45, 694–702.

Tang, K., Fu, D.-J., Julien, D., Braun, A., Cantor, C.R., Köster, H., 1999. Chip-based genotyping by mass spectrometry. Proc. Natl. Acad. Sci. 96, 10016–10020.

Theofilopoulos, A.N., Dixon, F.J., 1985. Murine models of systemic lupus erythematosus. Adv. Immunol. 37, 269–390.

Tiffin, N., Hodkinson, B. and Okpechi, I., 2013. Lupus in Africa: can we dispel the myths and face the challenges?. Lupus, p.0961203313509296.

Tobón, G.J., Renaudineau, Y., Hillion, S., Cornec, D., Devauchelle-Pensec, V., Youinou, P., Pers, J.-O., 2010. The Fms-like tyrosine kinase 3 ligand, a mediator of B cell survival, is also a marker of lymphoma in primary Sjögren's syndrome. Arthritis Rheum. 62, 3447–3456.

Toussirot, É., Roudier, J., 2008. Epstein–Barr virus in autoimmune diseases. Best Pract. Res. Clin. Rheumatol. 22, 883–896.

Tsokos, G.C., Kammer, G.M., 2000. Molecular aberrations in human systemic lupus erythematosus. Mol. Med. Today 6, 418–424.

Uramoto, K.M., Michet Jr, C.J., Thumboo, J., Sunku, J., O'Fallon, W.M., Gabriel, S.E., 1999. Trends in the incidence and mortality of systemic lupus erythematosus, 1950-1992. Arthritis Rheum. 42, 46–50.

Urra, J.M., De La Torre, M., 2012. Cytokines and systemic lupus erythematosus. INTECH Open Access Publisher.

Valouev, A., Ichikawa, J., Tonthat, T., Stuart, J., Ranade, S., Peckham, H., Zeng, K., Malek, J.A., Costa, G., McKernan, K., 2008. A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. Genome Res. 18, 1051–1063.

Van der Pouw Kraan, T., Wijbrandts, C.A., Van Baarsen, L.G.M., Voskuyl, A.E., Rustenburg, F., Baggen, J.M., Ibrahim, S.M., Fero, M., Dijkmans, B.A.C., Tak, P.P., 2007. Rheumatoid arthritis subtypes identified by genomic profiling of peripheral blood cells: assignment of a type I interferon signature in a subpopulation of patients. Ann. Rheum. Dis. 66, 1008–1014.

Vilar, M.P., Sato, E.I., 2002. Estimating the incidence of systemic lupus erythematosus in a tropical region (Natal, Brazil). Lupus 11, 528–532.

Voelkerding, K.V., Dames, S., Durtschi, J.D., 2010. Next generation sequencing for clinical diagnostics-principles and application to targeted resequencing for hypertrophic cardiomyopathy: a paper from the 2009 William Beaumont Hospital Symposium on Molecular Pathology. J. Mol. Diagn. 12, 539–551.

Voight, B.F., Scott, L.J., Steinthorsdottir, V., Morris, A.P., Dina, C., Welch, R.P., Zeggini, E., Huth, C., Aulchenko, Y.S., Thorleifsson, G., 2010. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. Nat. Genet. 42, 579–589.

Voss, A., Green, A., Junker, P., 1998. Systemic lupus erythematosus in Denmark: clinical and epidemiological characterization of a county-based cohort. Scand. J. Rheumatol. 27, 98–105.

Vyse, T.J., Kotzin, B.L., 1996. Genetic basis of systemic lupus erythematosus. Curr. Opin. Immunol. 8, 843–851.

Vytopil, M., Ricci, E., Russo, A.D., Hanisch, F., Neudecker, S., Zierz, S., Ricotti, R., Demay, L., Richard, P., Wehnert, M. and Bonne, G., 2002. Frequent low penetrance mutations in the Lamin A/C gene, causing Emery Dreifuss muscular dystrophy. Neuromuscular Disorders, 12(10), 958-963.

Wadee, S., Tikly, M., Hopley, M., 2007. Causes and predictors of death in South Africans with systemic lupus erythematosus. Rheumatology 46, 1487–1491.

Wallace, D.C., 1999. Mitochondrial diseases in man and mouse. Science, 283(5407), 1482-1488.

Wang, Z., Gerstein, M., Snyder, M., 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. 10, 57–63. doi:10.1038/nrg2484

Wang, K., Li, M. and Hakonarson, H., 2010. Analysing biological pathways in genome-wide association studies. Nature Reviews Genetics, 11(12), 843-854.

Watanbe-Fukunaga, R., Brannan, C.I., Copeland, N.G., Jenkins, N.A., Nagata, S., 1992. Lymphoproliferation disorder in mice explained by defects in Fas antigen that mediates apotosis. Nature 356, 314–317.

Watford, W.T., Hissong, B.D., Bream, J.H., Kanno, Y., Muul, L., O'Shea, J.J., 2004. Signaling by IL-12 and IL-23 and the immunoregulatory roles of STAT4. Immunol. Rev. 202, 139–156.

Weiss, K.M., Ward, R.H., 2000. James V. Neel, MD, Ph. D.(March 22, 1915–January 31, 2000): Founder Effect. Am. J. Hum. Genet. 66, 755–760.

Werneke, S.W., Schilte, C., Rohatgi, A., Monte, K.J., Michault, A., Arenzana-Seisdedos, F., Vanlandingham, D.L., Higgs, S., Fontanet, A., Albert, M.L., 2011. ISG15 is critical in the control of Chikungunya virus infection independent of UbE1L mediated conjugation. PLoS Pathog 7, e1002322.

Wheeler, D.A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.-J., Makhijani, V., Roth, G.T., 2008. The complete genome of an individual by massively parallel DNA sequencing. nature 452, 872–876.

Whitacre, C.C., 2001. Sex differences in autoimmune disease. Nature America Inc 345 Park Ave South, New York, NY 10010-1707 USA.

Wold, B., Myers, R.M., 2008. Sequence census methods for functional genomics. Nat. Methods 5, 19–21.

Wong, M., Tsao, B.P., 2006. Current topics in human SLE genetics, in: Springer Seminars in Immunopathology. Springer, pp. 97–107.

Xu, W.-D., Pan, H.-F., Xu, Y., Ye, D.-Q., 2013. Interferon regulatory factor 5 and autoimmune lupus. Expert Rev. Mol. Med. 15, e6. doi:10.1017/erm.2013.7

Xue, Y., Chen, Y., Ayub, Q., Huang, N., Ball, E.V., Mort, M., Phillips, A.D., Shaw, K., Stenson, P.D., Cooper, D.N. and Tyler-Smith, C., 2012. Deleterious-and disease-allele prevalence in healthy individuals: insights from current predictions, mutation databases, and population-scale resequencing. The American Journal of Human Genetics, 91(6), 1022-1032.

Yandell, M., Huff, C., Hu, H., Singleton, M., Moore, B., Xing, J., Jorde, L.B., Reese, M.G., 2011. A probabilistic disease-gene finder for personal genomes. Genome Res. 21, 1529–1542.

York, M.R., Nagai, T., Mangini, A.J., Lemaire, R., van Seventer, J.M., Lafyatis, R., 2007. A macrophage marker, siglec-1, is increased on circulating monocytes in patients with systemic sclerosis and induced by type i interferons and toll-like receptor agonists. Arthritis Rheum. 56, 1010–1020.

Zhang, X., Bogunovic, D., Payelle-Brogard, B., Francois-Newton, V., Speer, S.D., Yuan, C., Volpi, S., Li, Z., Sanal, O., Mansouri, D., 2015. Human intracellular ISG15 prevents interferon-[agr]/[bgr] over-amplification and auto-inflammation. Nature 517, 89–93.

Zhang, S., Fukuda, S., Lee, Y., Hangoc, G., Cooper, S., Spolski, R., Leonard, W.J., Broxmeyer, H.E., 2000. Essential role of signal transducer and activator of transcription (Stat) 5a but not Stat5b for Flt3-dependent signaling. J. Exp. Med. 192, 719–728.

Zhong, H., Yang, X., Kaplan, L.M., Molony, C. and Schadt, E.E., 2010. Integrating pathway analysis and genetics of gene expression for genome-wide association studies. The American Journal of Human Genetics, 86(4), 581-591.

Zikherman, J., Hermiston, M., Steiner, D., Hasegawa, K., Chan, A., Weiss, A., 2009. PTPN22 deficiency cooperates with the CD45 E613R allele to break tolerance on a non-autoimmune background. J. Immunol. 182, 4093–4106.

Electronic references (URL)

Applied Biosystems Home Page. www3.appliedbiosystems.com/index.htm (accessed on 1 August 2016).

ApplyRecalibration. https://software.broadinstitute.org/gatk/guide/tooldocs/org_broadinstitute_gatk_tools_walkers_variantrecalibration_ApplyRecalibration.php (accessed on 30 July 2016).

Babraham Bioinformatics. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed on 5 June 2016).

Base Quality Score Recalibration. https://software.broadinstitute.org/gatk/events/slides/1503/GATKwh6-BP-3-Base_recalibration.pdf (accessed on 30 July 2016).

Brandt, M., 2010. Research shows why lupus may be more common in black, Asian people. http://scopeblog.stanford.edu/2010/04/13/lupus_finding/ (accessed on 9 September 2014).

ExAC Browser (Beta) | Exome Aggregation Consortium. http://exac.broadinstitute.org/ (accessed on 30 July 2016).

Exome Variant Server. http://evs.gs.washington.edu/EVS/ (accessed on 30 July 2016).

Genome Analysis Toolkit. https://software.broadinstitute.org/gatk/ (accessed on 30 July 2016).

HaplotypeCaller. https://software.broadinstitute.org/gatk/guide/tooldocs/org_broadinstitute_gatk_tools_walkers_haplotypecaller_HaplotypeCaller.php (accessed on 30 July 2016).

Helicos Home Page. http://www.helicosbio.com/ (accessed on 1 August 2016).

Indel-based Realignment. https://software.broadinstitute.org/gatk/events/slides/1212/GATKwh0-BP-2-Realignment.pdf (accessed on 30 July 2016).

Information derived from Pathway Analysis.

http://bioinformatics.mdanderson.org/MicroarrayCourse/Lectures09/Pathway%20Analysis.pdf (accessed on 20 October 2016).

Ingenuity Pathway Analysis. www.qiagen.com/ingenuity (accessed on 30 July 2016).

Ingenuity Pathway Analysis. (http://repository.countway.harvard.edu/xmlui/bitstream/handle/10473/4740/Ingenuity%20Pathwahttp://lsl.sinica.edu.tw/Services/Class/files/20111228.pdfys.pdf?sequence=1 (accessed on 20 October 2016).

Ingenuity Pathway data curation. http://lsl.sinica.edu.tw/Services/Class/files/20111228.pdf (accessed on 20 October 2016).

Mapping and duplicate marking. https://software.broadinstitute.org/gatk/events/slides/1212/GATKwh0-BP-1-Map_and_Dedup.pdf (accessed on 30 July 2016**).**

Marzillier, J., 2013. DNA Sequencing and The Human Genome Project. https://www.lehigh.edu/~inbios21/PDF/Fall2013/Marzillier_11132013.pdf (accessed on 6 May 2016).

Moore, B., 2011. VAAST: Deciphering Genetic Disease with Next-Generation Sequencing. http://www.slideshare.net/barrymoore/vaast-deciphering-genetic-disease-with-nextgeneration-sequencing (accessed on 30 July 2016).

NOVOCRAFT. http://www.novocraft.com/userfiles/file/Novocraft.pdf (accessed on 25 June 2016).

Otogenetics Whole Exome and RNA Next Gen Sequencing Services. http://www.otogenetics.com/ (accessed on 5 June 2016)

Quality control / Read quality with FastQC. http://chipster.csc.fi/manual/fastqc.html (accessed on 7 April 2016).

Recalibrate variant quality scores. https://software.broadinstitute.org/gatk/guide/article?id=2805 (accessed on 30 July 2016).

Solexa Home Page. http://www.solexa.com/ (accessed on 1 August 2016).

The Ingenuity Pathway Database. http://www.ingenuity.com/ (accessed on 20 October 2016).

Tripp, S., Grueber, M., 2011. Economic Impact of the Human Genome Project. http://www.battelle.org/docs/default-document-library/economic_impact_of_the_human_genome_project.pdf (accessed on 10 July 2016).

USeq, MiSeq, WeAllSeq...to Seq. http://weallseqtoseq.blogspot.co.za/2013/10/gatk-best-practices-workshop-data-pre.html (accessed on 30 July 2016).

Variant Quality Score Recalibration (VQSR). http://gatkforums.broadinstitute.org/gatk/discussion/39/variant-quality-score-recalibration-vqsr (accessed on 30 July 2016).

VariantFiltration.
https://software.broadinstitute.org/gatk/guide/tooldocs/org_broadinstitute_gatk_tools_walkers_filters_VariantFiltration.php (accessed on 30 July 2016).

Appendix A. Parameters for variant filtration using recalibration model. The parameters were used to calibrate variants to ensure that only variants of high quality are retained for further analysis.

```
java -Xmx8g -jar $GATK_HOME/GenomeAnalysisTK.jar \
-T VariantFiltration \
-R $RESOURCES/ucsc.hg19.fasta \
-V $DATA/output.recalibratedINDEL.vcf \
--filterExpression "MQ0 >= 4 && ((MQ0 / (1.0 * DP)) > 0.1)" \
--filterName "HARD_TO_VALIDATE" \
--filterExpression "DP < 5 " \
--filterName "LowCoverage" \
--filterExpression "QUAL < 30.0 " \
--filterName "VeryLowQual" \
--filterExpression "QUAL > 30.0 && QUAL < 50.0 " \
--filterName "LowQual" \
--filterExpression "QD < 1.5 " \
--filterName "LowQD" \
--filterExpression "FS > 150.0 " \
--filterName "StrandBias" \
-o $Output/filtered_SNPs.vcf
```