The molecular genetic basis of the association of *TNFSF4* with SLE

Harinder Manku

Submitted for the degree of Doctor of Philosophy

September 2013

Division of Medicine

Imperial College London

Declaration of originality

The conceptualization of the work presented in this thesis, its realization and its documentation are my own unless mentioned. Data collection and analysis were performed in collaboration with colleagues as follows: Taqman genotyping presented in chapter 3 was completed at the JDRF/WT DIL, Cambridge, UK by Helen Stevens and colleagues. Cell-lines from the BDA-Warren repository samples held at the same laboratory were provided in collaboration with Prof. John Todd and colleagues. Phlebotomy of UK-European samples was by Dr. Andrew Wong, formerly a research co-ordinator at Imperial College. Genotyping and some QC presented in chapter 4 were completed at the OMRF under the auspices of the SLE Genetics consortium. The complement of researchers who took part in this genotyping effort are numerous, but I mention Dr. Jennifer Kelly, Dr. Kenneth Kaufman, Dr. Swapan Nath, Prof. Carl Langefeld and Prof. Betty Tsao for their additional contributions to the trans-ancestral project. Extra African-American samples were provided by Dr. Jeff. Edberg and Prof. Robert Kimberly at the University of Alabama. Additional UK-European samples were provided by Prof. Tim Vyse. The variant calling pipeline presented in chapter 5 was assembled by Dr. Michael Simpson of Kings College London who provided additional guidance and expertise to enable processing of the data presented in chapter 5 of this thesis. Professor Tim Vyse contributed to experimental design for data presented in chapters 3-5.

Copyright statement

Abstract

The tumour necrosis factor ligand superfamily member 4 gene (*TNFSF4*), also known as OX40L, is an established susceptibility locus in the autoimmune disease systemic lupus erythematosus (SLE). Genetic association studies map polymorphisms that associate with disease, but linkage disequilibrium often hinders the identification of the actual casual allele(s) at a disease susceptibility locus. At *TNFSF4* genetic association studies had shown that an extended 100kb haplotype upstream of the coding region of the gene was associated with SLE risk. The principle aim of the project was to conduct genetic association analyses in cohorts with different ancestry in an attempt to fine map the *TNFSF4* association signal and thereby identify the causal genetic variants that underlie the genetic risk. Utilizing >17,900 subjects of European, African-American, Hispanic-American and Southeast Asian ancestry a transancestral fine mapping analysis was performed. The results demonstrate the strong association of *TNFSF4* risk alleles in all populations tested. The most consistent and strongest evidence of association came from the single nucleotide polymorphism (SNP), rs2205960-T ($P = 7.1 \times 10^{-32}$, odds ratio = 1.63). This variant was also associated with autoantibody production in three independent cohorts. *In silico* analysis of the DNA sequence encompassing rs2205960-T predicts it to form part of a decameric motif, which binds the RelA (p65) component of the NF-κB transcription factor complex. A second associated SNP, rs16845607-A in *TNFSF4* intron 1 was identified in Hispanic-Americans ($P = 9.17 \times 10^{-9}$, odds ratio = 2.06). In an attempt to further refine the association, resequencing was performed in 80 individuals who were selected on the basis of their genotype to carry risk or non-risk haplotypes upstream of *TNFSF4*. This sequencing study identified >200 novel variants, mostly small insertion-deletion polymorphisms indels. The data presented in this thesis largely resolves the genetic basis of the immediate upstream association signal observed at *TNFSF4* with SLE and will facilitate the unraveling of the molecular basis of this genetic risk in systemic autoimmunity.

Contents

List of tables

Chapter 1

Chapter 2

Chapter 4

Chapter 5

Appendices

List of figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Acknowledgements

My sincere thanks to all colleagues who contributed to the research presented in this thesis. In particular, members of the Immunogenetics group past and present and the numerous collaborators who allowed access to their lupus cohorts under the collective banner of the SLE Genetics Consortium (SLEGEN). I thank the Wellcome Trust for funding these doctoral studies through programme grant 085492.

I express my gratitude to my supervisor, Professor Tim Vyse, for guidance, help and much patience and support, over a drink or three. Finally, thank you friends and family for your constant support during this research.

Abbreviations

3′ UTR	3′ untranslated region
3C	chromosome conformation capture
5′ UTR	5′ untranslated region
AA	African-American
ACR	American College of Rheumatology
ADT	Assay Design Tool
AI	Amerindian
AIM	ancestry informative marker
AMI	Amerindian
AMR	Amerindian/Native American
ANA	anti-nuclear antibody
APC	antigen-presenting cell
AS	East Asian
ASW	African ancestry in Southwest USA (HapMap project)
BAM	Binary Alignment/ Map
B-ATF	B-activating transcription factor
BCR	B-cell receptor
BDA	British Diabetic Association
bp	base pair
C1q	complement component 1, q subcomponent
CEU	United States trios of northern and western European ancestry (HapMap project)
CGF	continuous gene flow
CHB	Han Chinese individuals from Beijing (HapMap project)
CHD	Chinese in metropolitan Denver, Colorado, USA (HapMap project)
CHS	Southern Han Chinese
ChIP	chromatin immunoprecipitation-sequencing
CI	confidence interval
cM	centiMorgan
CLM	Colombians from Medellin, Colombia
CNV	copy number variation
DC	dendritic cell
DD	death domain
df	degrees of freedom
DIL	Diabetes and Inflammation Laboratory
DNA	deoxyribonucleic acid

DNM	de novo mutation
EAE	experimental autoimmune encephalomyelitis
EBI-EMBL	European Bioinformatics Institute - European Biology Laboratory
emPCR	emulsion polymerase chain reaction
ENCODE	Encyclopedia Of DNA Elements
eQTL	expression quantitative trait locus
ESP	Exome Sequencing Project
EUR	European
FACS	fluorescence-activated cell sorting
FGT	four-gamete test
FIN	Finnish in Finland
GC	genomic control
GC	germinal centre
GLADEL	Grupo Latino Americano de Estudio de Lupus
GBR	British in England and Scotland
GWAS	genome-wide association study
HDAC	histone deacytalase
HISP	Hispanic
HLA	human leukocyte antigen
HumDiv	human diversity training set
HumVar	human variation training set
IBD	identical by descent
IBS	Iberian population in Spain
Ig	immunoglobulin
IGV	Integrative Genomics Viewer
indel	insertion deletion
IQR	interquartile range
JDRF/WT	Juvenile Diabetes Research Foundation/ Wellcome Trust
JPT	Japanese from Tokyo (HapMap project)
kb	kilobase
LCL	lymphoblastoid cell line
LD	linkage disequilibrium
LINE	long interspersed element
LR	long range
LWK	Luhya in Webuye, Kenya (HapMap project)
MAF	minor allele frequency
Mb	megabase
MHC	major histocompatibility complex
MI	myocardial infarction
MKK	Maasai in Kinyawa, Kenya (HapMap project)
MRL	Murphy Roths Large (mouse strain)
mRNA	messenger ribonucleic acid
MS	multiple sclerosis

MSA	multiple sequence alignment
MUSC	Medical University of South Carolina
MXL	Mexican Ancestry from Los Angeles USA
NCBI	National Center for Biotechnology Information
NGS	next generation sequencing
NHLBI	National Heart, Lung, and Blood Institute
NK	natural killer
nsSNP	non-synonymous single nucleotide polymorphism
OMRF	Oklahoma Medical Research Foundation
OR	odds ratio
PBL	peripheral blood lymphocyte
PBMC	peripheral blood mononuclear cell
PCA	principal component analysis
PCR	polymerase chain reaction
PKB	protein kinase B
PTS	parallel-tagged sequencing
PUR	Puerto Ricans from Puerto Rico
PWM	position weight matrix
QC	quality control
QQ	quantile- quantile
RA	rheumatoid arthritis
RACE	5′ random amplification of cDNA ends
RIN	ribonucleic acid integrity number
RNA	ribonucleic acid
Ro	ribonuclear antigen
SAM	Sequence Alignment/ Map
SELEX	Systematic Evolution of Ligands by Exponential Enrichment
SINE	short interspersed element
siRNA	small interfering RNA
SLE	systemic lupus erythematosus
SLEDAI	Systemic Lupus Erythematosus Disease Activity Index
SLEIGH	Systemic Lupus Erythematosus in Gullah Health
SLICC	Systemic Lupus International Collaborating Clin- ics
Sm	Smith complex (nuclear antigen)
STR	short tandem-repeat
SV	structural variant
TBM	transcriptional base modification
TCR	T-cell receptor
TF	transcription factor
TFBS	transcription factor binding site

TLR	toll-like receptor
TNF	tumour necrosis factor
TNFRSF	tumour necrosis factor receptor superfamily
TNFRSF4	tumour necrosis factor receptor superfamily, member 4
TNFS	tumour necrosis factor superfamily
TNFSF4	tumour necrosis factor (ligand) superfamily, member 4
TSI	Toscani in Italia (HapMap project)
TSLP	thymic stromal lymphopoietin
TSS	transcription start site
UAB	University of Alabama, Birmingham
UCSC	University of California, Santa Cruz
URL	uniform resource locator
UV	ultraviolet
VCF	Variant Call Format
WTCCC	Wellcome Trust Case Control Consortium
YRI	Yoruba people of Ibadan, Nigeria c(HapMap project)

Chapter 1

Introduction

1.1 Systemic lupus erythematosus – a systemic disease

1.1.1 SLE: Prototype of systemic autoimmunity

Over a century has passed since Paul Erlich proposed "horror autotoxicus": Antibodies against self-antigens that injure one's cells are not produced through lack of purpose. However, research over the past several decades has demonstrated the reality of the autoimmune phenotype, which is characterised by an aberrant immune response against the organism's own cells and tissues due to failed recognition. An array of organ-specific and systemic autoimmune disorders are now known to manifest with varying population prevalence. Only few conditions are truly systemic and autoimmune, of these, systemic lupus erythematosus (SLE, abbreviated to lupus) is a prototype. Lupus is a rheumatic trait with the potential to affect all major organ systems with wide-ranging phenotypic heterogeneity. In common with many autoimmune disorders it is most likely to follow a benign course with intermittent or sporadic relapses (flares). Clinical manifestations of the disease are diverse, although high-affinity IgG autoantibodies to an array of nuclear antigens are a unifying feature of pathogenesis. Evidence suggests activation of complement and the components of immune regulation pathways to additionally characterise the global perturbation of the immune system found in SLE.

1.1.2 SLE: An overview of clinical features

Lupus is a systemic disorder with varied presentation of aetiologic features: Frequently observed benign disease manifestations include inflammation of the skin and joints, whilst impairment of the kidney and central nervous system are hallmarks of severe and active disease (Rozzo et al., 1996). The most extensively used SLE classification criteria were revised by the American College of Rheumatology (ACR) in 1982 and details of these criteria are found in Appendix A (Tan et al., 1982).

Patients need present four of eleven criteria during any interval of observation, a major obstacle in standardised clinical research, which has highlighted combinations of lupus sub-phenotypes more suited to classification independently. Given the poorly understood aetiology of SLE per se, sub-phenotypes may be amenable to study as more homogenous groups as they often result from single or related aberrations during pathogenesis. Clinical subsets that have been researched independently include specific autoantibody subsets, renal disease and age at diagnosis.

1.1.2.1 Autoantibody subsets

Serologic studies indicate antinuclear antibodies (ANA) are found in 95% of lupus individuals, whilst antibodies to double-stranded DNA (dsDNA-Ab) are found in 60% of cases. In a subset of patients, presence of dsDNA-Ab and a concomitant depression of complement C3 pro-activator levels are predictors of disease flares through immune complex formation (Tan et al., 1966). Autoantibody subsets are valuable predictors of sub-phenotype; Anti-Ro antibodies are associated with photosensitivity (Mond et al., 1989) and subcutaneous lupus erythematosus in Europeans (Lee et al., 1989) and Anti-Sm

antibodies are associated with renal disease in Afro-Caribbean, African-American (Alba et al., 2003) and Hispanic populations.

1.1.2.2 Renal disease

Autoantibody (IgM, IgG and IgA)-containing immune complexes deposit in, and cause damage to the kidneys leading to the lupus nephritis phenotype (Leavy, 2010). The most frequent diagnosis of renal disease in lupus is grade IV lupus nephritis, diagnosed by histopathology in a third of UK lupus individuals (Cortes et al., 2008) according to guidelines in the 2003 classification of lupus nephritis. Renal pathology is a hallmark of poor prognosis and an indicator of increased disease burden. This phenotype segregates with increased frequency in non-European SLE populations. Epidemiological data from the multi-ethnic LUMINA (Lupus in Minorities: Nature vs. Nurture) cohort illustrate the trend: After adjustment for clinical and demographic variables, there is decreased time to lupus nephritis in African-American and Hispanic SLE cases (Burgos et al., 2011). An epidemiological study of a Canadian multi-ethnic cohort also suggests increased incidence of lupus nephritis in African-American and South Asian lupus cases (Peschken et al., 2009). Resolution of the mechanisms which cause lupus nephritis in subsets of cases will improve the therapeutic treatment of severe disease.

1.1.2.3 Age at diagnosis

Subtypes of SLE can be categorised by age at diagnosis into neonatal, paediatric and late-diagnosis of disease (Simard and Costenbader, 2007; Tucker et al., 1995). Nearly 15% of cases present in children before the age of 16; paediatric cases have more frequent haematological and renal manifestations early after disease onset and a higher frequency of elevated dsDNA antibodies which predict a severe and active disease course (Tucker et al., 1995). Early expression of the trait could be a result of increased expression of genetic aetiologic features: Onset at age 50 or after is associated with reduced disease activity and evidence suggests reduced severity of the trait when diagnosed in post-

menopausal women (Urowitz et al., 2006; Simard et al, 2011). Lupus is diagnosed at a younger age in non-European populations: In a multi-ethnic UK cohort, mean age of onset of SLE was 28.9 years in South Asians, 32.9 years in Afro-Caribbean's and 36 years in UK Europeans (Chambers et al., 2007). The trend repeats in the aforementioned large multi-ethnic Canadian cohort, though the South Indian group presents with lower lupus damage scores, possibly due to earlier access to healthcare (Peschken et al., 2009).

1.1.3 SLE Epidemiology

Classical epidemiological studies in SLE find segregation of disease with gender, geographical location and race: New, large-scale datasets are powering definition of ever narrower subtypes within these categories. Current research is also directed in populations with little or no past definition with regards to the lupus phenotype, including Amerindian populations from North (Houghton et al., 2006) and South (Seldin et al., 2008) America. In the current era of global SLE research, better-defined epidemiological data are required from these populations: A research effort underway in Central and South America using the GLADEL multinational Latin American SLE cohort (Pons-Estel et al., 2004) is attempting to address unknown aetiology in Latinos. The updated global data do not include measurements of SLE prevalence in Africa and South Asia as the research is at a very early stage in these groups. Epidemiological studies in Northern and Western European SLE populations have directed the majority of studies on lupus pathogenesis.

There is a gender imbalance in lupus; the ratio of affected, reproductively-able females to males is 9 to 1. This pattern is less pronounced but still evident in early- and late-onset groups (Masi and Kaslow, 1978). The global and racial incidence and prevalence patterns follow a similar trend with a bias in non-European populations. The latter include groups with admixed ancestry (defined as populations with recent ancestry from two or more continents) such as the African-American and Latino/Hispanic groups. No clear North–South or East–

West pattern emerges from these data (Danchenko et al., 2006; Simard and Costenbader, 2007).

1.1.3.1 SLE prevalence gradient

Despite the high disease load in African-Americans, there is the perception that lupus is relatively rare in continental Africans: A 'prevalence gradient' between sub-Saharan Africa, where SLE prevalence is reportedly low, and western countries, where these same populations have more disease, has been described by Bae and colleagues (Bae et al., 1998). Increased competition from morbidity/mortality factors due to infection, and reduced survival time and difficulties diagnosing the lupus phenotype in Africa challenge the current estimates of African SLE prevalence. Thus, questions remain on the accuracy of this gradient. High prevalence and early-onset morbidity in African-admixed, Hispanic and South Asian populations probably reflect increased expression of genetic aetiology and socioeconomic factors related to poverty and limited access to care (Molina et al., 1997; Fernandez et al., 2007).

The majority of genetic association studies in Africans have been undertaken in East, Central and North African populations, which are not the ancestral areas of origin of most African-Americans. Significant genetic differences exist between native African populations (Gilkeson et al., 2011). To better define the SLE prevalence gradient, and to establish the impact of environmental factors on African SLE prevalence, parallelised studies of Europeans and West-Africans are required. However, defining the disease in West Africa is challenged by minimal health care systems (Kushner et al., 2010). A confounding factor when using African-American populations, which predominantly descend from West-Africa, is the significant and variable genetic admixture (10-30%) with Europeans. A second informative group is the Gullah population of the Sea Islands of South Carolina. The Gullah have lower genetic admixture (<10%) due to geographical isolation and strong cultural heritage. Anthropologic studies indicate a direct ancestral link between the Gullah and Sierra Leoneans

(Gilkeson et al., 2011). Epidemiological data collected by the Gilkeson group from Gullah and Sierra Leonean groups suggest the presence of autoantibodies to nuclear antigens in the sera of both groups but a lack of progression to disease in the Sierra Leone group.

1.1.4 Immune dysregulation

1.1.4.1 B-cells, autoantibodies and immune complexes

A perception for the role of B-cells as the main drivers of SLE pathogenesis has remained robust for the past several decades, perhaps due to the ability of terminally-differentiated plasma cells to produce auto-reactive antibodies. Auto-antibodies, as mediators of the immune pathology underlying the disease process, are substantive contributors to lupus. The (auto)-antigen- presenting capability of an activated B-cell is also likely to direct pathogenesis prior to auto-antibody production.

B-lineage cells induce or maintain SLE through the secretion of inflammatory mediators, presentation of auto-antigen to CD4+ T-cells and production of antibodies to nuclear self-antigens (Minton, 2011). The latter is mediated by plasma cells, final differentiated cells in the B-lymphocyte pathway, which produce autoantibodies to directly or indirectly interfere with cellular function and evoke immune pathology: Fc-mediated activation of complement followed by recruitment of inflammatory cells is an example of direct interference. Antibody-dependent mechanisms can indirectly mediate end organ damage through immune complex (IC) formation. In lupus, IC-activation of complement results in their deposition in the kidneys, and deposition correlates with progression to lupus nephritis. IC-activation of the FcγRIII-dependent pathway primes a range of cells which control immune functionality, including plasmacytoid DCs which secrete the pathogenic cytokine IFNα (Martin and Chan, 2006). IC can also activate a range of pathways of the innate and adaptive immune systems inappropriately in SLE.

1.1.4.2 T-cells

$CD4^+$ T-cells

T-cells perpetuate autoimmunity in several organ-specific diseases, however, for several years, B-cells were believed to be the predominant drivers of lupus pathogenesis. Evidence suggests T-cell contribution to the initiation and perpetuation of SLE, in addition to pathology (Engler et al., 2011; Xu et al., 2004). Cognate interaction between B-cells and T-cells is likely to feature in lupus pathogenesis and directly drive immune pathology in the trait: T-cells from animal models and human cases have altered attributes including differences in homing (Lyons et al., 2010), aberrant signalling and transcription factor binding. $CD4^+$ T-cells interpret (auto)-antigen recognition, steering the cells they influence to become regulatory or pro-inflammatory: Switching is likely to be aberrant in SLE and the overall T-cell compartment perturbed. As a result, $CD4^+$ T-cells inappropriately activate B-cells and dendritic cells to secrete cytokines which promote inflammation.

$CD4^+$ T regulatory cells

CD4+ T regulatory cells (T_{regs}) are necessary for maintenance of immunological tolerance: Absence results in severe autoimmunity (Feuerer et al., 2009). Distinct lineages of T_{regs} are altered in lupus and these include 'natural' $CD4^+CD25^+FoxP3^+$ cells and an inducible $T_r1$subset which tolerise by producing large quantities of IL-10 (Ito et al., 2006).

1.2 Complex trait analysis

Association or linkage disequilibrium (LD)-based mapping is widely used to efficiently locate genes that influence complex traits. The best possible mapping data are obtained if the genealogical history of the sampled individuals is explicit: An accurate kilobase scale map of the recombination rate improves mapping accuracy for pin-pointing causal variations and identifying multiple independent contributors to risk at a single locus: The genetic association study is closely tied with indirect LD-based recombination maps and aims to identify statistical associations between candidate genetic polymorphisms and complex disease. The HapMap (The International HapMap Consortium., 2003; 2005; 2010) and 1000 Genomes (1000 Genomes Project Consortium., 2010; 2012) projects accelerated the development of tools used to map causal contributors to disease risk: I describe the rationale for the aforementioned association studies, including use of single nucleotide polymorphisms (SNPs) and haplotypes to infer population-scale genotypes by imputation. I give a brief overview of the current and historical techniques used to map the established genetic associations in complex disease.

1.2.1 Human genetic variation

1.2.1.1 Single nucleotide polymorphisms

Though any two human genomes are >99% identical, analysis of human DNA sequences has revealed significant genetic differences within and amongst populations. These variations are responsible for heritable changes, including disease susceptibility in individuals (Kruglyak and Nickerson, 2001). About 90% of genetic variations in humans are single base-pair substitutions which occur at a minor allele frequency of >1%, these are defined as single nucleotide polymorphisms (SNPs) (Figure 1.1). Whole genome multiplexed sequencing of random clones has facilitated early discovery of SNPs. Large multidisciplinary efforts to identify and characterize SNPs have subsequently been undertaken by the International HapMap Project and the 1000 Genomes Project (The International HapMap Project., 2005; 1000 Genomes Project Consortium., 2010).

The 1000 Genomes Project has used multiplexed sequencing of the whole genome in 1092 individuals from 14 populations to generate the most detailed catalogue of human genetic variation to date (1000 Genomes Project Consortium., 2010; 1000 Genomes Project Consortium., 2012). The total number of SNPs identified by this project is >30 million owing to individuals from different populations having distinct SNP profiles for both rare and common variants. Up to 98% of accessible single nucleotide polymorphisms are captured at a frequency of 1% in related ethnic groups as part of this project (Figure 1.2). In addition to SNPs, there are many novel or '*de novo*' single nucleotide variants and rare variants; in some cases these variants have been identified within a single nuclear family (Frazer et al., 2009).

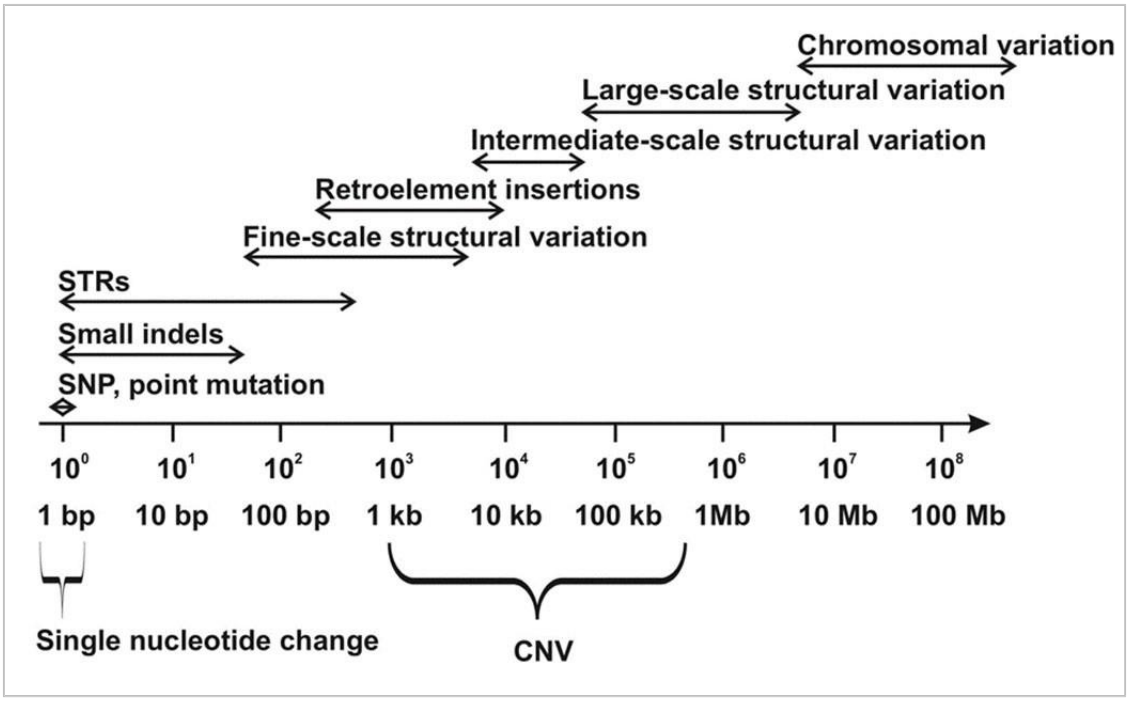
1.2.1.2 Structural variations

Structural variants (SVs) are the other pervasive class of inherited variations: Included are 1.4 million short insertions or deletions (indels) of nucleotide bases in the DNA sequence (1000 Genomes Project Consortium., 2012). Larger structural variants, including inversions and copy number variants (CNV), are also major contributors to human genetic variation (Figure 1.1). Although they occur at lower frequency than SNPs, the fraction of the genome SVs affect is comparatively large (Conrad et al., 2010). Structural variants have significant consequences on phenotypic variation: A genome-wide map of CNVs, based on sequencing data from 185 whole human genomes, encompassing 22,025 deletions and additional insertions and tandem duplications, has facilitated mapping of these variants in disease traits (Mills et al., 2011).

1.2.2 Heritability and genetic variance

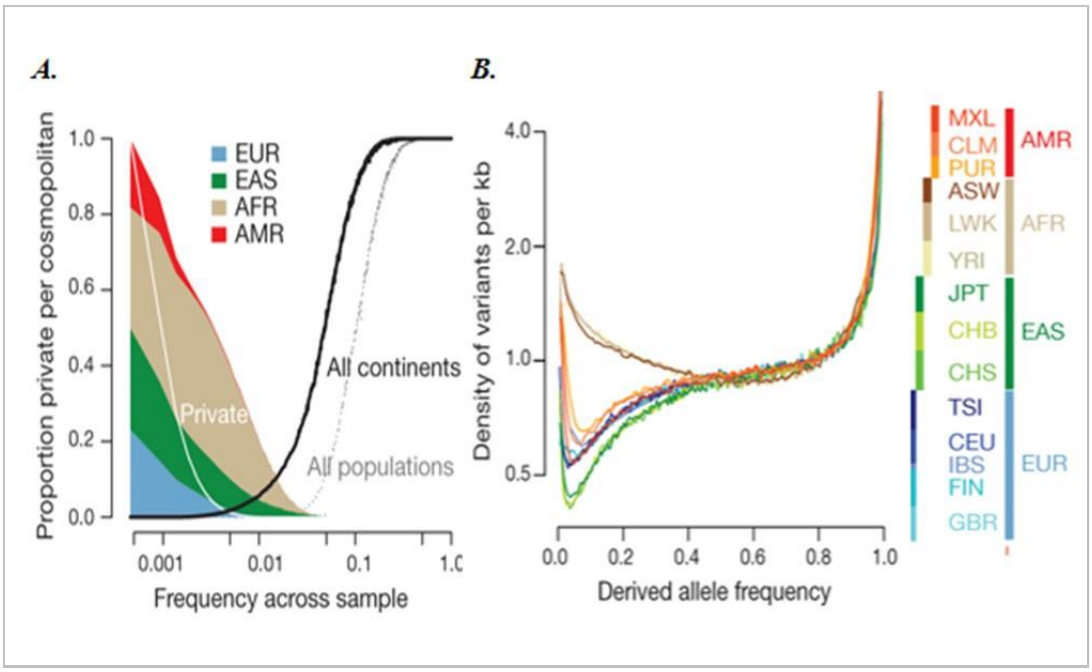
The heritability (h^2) of a phenotype is a useful indicator of the genetic aetiology for the trait and is defined as the ratio of the genetic variance to the total phenotypic variance amongst individuals with a common shared environment. Estimating h^2 for human phenotypes often involves comparison of the concordance rates for monozygotic twins against dizygotic twins. Using the h^2 parameter as a tool to assess genetic susceptibility to complex traits is useful, however accurate estimation can be impacted by increased hidden MZ shared environment, or if high risk pedigrees are used (Visscher et al., 2008). Heritability is sometimes incorrectly equated to the total genetic variance: Only a small proportion (<10%) of genetic variance that explains the genetic susceptibility to most complex diseases has been identified (Frazer et al., 2009). This may be because the allelic effects of the variants are small, because there are gene-gene and gene-environment epistatic interactions or because the causal variations are rare but highly penetrant (Eichler et al., 2010).

Figure 1.1 The spectrum of variation in the human genome



Depiction of genetic variations with their size range (double-headed arrows). SNP indicates single-nucleotide polymorphism; indels, insertions and deletions; STR, short tandem repeat and CNV, copy number variation. SNPs and point mutations apart, the size ranges of the variations are not definitive. A logarithmic x-axis measures the number of nucleotides, from 1 bp to ≥100 Mb. (Figure by Pollex and Hegele., 2007).

Figure 1.2 The distribution of rare and common variants identified by the 1000 Genomes Project in *A*. populations and *B*. sub-populations



A. The fraction of variants identified across the 1000 Genomes Project that are identified in one population (white line), identified in one ancestry-based group, found in all groups (solid black line) and found in all populations (dotted black line). *B*. The density of the expected number of variants per kilobase per individual genome drawn from each population, as a function of variant frequency. Key with colour-coding for all groups and populations to the right of figure) (The 1000 Genomes Project Consortium., 2012).

1.2.3 Recombination: An overview

The Holliday model of recombination suggests homologous pairing of chromosomes is followed by crossing-over of sections of DNA (Holliday, 1964). Cross-over events cluster into narrow 2kb regions coinciding with a breakdown of LD (hotspots) and point to an LD pattern consisting of blocks of correlated DNA bases termed haplotypes (Amos et al., 1968; Piazza et al., 1969; Gabriel et al., 2002; Jeffreys and May, 2004). Extending the pattern to the entire human genome has allowed inference of the recombination pattern (Gabriel et al., 2002). The boundaries of blocks and the specific haplotypes they contain are correlated across populations to provide statistical power in association studies of common genetic variation across each region. Classical quantification of recombination rate has used pedigree-based data (Clark et al., 2010). The first comprehensive genome-scale linkage maps have been constructed using STRP markers in eight CEU families (Dib et al., 1996). The CEU map has greatly enhanced the ability to localize and identify genes for inherited disorders and revealed extensive variation in the recombination rate per unit physical length across the genome. Significant variation in the recombination rate per meiosis has been found in females but not males with peak differences at metacentric centromeres (Broman et al., 1998, Murray et al., 1994). The sex-specific variations identified have been consistent across all chromosomes (Dib et al., 1996).

1.2.4 Estimation of fine-scale recombination rates from human population variation data

The expense of large-scale crossing-over experiments has limited the resolution of recombination to the megabase scale. Improvements in genotyping efficiency have enabled genome-wide experiments including the HapMap diversity study and the first inferred recombination maps. These maps present recombination at the kilobase scale; they find increased local rate variation, with the majority of recombination events in small two kilobase segments of sequence called hotspots.

The first recombination maps in non-Europeans have been inferred and illustrate features that are unique in each population.

1.2.5 Model-based inference

The program *rhomap* adopts the coalescent model (Stumpf and McVean, 2003), which treats pairs of adjacent SNPs as two bi-allelic loci: The loci are used to score a recombination event using the four gamete test (FGT). *Rhomap* infers the local recombination rate from the decay of LD between adjacent SNPs. However, LD is shaped by many additional factors: Incorporation of a composite likelihood method allows a range of recombination rates for pairs of SNPs and priors are included to avoid over-fitting and to support short-range smoothness (McVean et al., 2004). *Rhomap* uses the product of per-generation recombination rate, r, and the effective population size, N_e to gives the estimated population recombination rate, ρ. Simulations of HapMap samples using *rhomap* establish recombination as a general feature at the kilobase scale with hotspots occurring ubiquitously at 200kb intervals (Stumpf and McVean, 2003). The inferred hotspots are likely to be located in intergenic regions as opposed to genes.

1.2.6 Comparison of recombination maps across ancestral groups

Recombination maps built in people of European descent have limited use when extended to Non-Europeans: In a 2005 study, microsatellite-based genetic maps constructed in Europeans, African-Americans, Mexican Americans and East Asians found excess map length in African- Americans and East Asian females (Jorgenson et al., 2005). Locus-specific factors influenced by demographics, including natural selection, may also bias LD-based estimates: The next generation deCODE recombination map, which has been constructed with family-based genome-wide SNP data to identify 15,257 meiosis, has been contrasted with the 2005 CEU and YRI HapMap LD-based maps. Using regression to map differences, clear differences exist between the European and West African recombination pattern (Kong et al., 2010; The International HapMap Project., 2007). Decay of LD indicates increased recombination and reduced concentration of crossovers in the West African genome (Hinch et al., 2011). In this population

70% of recombination occurs in 10% of the sequence rather than 80% of the recombination in 10% for Europeans and East Asian individuals (Leavy, 2010).

1.2.7 Mapping causal genes in disease: parametric linkage-based mapping

Flanking restriction fragment length polymorphisms (RFLPs) or highly polymorphic microsatellite markers are typed in multigenerational clear-cut pedigrees with two or more disease-affected family members to identify sections of the genome which co-segregate with disease (Lander and Schork, 1994). Linkage studies have successfully mapped simple Mendelian traits, including Cystic Fibrosis (Botstein and Risch, 2003). Positive selection of the inherited phenotype is a key requirement for this positional method: Researchers had little or no prior knowledge of the biology behind the trait. The first comprehensive tools for correlating phenotype with DNA sequence were provided by the human genetic linkage map. However, identification of the underlying disease gene has powered limited progress in therapeutics by informing on molecular/physiological mechanisms of mainly rare, highly penetrant diseases.

Linkage disequilibrium (LD), is the non-random association of alleles at two or more genetic loci, which is dependent on identity by descent (IBD), has been used to narrow defined linkage regions: The inherited DNA progressively shortens over many generations due to increased recombination. Markers in LD with disease-associated alleles segregate at the same frequency as the mutant allele non-randomly in the founder population. LD places the risk variant in a much narrower interval relative to traditional linkage methods, thereby greatly increasing the resolution of the linkage map. The first studies using this method identified loci in rare, autosomal recessive disorders as the associated alleles are less likely to be negatively selected (Friedman et al., 1995).

1.2.8 LD-based mapping: haplotypes

Resolution of haplotypes across the human genome has accelerated evaluation of the more elusive genes which underlie complex disease (Service et al., 1999). The first haplotypes were identified as early as 1968 on chromosome 6p, spanning the HLA super locus (Amos et al., 1968): These are series of adjacent markers, non-randomly inherited together, thus in strong LD, and considered an independent entity. Mapping using haplotypes involves genotyping many adjacent markers which span a genomic region, followed by analysis of haplotype frequency between disease-affected and non-affected groups. Historic cross-over points can segregate with disease-associated loci: Decay of haplotype sharing permits their identification as causal contributors to disease (McPeek and Strahs, 1999). The identification of genes in these early studies has heavily depended on low etiologic heterogeneity within the affected samples: In complex diseases, including lupus, the trait is usually the product of several independent genetic loci or alleles and successful resolution of contributing loci has been limited.

1.2.9 Pattern and structure of SNP-based haplotypes in the human genome

The conserved pattern and the structure of SNP-based haplotype blocks across the human genome offer a powerful approach for genetic mapping studies: Disease-associated SNPs, on specific ancestral haplotypes, are transmitted to the next generation and are conserved within each population, though they can be altered by mutation or meiotic recombination (Amos et al., 1968; Gabriel et al., 2002). The boundaries of haplotype blocks coincide with recombination hotspots and so they have variable length. Within each block, for most populations, only a few common haplotypes are observed. Selecting SNPs which best tag each haplotype allows inexpensive yet efficient study design to capture all common sequence variation within the target genomic region (Johnson et al., 2001).

1.2.10 Variation of human haplotypes across population samples

A collective effort to decipher human haplotype variation amongst populations has significantly contributed to genetic research into complex disease. Data generated by Gabriel and colleagues provides a basic framework for genetic mapping studies: Candidate SNPs in 51 autosomal regions across populations of distinct ancestry were surveyed by this research. Significant genome-wide differences in SNP incidence and allele frequency are found across ancestral populations (SNPs incidence, 70% (East Asians) up to 86% (African-Americans)) (Gabriel et al., 2002). The African-American population have the proportion of adjacent SNPs with most reduced proximal distance; indirect evidence for an increased rate of historical recombination in this group. The pattern is reversed for adjacent SNPs separated by 22kb or more. Collectively, these data suggest the African-derived population to have shorter haplotypes. Differences in LD amongst the aforementioned populations translate to a minimum average span of 9kb for African-derived haplotype blocks compared to 18kb for East Asians and Europeans.

1.2.11 Genetic association (candidate gene) studies

The past 15 years has increased the number of SNP variants from the thousands to the tens of millions: The first dense maps of SNPs, which uniquely tag common haplotypes, were generated for the human genome in 1999-2000. SNP maps have driven high-resolution identification of candidate genes in complex disease. (Altshuler et al., 2000; Mullikin et al., 2000; International HapMap Consortium, 2005; Cunninghame Graham et al., 2008). The genetic association study tests allele or genotype frequencies of one or more polymorphism between two groups of individuals with common ancestry. Classically, these groups are diseased subjects and healthy controls who are assumed to be independent non-related individuals. The allele/genotype is associated with the trait if it occurs at higher frequency in the cases than the controls. Three likely scenarios result in association between a genetic polymorphism and a trait in a given population: The polymorphism has a causal role; the polymorphism has no direct causal role

but is a surrogate/proxy marker in LD with a nearby causal variant (indirect association); underlying stratification or admixture of the population cause spurious association of the polymorphism with the trait (Cordell and Clayton., 2005).

1.2.12 Addressing admixture

Increased prevalence of complex disease, including SLE, in admixed populations has been documented by many studies (Burgos et al., 2011). Owing to this, multi-ethnic cohorts have become increasingly common over the past several years. Overrepresentation of disease in admixed populations presents the additional challenge of controlling stratification to allow close genetic matching of cases with controls. Stratification occurs due to systematic differences in allele frequency and phenotype distribution between subgroups of the population. In admixed populations, this variability might be ascribed to different proportions of the source genomes between cases and controls. Exploring the distribution of admixture in the AA and Hispanic groups finds that it occurs at a steady rate in each generation, resembling patterns observed in the continuous gene flow (CGF) model of admixture (Pfaff et al., 2001).

African-American and Hispanic chromosomes contain segments of their respective source genomes which can be distinguished as local ancestry. Often, the ancestral populations have risk alleles with large enough differences in frequency to easily be mapped at a causal locus by using the genotypes of ancestry informative markers (AIMs) or GWAS chips. AIM panels are designed to include markers with large differences between ancestral groups which are not linked within each group (Pfaff et al., 2001).

The mosaic-like structure of the aforementioned admixed genomes might better delineate causal variants in complex disease. However, these genomes can also give false positive associations. Categorising populations by self-reported

ancestry alone is an imperfect control for genetic heterogeneity as any underlying population substructure would inflate the association test statistic. Even modest levels of bias can distort the null distribution to overwhelm evidence of true association. Systematic bias can be visualised by plotting a quantile-quantile plot of the observed against the expected distribution of the p-values using AIMs to generate test statistics for the data. To resolve the potential confounding effects which might undermine novel associations, the first analytic method frequently used to correct raw genotype data is genomic control (GC). GC is the median χ^2 (1 degree of freedom) association statistic across SNPs divided by its theoretical median under the null distribution, a value of $\lambda_{GC} > 1$ indicates stratification (Devlin and Roeder, 1999; Devlin et al., 2004). Within a tightly matched ancestral group the allele frequencies are within narrow confines, these data are suited to GC alone: Genomic control is not sufficient to control for stratification in groups with mixed continental ancestry owing to increased deviations of allele frequencies, and should be used as a checkpoint after better-suited methods have been used (Lohmueller et al., 2006).

1.2.12.1 Correcting population substructure in admixed populations

Population substructure can be corrected intuitively using clustering algorithms followed by the use of regression as a covariate within the analysis framework (Sankararaman et al., 2008). Estimates of local ancestry by programs including STRUCTURE (Pritchard et al., 2000) and LAMP (Sankararaman et al., 2008) can be used for admixture mapping. The accuracy by which the clustering algorithm corrects stratification is difficult to predict when applied to the composite genomes of recently admixed Hispanic populations; they often have four-way admixture and thus available source information can lack clarity.

The prevailing method for dealing with population stratification in admixed populations over the past several years has been by a principal components (PC)-based approach such as that used by the Eigenstrat tool (Price et al., 2006). Continuous axes of genetic variation are inferred from population-scale genetic data, using AIM or GWAS genotypes. The variability is ascribed a number of

dimensions, the first dimension corresponds with the first principal component which describes the biggest change in allele frequency between subgroups. If there are ancestral differences within a population, plotting the first two principal components illustrate major differences between the groups, this might correspond to a significant geographical change. The second, third and even fourth principal components might be compared with each other to correct for subtle but significant population differences.

1.2.13 Inferring population-scale genotypes: Imputation

Missing genotype data in a candidate gene association study can challenge modelling the effects of multiple genetic variants (D'Angelo et al., 2010). The prevailing approach for dealing with these incomplete data is by imputation. Genotype imputation is defined as the prediction of polymorphic variants that have not been assayed in an association study. High-throughput, low-cost genotyping has accelerated the development of several powerful imputation approaches to estimate genotypic or haplotypic effects in large datasets (Li et al., 2006; Marchini et al., 2007; Marchini and Howie, 2008; Marchini and Howie, 2010). Combining observed and missing genotypes (MAF>1%) and predicting the missing data from the observed genotypes in the presence of a fine-scale recombination map and high-density reference genotype panel generates allows inference of genotypes which concur with high confidence.

The IMPUTE program, devised by Marchini and colleagues, is reliant on the genotype calling algorithm CHIAMO, and has been widely applied to impute missing data for SNPs in genetic association studies (Marchini et al., 2007). Prior to imputation, genotyped SNPs are phased and resulting haplotypes are then compared to dense reference haplotypes such as those from the HapMap phase III or 1000 Genomes panels. The comparison involves modelling a mosaic of haplotypes of other individuals and imputing missing genotypes to match them. The uncertainty and probability distributions over three possible genotypes are used to evaluate SNPs for quality. Well-imputed genotypes progress to association analyses. The incorporation of a flexible modelling framework has

resulted in increased accuracy of imputation by accounting for phasing uncertainty independently. IMPUTE2 has been used to combine information across multiple reference panels potentially enabling the inclusion of thousands of chromosomes to reduce errors at common and lower frequency, but not rare SNPs (Howie et al., 2009).

Although power is not greatly boosted by imputing sporadic missing data, false positives at difficult to genotype SNPs are controlled (Marchini and Howie, 2010). Imputation has been used to infer high frequency SNPs from 1958 British Birth Cohort genotypes: Imputation gave a maximum posterior genotype call rate of 0.998, illustrating the high-confidence of well-imputed genotypes (Marchini et al., 2007). This method fine-maps a panel of markers so that SNPs which have not been genotyped but which may be causal, are more likely to be detected. Imputed, associated SNPs can have slightly overinflated association values and the quality of the inferred data is limited if there are increased or unexpected recombination events in the assayed region. The tool has worked well to equate SNPs from independent datasets to compare genotypes across different platforms. Equating the variants by imputation has facilitated meta-analysis to increase the power to detecting causal variation for complex traits.

1.2.14 Statistical power to detect associations

The power of a statistical test is the probability that it will reject the null hypothesis and detect a statistically significant association. A number of factors constrain the power to detect a genetic association: The genotype relative risk (GRR) (the ratio of the risk of disease between individuals with and without the genotype), the sample size and study design can degrade power (Evans and Purcell., 2007). *A priori* power calculations are often used to determine the sample size required to detect the effect. Sample size is under the investigator's control; however effect size can also be increased through genotyping a region of interest more densely.

Examining all observed haplotypes at a locus by capturing their tag SNPs has the trade-off of reduced power to detect common causal alleles: This is because significance levels are set at a higher level to compensate for the increased number of statistical tests. The number of tests can be decreased by ranking tag SNPs according to the number of other SNPs for which they can act as proxy and analysing only the highest ranked variants to maintain the relative power. Tag SNPs selected from high density data give a distribution shifted towards higher χ^2 values as causal SNPs are likely to be captured (Ardlie et al., 2002; de Bakker et al., 2005). Selecting the best tag SNP from a reference genotype panel can be undertaken using statistical programs developed for haplotype analysis (Service et al., 1999; McPeek and Strahs, 1999).

1.2.15 Pair-wise correlations between polymorphic variants

The popular Haploview software enables the selection of tag SNPs with a rank indicator viewed with the generated haplotypes on the interface. The quality metrics can be adjusted for different datasets using a range of adaptable software tools to generate single variant and haplotype association statistics. Haploview calculates several pairwise measures of LD. The multi allelic D′ represents the degree of LD between two loci and is dependent on the frequencies of the alleles (Devlin and Risch, 1995): A D′ value of 0 indicates complete independence whilst a D′ of 1 indicates complete LD. An alternative, widely-used measure of LD, also included in Haploview, is the pair-wise correlation coefficient r^2 (Weiss and Clark, 2002). The r^2 measure has the advantage of adjusting for loci having different allele frequencies. The scale used for pair-wise LD representation of r^2 is the same as for D′.

1.2.16 Modelling the genetic association

The association of SNPs with disease status can be tested using allelic and genotypic tests: The genetic model selected must best suit the underlying correlation. The dominant genetic model dichotomizes SNP genotypes by treating heterozygotes and one of the homozygote genotypes as a single category: The single dominant allele is sufficient to confer risk so the two groups are modelled

as having the same risk (Lunetta., 2008). In the additive genetic model, each additional copy of the variant allele increases the risk of disease by the same amount. Thus, in Table 1.1, the homozygous A/A genotype confers double the risk compared to the heterozygous G/A genotype. Both these models require 1 degree of freedom (*df)* for association testing. The degrees of freedom can be defined as the arbitrary number of coefficients in the regression model. The general genetic model retains the three distinct genotype classes and makes no assumptions on how the risk varies between these classes, but has the trade-off of requiring 2*df* for association-testing.

	Genotype					
	GG	GA	AA	χ^2	*df*	*P*
Crude OR (vs AA)	1	0.79	2.12	11.36	2	3.40×10^{-3}
Additive model OR (vs AA)	1	1.52	2.31	5.00	1	0.03
Dominant G allele OR (vs AA)	1	1.17	1.17	0.27	1	0.60
Dominant T allele OR (vs GG)	1	1.00	2.43	10.99	1	9.00×10^{-4}
Allele table, G vs. A allele	1 (A)	1.52 (G)		4.99	1	0.03

Table 1.1 Genotype vs. outcome for tests of association between SNP genotype and trait (adapted from Lunetta., 2008)

1.2.17 Testing the association of variants with trait

Uncertainty owing to imputation can be accounted for by weighting the probability of imputed genotypes: Score tests are an asymptotic test of hypothesis, and have been used to rapidly evaluate the likelihood of probablistic genotypes under the null hypothesis (Marchini and Howie., 2010). The score test has worked well to test association of variants with a binary or quantitative trait when fully genotyped or high certainty, well-imputed variants are used, as the log-likelihood is close to the quadratic function of the regression model. In this scenario the score test is a close approximation of the Cochran-Armitage trend test which exploits the suspected direction of the effect to increase the power to detect association (Sasieni., 1997). P-values are then used to interpret the quantile of the score test statistic. The level of significance which rejects an incorrect null hypothesis to preserve the type I error is dependent on the number of statistical

tests. Conventionally, 5% has been chosen as the significance level for the overall analysis at which the type 1 error is conserved, this is likely to consist of many tests with much lower significant levels (Balding., 2006). SNPTEST v2 is used to implement the aforementioned score test and can also condition on user-specified covariates to detect independent effects (Marchini and Howie., 2010).

1.2.18 Assessing the functional potential of risk-associated variants

The functional potential of putative SLE-associated risk variants should be evaluated to better assess their relevance prior to timely expression studies. Susceptibility markers at a single locus are often constrained by high pair-wise LD so this strategy would refine these variants to better inform the study. The ENCODE (Encylopedia of DNA Elements) pilot project (Birney et al., 2007) established pervasive transcription of the genome: These data comprehensively related transcription start sites (TSS) to their specific regulatory sequences and highlighted regions of accessible chromatin and histone modifications associated with transcription. In addition to these data, evolutionary and computational analyses have generated scores based on mammalian conservation (Siepel et al., 2005) which can also be used to assess the regulatory potential of variants.

The overlap of susceptibility variants with regulatory elements can be evaluated using NCBI (URL: http://www.ncbi.nlm.nih.gov/), UCSC (http://genome.ucsc.edu/) and Ensembl (http://www.ensembl.org/index.html). These databases incorporate many sources of regulatory data, including that from the ENCODE project. The regulatory potential and conservation scores for detecting cis-regulatory modules can be generated for aligned mammalian genome sequences (King et al., 2005).

1.2.19 Examining causal variants for regulatory potential: motif inference

Once a causal variant has been identified, the encompassing DNA sequence can be examined for interaction with regulatory proteins including transcription factors (TFs) to predict binding to regulatory proteins with high confidence. For a causal polymorphism, each allele can be investigated for its impact on binding affinity of the motif for the target protein: SELEX binding data and position weight matrix (PWM) profiles stored in the Jaspar core database are used to investigate the DNA sequence for degeneracy of the motif (Portales-Casamar et al., 2010). Binding of the regulatory protein can be confirmed using publically available genome wide ChIP-seq data generated in EBV-B-cell lines as part of the ENCODE project (ENCODE Project Consortium., 2010).

Causal variants located in coding regions can be investigated for impact on amino acid substitution and hence structure and function of the gene using the PolyPhen-2 (Polymorphism Phenotyping v2) tool. Orthologs and paralogs of the gene sequence are used to increase the accuracy of the predicted effect in the multiple sequence alignment (MSA). After identification and alignment of homologs, the putative coding variants are interrogated for their predicted functional impact with respect to the translated protein. Polyphen-2 replaces amino acids where the variant causes a non-synonymous change and a naive Bayes classifier is used in two datasets (HumDiv and HumVar) to predict and classify the functional impact of each coding variant (Adzhubei et al., 2010).

1.3 Genetic susceptibility to SLE: a complex trait

1.3.1 Genetic basis to SLE

The genetic basis to lupus is established as a key element in disease susceptibility: Increased heritability, familial aggregation (as illustrated by high sibling recurrence risk ratios) and increased concordance for monozygotic compared to dizygotic twins denote lupus as a complex genetic trait (Deapen et al., 1992; Jarvinen and Aho, 1994; Alarcon-Segovia et al., 2005). Multiple etiologic genes determine SLE susceptibility and the number of established genetic associations has increased sharply as a result of advances in genotyping methodologies: Variations in 50 to 80 loci with modest effect sizes are thought to explain the genetic component to disease (Rhodes and Vyse, 2008), although the current number of established loci stands at around 30. No particular gene is necessary or sufficient for disease expression, however, major histocompatibility complex (MHC) genes confer the greatest risk with modest contributions from multiple non-MHC genes (Vyse and Kotzin, 1998). The mechanisms by which these genes predispose to SLE are not completely understood; an established locus where this is the case is the Fc receptor locus (Fanciulli et al., 2007). However, the known SLE loci can be broadly categorised by signalling pathway (Figure 1.3) (Harley et al., 2009).

In lupus, whole genome linkage scans have identified several large regions containing many genes with immune-related function. The *MHC* locus and the classical complement gene *C1q* (Bowness et al., 1994) were amongst the first polymorphic loci related to the human disease, the latter being a rare but highly penetrant contributor to risk**.** Linkage analysis in SLE has proved unsuccessful in the majority of studies: Underlying complex genetics and clinical heterogeneity have resulted in linkage studies beset with background noise. Early studies had poor resolution (5 to 20cM) due to the analysis of too few meiotic events in rare forms of lupus (Botstein and Risch, 2003). The initial linkage of the *MHC* risk locus on chromosome 6 has later been re-visited in the same pedigrees to map polymorphic variants with increased density and narrow the linked interval to class II (Tsao, 2004). Linkage studies in lupus have been inadequately powered to pin-point non-MHC loci in SLE as they contribute modestly to disease risk.

1.3.2 Modelling SLE risk loci

Highly penetrant monogenic forms of disease are rare; the genetic risk in most cases of SLE is derived from multiple variants at mostly independent susceptibility loci (Ramos et al., 2011; Harley et al., 2009). Most loci have modest effect on pathogenesis, in common with other complex autoimmune and inflammatory traits (Barrett et al., 2008). To date, replicated loci explain 10-20% of the genetic component to SLE (Harley et al., 2009). Modelling the biological processes underlying mechanism is uncertain in SLE as most causal processes are not completely delineated. For known risk loci, SLE is highly heritable but with large variance in risk between individuals, the genetic model which best fits the majority of established risk loci is the additive model (Madsen et al., 2011; Slatkin, 2008). The overall risk is the sum of the contributions from each locus.

1.3.3 Rare genetic forms of SLE

Although SLE is usually modelled with several genetic loci under the additive model, rare more penetrant forms exist. These include complete deficiency of the early classical complement pathway component C1q (Botto et al., 1998), mutations in *DNASE1* (Yasutomo et al., 2001) and *TREX1* (Lee-Kirsch et al., 2007). Rare forms of SLE have clarified disease pathogenesis, and the 'rare variant, rare disease' paradigm applies to the aforementioned forms of disease. The extent to which SLE heritability is explained by rare disease remains broadly undefined.

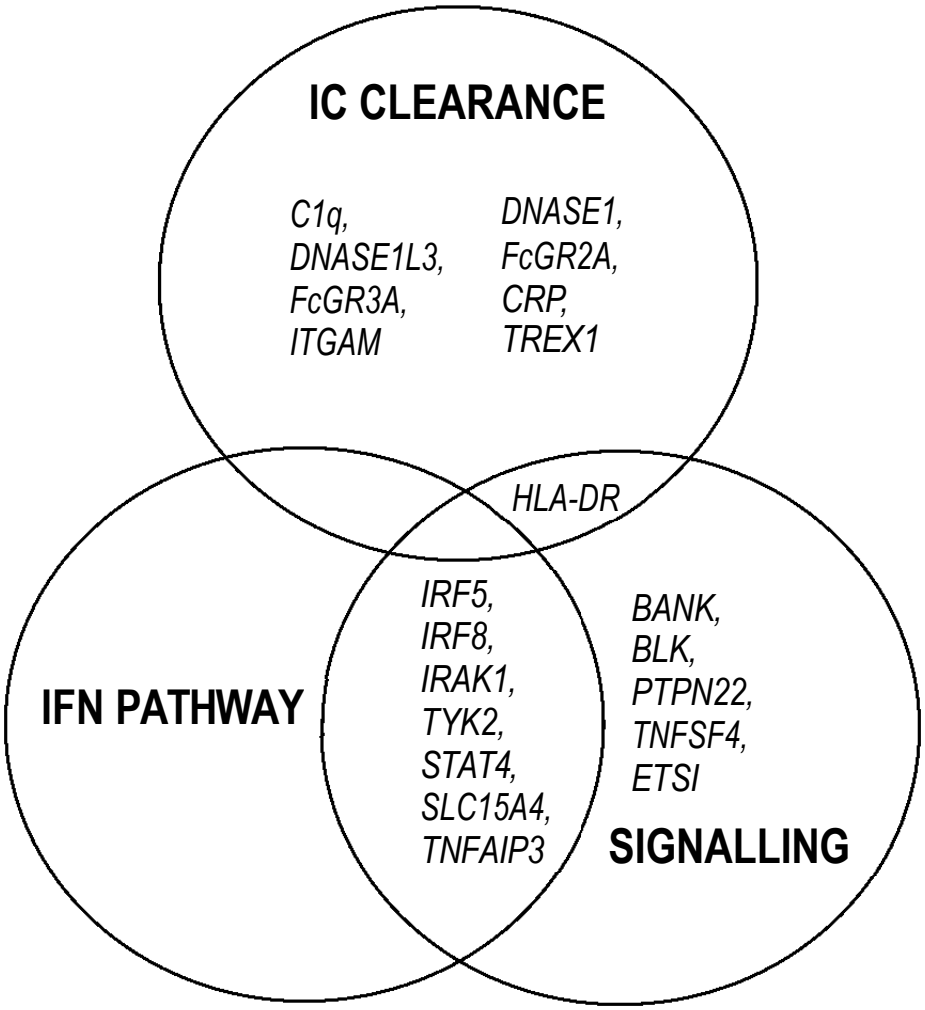
1.3.4 Combined linkage and linkage disequilibrium-based mapping in SLE

Early successes using the combined method in lupus were limited; the higher frequency polymorphisms which typically predispose to autoimmune disease are also common in healthy populations. A key SLE susceptibility gene identified using combined linkage and LD is *IRF5* (Sigurdsson et al., 2005). The genetic association of *IRF5* with SLE is now validated by candidate gene association mapping (Graham et al., 2006) and hypothesis-free genome-wide association study (GWAS) (Hom et al., 2008). I address the rationale behind these aforementioned tools for mapping genetic signals in disease in the following sections.

1.3.5 Cross-population candidate gene association studies in lupus

The number of loci correlated with SLE has been limited by the aforementioned constraints owing to heterogeneity in complex disease. The discovery of SNPs, distributed at high frequency across the genome, has greatly accelerated identification of SLE susceptibility loci. Most of the early signals identified in SLE by association testing are now established lupus risk loci. Associated variants which fail to replicate are a much reduced feature of modern genetic research in SLE, due in part to the accumulation of large, well-matched SLE-

Figure 1.3 Pathways that contain established SLE susceptibility loci. Updated and adapted from a figure in Harley *et al*. 2009



control cohorts. Failed associations are explained by poor study design or over-interpretation of data due to failed detection of LD. Inflated associations might also result from inconsistencies in allele frequency due to population stratification.

Several high-throughput genome-wide association studies (GWAS) have been undertaken in lupus (Harley et al., 2008; Hom et al., 2008). GWAS has corroborated the association of immunologically relevant loci (Rhodes and Vyse, 2008) previously identified by linkage and candidate gene association mapping studies (Table 1.2). These large-scale candidate gene association and GWAS studies have identified over 30 robust SLE-associated loci (Hom et al., 2008; Harley et al., 2008; Yang et al., 2010; Han et al., 2009). Most disease-associated genes are involved in immune regulation or signalling pathways (Figure 1.3), and have informed, to an extent, on the pathogenesis mechanisms which underpin lupus. More recently, robust lupus-associated genes discovered in European cohorts have been tested for association across ancestral groups as described in this section.

1.3.5.1 MHC class II

The strongest genetic contributors to SLE risk reside within the human MHC human leukocyte antigen (HLA) region on chromosome 6p. The HLA super-locus spans over 3Mb and is the location of many candidate genes with immune function. Association data from Europeans suggest common extended HLA haplotypes are associated with SLE: The HLA class II alleles involved in antigen presentation are included on these extended associated haplotypes. The major association in Europeans is ascribed to HLA-DR3 (DRB1*03:01). Within class II, HLA-DR2 (DRB1*15:01) and HLA-DR3 (DRB1*03:01), are independently associated with at least a two-fold increased risk of lupus (Yang et al., 2010).

Assessing the precise contribution of individual MHC genes to lupus susceptibility is challenging, owing to the long-range LD exhibited by alleles at

the locus. The HLA association of DRB1*03:01 with European SLE is the most consistently replicated signal by candidate gene and GWAS association study and one of the few loci identified by linkage (Graham et al., 2002; Fernando et al., 2008) which has consistently replicated by association study. The association of the MHC with SLE is strongly corroborated by signals in East Asian populations and large-scale studies in African-Americans, Hispanics and South Asian SLE-control cohorts are currently underway. Data from a multi-ethnic cohort suggests distinct alleles at HLA-DR3 best explain the aforementioned association of MHC class II in each population, indicating heterogeneity in these different populations: The European-associated HLA-DR3 (DRB1*0301) haplotype is not associated in Hispanics or African-Americans (AA). Conversely; HLA-DR3 (DRB1*15:03) is best-associated in AA and HLA-DR3 (DRB1*08:01) in Hispanics (Mihas et al., 1981; Fernando et al., 2007; Barcellos et al., 2009; Yang et al., 2010; Morris et al., 2012).

1.3.5.2 IRF5

Type I interferons (IFNs) are a class of cytokine with pleiotropic functions with regards to immune cell activity. IRF5 is a transcription factor which promotes inflammatory macrophage polarization and T(H)1-T(H)17 immune responses (Krausgruber et al., 2011). Elevation of IFNs and related proteins in the sera of lupus individuals has been reported as early as 1983 (Preble et al., 1983). As a result, they have been evaluated as contributors to disease development and progression prior to the discovery of the relevant genetic association (Ronnblom and Alm, 2002). A joint linkage and association scan by Sigurdsson and colleagues has identified interferon regulatory factor 5 (*IRF5*) to be associated with European SLE (Sigurdsson et al., 2005). Independent cis-acting variants which tag a common *IRF5* haplotype drive differential expression of distinct *IRF5* splice variants. These data suggest the rs2004640-T allele to create a 5' donor splice site in an alternate exon 1 of *IRF5*, allowing expression of several unique *IRF5* isoforms (Graham et al., 2006). *IRF5* splice variants are associated with increased expression of *IRF5* transcript to elevate risk of disease (Graham et al., 2006; Graham et al., 2007).

Association of the *IRF5* gene with lupus is robust across genetic studies in East Asian, Mexican and African-American SLE-control populations: the intronic risk-associated allele rs2004640-T is best-associated in European and African-American SLE-control groups (Kelly et al., 2008). The observed association of rs2004640-T is strengthened in independent cohorts of Amerindian families: the frequency of the rs2004640-T-T homozygote genotype is higher in Amerindian cases compared to Europeans and African-Americans (Reddy et al., 2007). Not all signals at *IRF5* in SLE are preserved across populations: a high frequency *IRF5* haplotype which confers protection against SLE in Northern Europeans is absent or observed at very low frequency in AAs (Graham et al., 2007).

Table 1.2 Best evidence of association, SLE risk loci by population

Locus associated with SLE	Population	Best evidence of association Marker, Odds ratio (95% CI)/ Nominal p-value	Case/control frequencies	Study type	Pathway	Reference
BANK1	AA	rs548234, 0.78(0.7-0.88)/ $5.9x10^{-5}$	1724, 2024	Candidate gene association	Immune regulation/ signalling	Sanchez et al.
	European	rs10516487, 1.38(1.25-1.53)/ $3.7x10^{-10}$	2003, 1968	Candidate gene association		Kozyrev et al.
	East Asian	rs4522865, 0.73/ $3.56x10^{-4}$	3620, 5700	GWAS and replication		Yang et al.
BLK	Han Chinese	rs7812879, 0.69(0.64-0.74)/ 2.09×10^{-24}	4199, 8255	GWAS and replication	Immune regulation/ signalling	Han et al.
	European	rs13277113, 1.39 (1.28-1.51)/ 1×10^{-10}	2104, 4197	GWAS and replication		Hom et al.
	Japanese	rs13277113, 2.44(1.43-4.16)/ 4.75×10^{-7}	327, 322	Candidate gene association		Ito et al.
	AA	rs13277113, 1.36(1.19-1.55)/ 6.4×10^{-6}	1724, 2024	Candidate gene association		Sanchez et al.
DNASE1L3	Middle Eastern Arabian	C643delT, LOD score of 6.6	N/A	Autozygome linkage analysis	Defective clearance of DNA	Al Mayouf et al.
ETS1	Han Chinese	rs1128334, 1.37(1.29-1.45)/ $1.77x10^{-25}$	4199, 8255	GWAS and replication	Immune regulation/signalling	Han et al.
FCGR2A	European	rs1801274, 0.74(0.65-0.83)/ $6.78x10^{-7}$	3137, 6456	GWAS and replication	Immune regulation	Harley et al.
FCGR2B	European	rs1050501, 2.06/ 0.014	326, 1296	Candidate gene association,	Defective clearance of DNA	Willcocks et al.
	H/K Asian	rs1050501, 1.7/$8x10^{-5}$	819, 1026	Candidate gene association,		Willcocks et al.
FCGR2B and 3B	Thai Asian	Na2/Na2 and Thr232 OR=2.55	187, 87	Candidate gene association	Defective clearance of DNA	Siriboonrit et al.
FCGR3B	European	CNV (risk if 0 or 1 copy) 2.23/ $2.7x10^{-8}$	536, 312	Candidate gene study, qPCR	Defective clearance of DNA	Fanciulli et al.
	Japanese	Allelic (Na2/Na2)/ $8x10^{-3}$	81, 217	Candidate gene study, qPCR		Hatta et al.
	European	Allelic (Na2/Na2)/ 0.014	365 trios	Paralog ratio test		Morris et al.
HLA-DRB1*	AA	1501	N/A	Family-based TDT	Immune regulation	Fernando et al.
	European	0301, 2.3/ $4x10^{-8}$	365 trios			
	Hispanic	0801	N/A			
	East Asian	rs9271100, 1.9(1.59-2.27)/ $1.42x10^{-12}$	N/A			
IKZF1	European	rs2366293, 1.23/ $2.33x10^{-9}$	8710, 5510	Candidate gene association	Immune regulation	Cunninghame Graham et al.
	East Asian/	rs10276619, 0.77(0.73-0.82)/ $1.19x10^{-16}$	4199, 8255	GWAS and replication		Han et al.
IL10	European	rs3024501, 1.19(1.11-1.28)/ $4x10^{-8}$	1963, 4329	Custom targeted chip and	Immune regulation	Gateva et al.
IRAK1	European	rs763737, 1.19/ $1.05x10^{-3}$	3123, 3114	Candidate gene association	Immune regulation/signalling	Jacob et al.
	East Asian	rs763737, 1.41/ $2.29x10^{-8}$	945, 869	Candidate gene association		Jacob et al.
	Hispanic	rs763737, 1.68/ $6.45x10^{-3}$	845, 265	Candidate gene association		Jacob et al.
IRF5	European	rs2004640, 1.47(1.36-1.60)/ $4.4x10^{-16}$	1661, 2508	Candidate gene association	Immune regulation/ signalling	Graham et al.
	East Asian	rs2070197, 1.43(1.32-1.54)/ $8.14x10^{-19}$	4199, 8255	GWAS and replication		Han et al.
	Amerindian	rs2070197, 2.06(1.63-2.6)/ $1.65x10^{-9}$	804, 667	Candidate gene association		Sanchez et al.

Locus associated with SLE	Population	Best evidence of association Marker, Odds ratio (95% CI)/ Nominal p-value	Case/control frequencies	Study type	Pathway	Reference
ITGAM	Colombian	rs1143679, 2.53(1.75-3.68), $3.6x10^{-7}$	4199, 8255	Candidate gene replication	Immune regulation	Han et al.
	European	rs9888739, 1.62(1.47-1.78), $1.61x10^{-23}$	3137, 6456	GWAS and replication	Immune regulation	Harley et al.
	Hispanic American	rs1143679, 2.06(1.44-2.97)/ $8.74x10^{-5}$	657, 227	Candidate gene association	Immune regulation	Molineros et al.
	European	rs1143679, 1.78/ $1.7x10^{-17}$	3818, unclear	GWAS and replication	Immune regulation	Nath et al.
	Amerindian	rs1143679, 2.23(1.77-2.82), $6.22x10^{-11}$	804, 667	Candidate gene replication	Immune regulation	Sanchez et al.
JAZF1	European	rs849142, 1.19(1.13-1.26)/ $1.54x10^{-9}$	1963, 4329	Custom targeted chip and	Immune regulation	Gateva et al.
LYN	European	rs7829819, 0.77(0.7-0.84)/ $5.4x10^{-9}$	3137, 6456	GWAS and replication	Immune regulation	Harley et al.
PHRF1	European	rs4963128, 0.78(0.71-0.85)/ $1.3x10^{-7}$	3137, 6456	GWAS and replication	Immune regulation	Harley et al.
PRDM-	European	PRDM1, rs6568431, 1.20(1.14-1.27)	1963, 4329	Custom targeted chip and	Autophagy	Gateva et al.
	Han Chinese	rs548234, 1.25(1.2-1.3)/ $5.18 x10^{-12}$	4199, 8255	GWAS and replication		Han et al.
PTPN22	European	rs2476601, 1.35(1.24-1.47)/ $3.4 x 10^{-12}$	1963, 4329	Custom targeted chip and	Immune regulation	Gateva et al.
PXK	European	rs6445975, 1.27(1.15-1.39)/ $9.2 x 10^{-7}$	3137, 6456	GWAS and replication	Immune regulation/ signalli ng	Harley et al.
RASGRP3	Han Chinese	rs13385731, 0.7(0.64-0.76)/ $1.25 x 10^{-15}$	4199, 8255	GWAS and replication	Immune regulation/ signalli ng	Han et al.
SCUBE1	European	rs2071725, 0.78(0.72-0.86)/ $1.21 x 10^{-7}$	3137, 6456	GWAS and replication	Defective clearance of DNA	Harley et al.
SLC15A4	Han Chinese	rs1385374, 1.26(1.18-1.35)/ $1.77x10^{-21}$	4199, 8255	Custom targeted chip and	Unclear	Han et al.
STAT4	European	rs7574864, 1.57(1.49-1.69)/ 1.4x10-^{41}	1963, 4329	Custom targeted chip and	Immune regulation/ signalli ng	Gateva et al.
	Han Chinese	rs7574864, 1.51(1.43-1.61)/ $5.17x10^{-42}$	4199, 8255	GWAS and replication, meta		Han et al.
	Amerindian	rs7574864, 1.41(1.2-1.66)/ $5.54x10^{-5}$	804, 667	Candidate gene replication		Sanchez et al.
TNFAIP3	Han Chinese	rs2230926, 1.72(1.52-1.94)	4199, 8255	GWAS and replication	Immune regulation	Han et al.
	European	rs5029939, 2.29/ 2.89 x 10^{-12}	2104, 4197	GWAS and replication	Immune regulation	Hom et al.
TNFSF4	Han Chinese	rs2205960, 1.46(1.4- 1.6)/ $2.53x10^{-32}$	4199, 8255	GWAS and replication	Immune regulation	Han et al.
	African-American	rs2205960, 1.49(1.2-1.8)/ $3.79x10^{-5}$	1680, 2170	Candidate gene replication		
	Hispanic	rs2205960, 1.62(1.4-1.9)/ $3.79x10^{-12}$	1348, 717	Candidate gene replication		
	European	rs2205960, 1.34(1.3-1.4)/ $4.6x10^{-15}$	3432, 3640	Candidate gene replication		
TNIP1	European	rs7708392, 1.27(1.1-1.35)/ $3.8x10^{-13}$	1963, 4329	Custom targeted chip and	Immune regulation	Gateva et al.
	Han Chinese	rs10036748, 0.81(0.75-0.87)/ $1.67x10^{-9}$	4199, 8255	GWAS and replication	Immune regulation	Han et al.
UBE2L3	European	rs5754217, 1.22(1.14-1.32)/ $7.53x10^{-8}$	3137, 6456	GWAS and replication	Ubiquitination	Harley et al.
UHRF1BP1	European	rs11755393, 1.17 (1.1-1.24)? $2.2x10^{-8}$	1963, 4329	Custom targeted chip and	Unclear	Gateva et al.
	Hong Kong	rs13205210, 1.49(1.3-1.7)/ $2.8x10^{-9}$	1230, 3144	GWAS and replication meta		Zhang et al.
WDFY4	Han Chinese	rs1913517,1.24(1.17-1.32)/ $7.22x10^{-12}$	4199, 8255	GWAS and replication	Unclear	Han et al.
XKR6	European	rs6985109, 1.23(1.16-1.3)/ $2.51x 10^{-11}$	3137, 6456	GWAS and replication	Immune regulation	Harley et al.

1.3.6 African-American and Hispanic chromosomes in SLE

Differences in complex disease prevalence across ancestral populations is explained by the environment and are also due to genetic drift and positive selection of unique sets of variants per continent (Pfaff et al., 2001). Some of these variants favour predisposition to the disease more than others, although only a small fraction of genetic variation represents the differences between populations. Admixed populations including African-Americans (AAs) and Hispanics are disproportionately burdened by SLE and other complex diseases.

1.3.7 Imputation-based association analysis in lupus

Imputation-based strategies have been used in lupus association studies to increase power: re-visiting the established Integrin-α-M (*ITGAM*) association with SLE, combining genotypes from multiple studies have allowed investigators to delineate causal variation at *ITGAM* (Nath et al., 2008; Han et al., 2009). The authors have used imputation to assess the *ITGAM* association across independent cohorts of UK-Europeans, Columbians and Mexicans by the aforementioned meta-analysis methods. Their trans-ancestral approach has identified and replicated the association of the exonic (*ITGAM*) variant in SLE individuals of European and African-American descent: the research has additionally identified an independent signal to explain association of *ITGAM* in an Asian population (Han et al., 2009).

1.3.8 Epigenetic modifications and SLE

Recent research suggests that additional heritable factors over and above the aforementioned genetic loci influence SLE susceptibility: these can be epigenetic changes which are defined as functionally relevant heritable modifications caused by elements other than the DNA sequence. These changes modulate gene expression in SLE: the most extensively studied epigenetic modifications which

influence gene activity are DNA methylation and histone post-translational modifications (Ballestar, 2011).

1.3.9 Functional Assessment of SLE-risk loci in human populations

Examination of the genomic sequence at the locus better directs functional experiments linked to the association of candidate susceptibility genes in disease: Functional studies of the *TNFAIP3* gene in SLE have been refined by *in silico* assessment of multiple risk-associated variants prior to expression studies using *ex vivo* samples (Adrianto et al., 2011). Examining the sequence of the aforementioned lupus susceptibility gene *IRF5* has identified a candidate variant (rs2004640-T) with potential to regulate expression of *IRF5* transcription. Rs2004640-T is located 2bp downstream of the intron-exon border of an alternative first exon; the associated T allele creates a consensus GT donor splice site, thus providing a regulatory mechanism. Quantitative real-time PCR analysis of total RNA from human PBMCs has confirmed genetic association of this variant with SLE: The investigators have found increased expression of *IRF5* in cases compared to controls (Graham et al., 2006).

1.3.10 Functional Assessment of SLE-risk loci: specific cell populations

In common with other complex diseases, evidence from mostly *in vitro* studies have suggested the involvement of multiple immune cell types in lupus pathogenesis: APCs, including activated B lymphocytes, and CD4+ T-cell sub-sets, are likely causal contributors. Experimental assessment of the cellular subtypes which best expresses the associated gene is a requisite.

1.3.11 Murine models of SLE

Many of the loci implicated in SLE do not directly alter the coding sequence but alter regulation of the associated gene, usually in a cell-specific manner: Identifying the cell types that are pathogenic in lupus as a result of these risk loci is challenging in an experimental setting, but necessary for accurate assessment of

functional impact. However, results obtained from *in vitro* cellular model systems are often difficult to translate *in vivo.* There has been less evidence that directly implicates a mechanism of auto-reactive B-cell activation in human SLE than comes from mouse models, which provide a direct cellular link to disease (Lee et al., 2002).

Multiple, independent sources suggest that pathogenic B and T-cells are causal contributors to risk of lupus in mouse models of the trait. This contrasts to the largely circumstantial evidence for the role of these cells in human disease. Autoimmune MLR mice homozygous for the *lpr* mutation lack B-cells and have attenuated disease (Shlomchik et al., 1994). Research also suggests CD8+ T-cells are activated in lupus-prone MRL-Fas^{lpr} mice, though not directly by B-cells (Chan and Shlomchik, 2000) and inhibiting CD4+ cell –dependent B-cell help in SLE-prone NZB/NZW F1 mice results in absence of the immune response (Mihara et al., 2000). The double-stranded DNA autoantibodies found in most patients with disease (Monestier and Kotzin, 1992) which contribute to disease pathogenesis are also found in spontaneous mouse models of disease (Frese and Diamond, 2011). The immune systems of the mouse and human are similar in lineage and structure but also have fundamental differences (Hu et al., 2011). Murine models have directed functional studies using human cells with limited success, such as the association of the Fc receptor locus in SLE (Fanciulli et al., 2007).

.

1.4 Next generation sequencing in complex disease

Determining the sequence of a complete set of chromosomes allows study of global genetic properties of an individual, organism or related species and this is core to the discipline of genomics. The first sequenced genome, that of the bacteriophage *phiX*174 virus, was completed three decades ago (Sanger et al., 1977) and determining the sequence of bases of DNA has remained at the heart of genomics since. Automated Sanger sequencing has dominated the landscape of genome sequencing for several decades: This method is still used for small-scale projects and to confirm the results of alternative sequencing technologies, it is considered a first-generation technology. The Sanger approach has relied on the use of uniquely fluorescent chain-terminating nucleotide analogues for each genetic base followed by capillary electrophoresis to determine nucleotide order.

The most successful application of the Sanger method has been the complete sequencing of the first human genome (Lander et al., 2001; Venter et al., 2001) at a cost of roughly $3.3 billion over approximately 14 years. Over a decade later, human genomes are now rapidly sequenced due to technologies collectively termed Next-Generation Sequencing (NGS). Collectively, these platforms offer the major advantage of producing vast amounts of sequencing data at low production cost (Metzker, 2010). These rapid *de novo* sequencing methods are greatly accelerating biological and medical research and discovery: Whole genome sequencing studies in humans, unimaginable a few years ago, are performed routinely today.

Multiple NGS platforms are routinely used today, each with subtle advantages to address differing biological questions. The main technologies which co-exist are Roche-454 (Margulies et al., 2005) which relies on pyrosequencing; the Illumina-Solexa platform uses sequencing by synthesis of immobilised templates, and Applied Biosystem's SOLiD (Sequencing by Oligonucleotide Ligation and Detection) (Bentley., 2006; URL: http://www.appliedbiosystems.com/absite/us/en/home.html). All methods require a robust sequencing template; this is a crucial determinant of the quality of sequencing reads.

Sequencing metrics differ for each platform, though quality and accuracy scores are generated per run for each method. For all platforms, template preparation is followed by sequencing, genomic alignment and assembly of reads. The read lengths generated by NGS platforms tend to be shorter compared to Sanger reads; in NGS reads are generated with higher coverage depths to compensate for sequence length, a requisite for accurate assembly. The Roche-454 GS FLX Titanium platform generates long read-lengths, at approx. 450bp average length per read; sequencing of 400-600 million bases per 454 run takes around 10 hours. This method is expensive and used less than other technologies.

1.4.1 NGS sequencing platform: Roche- 454

The Roche 454 method is the first of the established NGS platforms that gained prominence by describing the first million base pairs of the Neanderthal genome (Green et al., 2006). The advantage of this platform over competing methods has been the longer sequencing read lengths which has allowed improved mapping in repetitive regions, themethod also has relatively rapid turnaround times. However, the high reagent costs of this method and error rates in homopolymer regions have contributed to the competing Illumina Solexa platform being the

most widely used platform in the field. The overall error rate during a 454 run is less than that for a Solexa run (Gilles et al., 2011).

1.4.1.1 Roche 454 and read amplification

The 454 approach uses fragmentation of DNA into randomly sized pieces to build a template library. Immobilisation of DNA fragments to a support surface facilitates parallel sequencing of many hundreds of thousands of reactions simultaneously. To detect these templates requires amplification and the method used in 454 is emulsion PCR (emPCR); universal primers are ligated to sites common to all fragments followed by amplification. DNA is separated under conditions favouring a single molecule per bead and amplified. After enrichment, the emPCR beads are deposited into individual pico-sized wells of a specially-adapted plate (pico-titre plate) and sequenced.

1.4.1.2 Roche 454: read assembly and frameworks for identifying novel variants

After read generation comes assembly; the tools used to map Sanger sequences use an overlap-consensus-layout paradigm to align overlapping identical ends of adjacent sequencing reads. Where an accurate reference sequence exists, these tools are well-equipped to map the sequence. However, the reference still has coverage gaps and ambiguities which can make alignment challenging, these gaps often arise from sequencing errors, conserved interspersed repeat elements and copy-number variants.

The huge volumes of NGS data generated have posed unparalleled data handling questions. These have accelerated the development of storage and analysis frameworks. The latter examine data without bias, so that true variants are better distinguished from aberrant calls. The analysis framework has to map reads to an often imperfect reference genome. The local genome is realigned around complex

variations such as insertion/deletion (indel) polymorphisms and CNV. Machine artefacts vary with sequencing platform, and there are at least five different platforms, so multiple biases are factored into a single per-base error estimate.

NGS data can suffer high per-base error rates in addition to errors of alignment. To quantify the error rate associated with a SNP or indel accurately, the sequence must be discriminated well. This is usually accomplished within an analysis framework by probabilistic algorithms. Within the framework, base calling and alignment errors are modelled with priors of established local variants and fitted against LD patterns for the genomic region. The base-calling algorithm Phred is often used to estimate the probability of error for each base-call as a function of the parameters computed from the data (Ewing and Green, 1998). Phred places an emphasis on discrimination within the high quality range (error rates <0.01) for the data, the error probability is then log transformed so that rates closest to 0 can also be resolved. A base-call having a probability of 1/1000 of being incorrect is assigned a value of 30, 1/10,000 assigned a rate of 40, and so on.

Error rate determination at the base level is followed by local refinement of the aligned sequence. To do this, the reference file is converted into a technology-independent file format such as SAM/BAM. A final stage of analysis involves accurate variant calling in the presence of covariates which reflect features specific to the local genome; some plasticity is therefore required to refine true variants from false positive calls (DePristo et al., 2011). Large-scale and multiplexed sequencing studies need standardized formatting for storing sequence variations from NGS. The variant call format (VCF) incorporates meta-information tailored to the specific data into a standardised format (Danecek et al., 2011). VCF was developed to represent human genetic variation against a single reference sequence for the 1000 Genomes project but has been adopted for large-scale studies including the NHLBI exome project (URL: http://evs.gs.washington.edu/EVS/).

The large number of genome-scale sequenced datasets which vary with considerable diversity require scalable, intuitive visualisation tools. Evaluation of identified variants is followed by visualisation of reads at the nucleotide scale against a reference sequence. The integrative genomics viewer (IGV), developed at the Broad Institute, functions to do so on a standard desktop computer. IGV has low computational load and supports the integration of aligned sequence reads, mutations and CNVs across multiple individuals or projects. The IGV enables exploration of the dataset at a range of scales with relative ease (Robinson et al., 2011).

1.4.2 A Pilot whole-genome study in humans: 1000 Genomes CEU and YRI trios

The pilot phase of the 1000 Genomes Project has sequenced DNA from LCLs from human HapMap CEU and YRI trio individuals. (1000 Genomes Project Consortium, 2010; Conrad et al., 2011). These data have identified increased paternal *de novo* mutations (DNMs), with non-overlapping ranges, and rate variation within and between the two families sequenced. Although the number of somatic DMRs that have been identified in the CEU trio far exceeded those identified in the YRI trio, the number of germ line DNMs have illustrated the opposite trend: A three-fold increase in the inherited variants have been identified in the YRI pedigree. An almost equal number of the inherited DNMs are located in introns and intergenic regions in the YRI trio. As expected, a significantly reduced number of variants have been discovered in coding regions (Conrad et al., 2011). A striking feature of the 1000 Genomes pilot study is the discovery of a 20 fold increase in the number of somatic DNMs relative to historical studies. The investigators have proposed the observed difference owing to the age, mutagenic culture and/or clonality of the cell lines. Age-related metrics were unavailable to the authors and have not been factored into the published data. The size of the study and use of cell-lines has limited the observations to humans.

Re-visiting the 1000 Genomes pilot study, whole genome x2-4 coverage data from HapMap phase II European (CEU), West African (YRI) and East Asian (JPT and CHB) groups has been combined with high coverage exome sequencing data from HapMap phase III samples. This has leveraged the strengths of each study to catalogue novel genomic variants (Gravel et al., 2011). These data have demonstrated capture of low-coverage intergenic common variation, though many rare variants in the non-coding genome have been lost. The allele frequencies for the diploid population have been difficult to assess with the low coverage data: The investigators have found a majority of human genome-variable sites to be rare with low sharing amongst diverged populations.

1.4.3 Exome sequencing studies

Selective sequencing of genomic coding regions is a cheaper alternative to whole genome studies since less than 5% of the capacity is required. This approach uses targeted exome-capture followed by short-read NGS sequencing at high coverage to enrich for the discovery of highly penetrant variants. The strategy has resolved the genetic basis of rare Mendelian disease (Biesecker, 2010) with small numbers of unrelated, affected individuals (Ng et al., 2009). Using large sample sizes may extend the strategy to complex disease: Control data-sets from exomes of European-American and African-American individuals, from large well-phenotyped cohorts, are available in the NHLBI ESP Exome Variant Server (URL: http://evs.gs.washington.edu/EVS/). This set of exomes might be used to extend and enrich the discovery of novel loci and mechanisms in SLE.

1.4.4 Trans-ancestral mapping and sequencing in lupus

Genetic association by re-sequencing has successfully refined the *TNFAIP3* association in lupus. The *TNFAIP3* gene encodes A20, an ubiquitin-modifying enzyme which regulates NF-κB. A20 modifies RIP and TRAF6 downstream of the TNFα or Toll-like receptor (Graham et al., 2008). GWAS in European and Asian SLE-control cohorts has identified *TNFAIP3* variants which are strongly

associated with risk of disease (Graham et al., 2008; Han et al., 2009; Yang et al., 2010). Re-visiting these association data, logistic regression has been used to model the association of typed variants and proxies identified that are in LD with untyped polymorphisms on the risk haplotype: Further fine-mapping and genomic re-sequencing in European and Korean lupus cases have fully characterized this haplotype. Nine *TNFAIP3* risk chromosomes from seven carriers of European ancestry (two homozygotes and five heterozygotes) have been sequenced; no additional informative SNPs have been identified at the locus, instead the investigators have found a novel single base deletion on all nine risk chromosomes. The final analysis identified this as a TT>A polymorphic dinucleotide (deletion T followed by a T to A transversion) strongly associated with SLE in both Europeans and Koreans (Adrianto et al., 2011).

1.5 Tumour necrosis factor (ligand) superfamily, member 4

1.5.1 TNFSF/TNFRSF superfamily

The tumour necrosis factor (TNF) and TNF receptor super families (TNFSF and TNFRSF) consist of approximately 50 membrane and soluble proteins that can modulate diverse aspects of immune function (Croft. 2012). These molecules mostly evolved with, or closely after, the adaptive immune system 350–450 million years ago (Croft. 2012). The control of immunity, which includes cell-survival, occurs upon TNFSF engagement of complementary TNFSFR (Croft, 2010). TNFSF-TNFRSF interactions strongly regulate conventional CD4 and CD8 T-cells: Although the specificity of the T-cell response is controlled by the TCR, complete activation is only achieved after interaction between accessory co-stimulatory receptor-ligand pairs.

Primary co-stimulation occurs on interaction between CD28 and B7, however at a later stage when a sustained or memory response is required, there is a secondary co-stimulatory event, and the interacting pair often belongs to the TNFSF-TNFRSF superfamily. This interaction directly influences adaptive immunity through T-cell signalling. TNFSF-TNFRSF members are linked via a series of membrane proximal events to NF-κB and stress kinase signalling, resulting in cytokine production and cellular proliferation (Watts, 2005). Binding also influences the innate immune response indirectly through activation of antigen presenting cells (APC).

1.5.1.1 TNFSF/TNFRSF pairs involved in activation and survival

TNFSF/TNFRSF pairs involved in activation and survival, that prevent excessive apoptosis of T-cells, are TNFSF4/TNFRSF4, 4-1BBL/4-1BB, CD30L/CD30, LIGHT/HVEM, CD70/CD27, and GITRL/GITR (So et al., 2006). Within each pair, the TNFSF ligands are type II membrane glycoproteins. Most are homotrimers with unique and overlapping function and shared intracellular signalling. The TNFRSF members also have a similarly shared structural aspect in that they are all type 1 membrane glycoproteins. Other TNFSF/TNFRSF pairs can induce pro-apoptotic pathways, illustrating their role as double-edged swords in the immune response (Aggarwal. 2003).

1.5.2 TNFSF-TNFRSF in autoimmunity

TNFSF and TNFRSF family members can exacerbate or ameliorate disease depending on the prevailing circumstances (Watts, 2005). Evidence suggests they are key mediators of organ-specific autoimmune disorders including inflammatory bowel disease and rheumatoid arthritis: Blocking agents against specific family members used in patients with these conditions show beneficial results in the majority. TNFSF/TNFRSF are implicated in systemic autoimmunity: TNFRSF6 has long been implicated in SLE - the soluble form appears to increase in the serum of patients with active disease (Courtney et al., 1999). There is also evidence for association of *TNFRSF6* with SLE: A novel nucleotide insertion in *TNFRSF6* mRNA alters the reading frame, causing cells to have mRNA editing refractory to apoptosis, thereby providing a mechanism for defective clearance of auto-reactive immune cells (Wu et al., 2011).

The TNFS member, TNFSF13b (BAFF), is found at significantly higher levels in SLE patients (Roschke et al., 2002). This molecule regulates B-cell differentiation, including follicular B-cell development, and the conversion of memory cells to antibody-producing plasma cells (Marston and Looney, 2010). In two moderately successful clinical trials, blockade of TNFSF13b by the humanized antibody Belimumab has proved efficacious against moderately active

SLE, with disease prevention in early or preclinical cases. Blocking the TNFSF13b molecule probably censors auto-reactive B-cells prior to disease progression. In 2011 Belimumab was granted FSA approval, the most recently approved drug specifically for treatment of lupus in 50 years.

1.5.3 TNFSF4

1.5.3.1 TNFSF4

Another TNFS member, *TNFSF4* (also known as OX40L, CD252), is located on chromosome 1q25, within a genetic interval replicated to show significant linkage with SLE, making it a plausible candidate susceptibility gene (Shen and Tsao, 2004). TNFSF4 forms a functional trimer which uniquely binds its receptor, monomeric TNFRSF4 (OX40, CD134), on T and NK lineage cells (Compaan and Hymowitz, 2006) to provide a late-stage co-stimulatory signal at the APC- T-cell interface. In common with related molecules, expression of the TNFSF4-TNFRSF4 pair is not ubiquitous for the most part, but can be induced following activation at the surface of a wide-range of cells which control immune functionality.

1.5.3.2 TNFSF4 expression

The TNFSF4 homotrimer is induced on antigen-presenting cells (DCs, B-cells and macrophages) by innate (TLR) and adaptive (BCR, Ig) signals prior and during engagement with naive and memory T-cells; expression of the ligand is transient. Aberrant expression of the resulting effector T-cells are pathogenic in autoimmunity and protective in infection and cancer (Croft, 2010). TNFSF4 can be induced, under physiologic conditions, on $CD4^+$ and $CD8^+$ T-cells, suggesting T-cell-T-cell interactions could ramp up inflammation. There is also evidence for constitutive expression of the ligand by lymphoid tissue inducer cells (LTi), an immune accessory cell type which interacts with B-cells at the B-T junction in secondary lymphoid organs (Kim et al., 2003). Regulatory T-cells subsets are able to express the complementary receptor, TNFRSF4. Engagement of this molecule by TNFSF4 may inhibit proliferation of regulatory cells or prevent their

suppressive effects to further promote effector T-cell responses (Ito et al., 2006). In common with other TNF ligands, TNFSF4 can also be induced on non-immune structural cells, including vascular endothelial and smooth muscle cells, which are proximal to activated T-cells; interactions may contribute to tissue damage at sites of inflammation.

1.5.3.3 TNFSF4 *and human disease*

Several lines of evidence published over the last 15 years suggest signalling specifically through TNFSF4–TNFRSF4 in humans is required for the induction of adaptive anti-tumour immunity, allergy and autoimmunity (Gri et al., 2008; Cunninghame Graham et al., 2008; Seshasayee et al., 2007; Zaini et al., 2007). The TNFSF4 homotrimer inhibits generation of natural and adaptive T regulatory (TR1) cells (Ito et al., 2006). There is evidence to suggest that the extent of the inhibitory effect, and predominance of T-helper subtype, influences progression to disease: Receptor-ligand signalling is protective against tumours in malignant Hodgkin's Lymphoma (Buglio et al., 2011). The Thymic stromal lymphopoietin (TSLP) molecule is an important mediator of many allergic diseases (Kaur and Brightling, 2012). TSLP increases the expression of TNFSF4 on immature dendritic cells (DC), which respond by promoting T helper (Th) 2 polarization of naive T-cells within the lymph node. These polarised T-cells then produce the cytokines typically implicated in the allergic response. The TNFSF4 molecule influences the severity of allergic phenotypes: Airway Smooth Muscle increases expression of TNFSF4 in asthmatic individuals (Krimmer et al., 2009).

In the aforementioned traits involving TNFSF4, T-lymphocyte activation modulates the severity of the response and/or influences progression to disease. T-cell activation is not a single event and requires a primary signal (antigen recognition from the APC-TCR interaction), followed by a secondary costimulation event (TNFSF4 ligation of TNFRSF4). Signal two strongly influences the protein kinase B (PKB)-signalling pathway to augment TCR-

dependent activation of NF-κB (So et al., 2011). In the absence of signal 2, T-cells do not expand efficiently in response to the antigen and the long-lasting memory response is impaired.

1.5.4 Mouse models of human pathologies: Autoimmunity and TNFSF4

The outcome of the TNFSF4-TNFRSF4 interaction is not limited to human disease: Knockout mouse models have impaired accumulation of antigen-specific T-cells, reduced cytokine production and inflammation (Croft, 2010). Blockade of the TNFSF4-TNFRS4 interaction also has ameliorative effects in animal models of T-cell pathologies (Compaan and Hymowitz, 2006) including allergic and autoimmune manifestations (Nohara et al., 2001). Weinberg and colleagues illustrate this well: Depletion of TNFRSF4+ effector T lymphocytes at inflammatory sites by anti-TNFRSF4 improved autoimmune sequelae in an EAE model of MS (Weinberg et al., 1996). The signalling complex induced by the TNFSF4-TNFRSF4 interaction includes the TNFR-associated factor 2 (TRAF2). $Traf2^{(-/-)}$ knockout mice develop fatal autoimmunity characterised by autoantibody production and organ infiltration by T-cell subsets including activated, effector, and memory cells (Lin et al., 2011).

1.5.5 Genomic organisation at 1q25.1

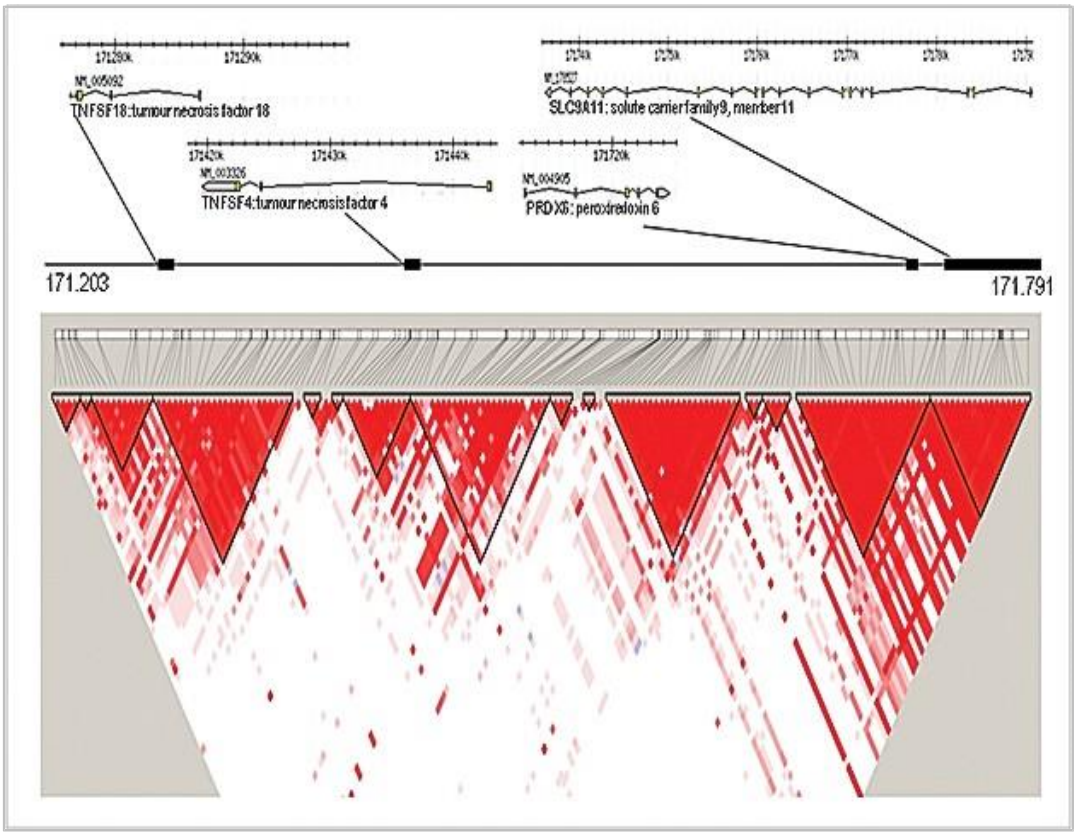
TNFSF4 maps to human chromosome 1q25.1, where the proximal adjacent gene *TNFSF18* (GITRL) has overlapping functional effects and is likely to have arisen by ancestral gene duplication. The gene adjacent to *TNFSF4* distally, *PRDX1*, bears no structural or functional resemblance to *TNFSF4*. Figure 1.4 demonstrates *TNFSF4* in relation to neighbouring translated genes on chromosome 1q25.1. Figure 1.4 also demonstrates that genetically, *TNFSF4* is situated within a bipartite structure of two linked haplotype blocks which have minimal long range LD with neighbouring genes. The *TNFSF4* gene spans 23.57kb: it has 3 translated exons, 2 introns and a 3kb 3′-untranslated region with many known variations (Figure 1.5). Figure 1.6 demonstrates the three transcripts of the human *TNFSF4*

gene which are translated into protein. The high level of sequence conservation of TNFSF4 protein between humans and primates, as depicted in Figure 1.7, is expected. Conservation of DNA sequence between humans and the other Eutheria is unexpectedly low for non-translated regions; this includes the 3′UTR of *TNFSF4*, a region rich in common polymorphisms with the potential to regulate expression of the *TNFSF4* gene (Figure 1.5).

1.5.6 TNFSF4- genetic variation and gene expression studies

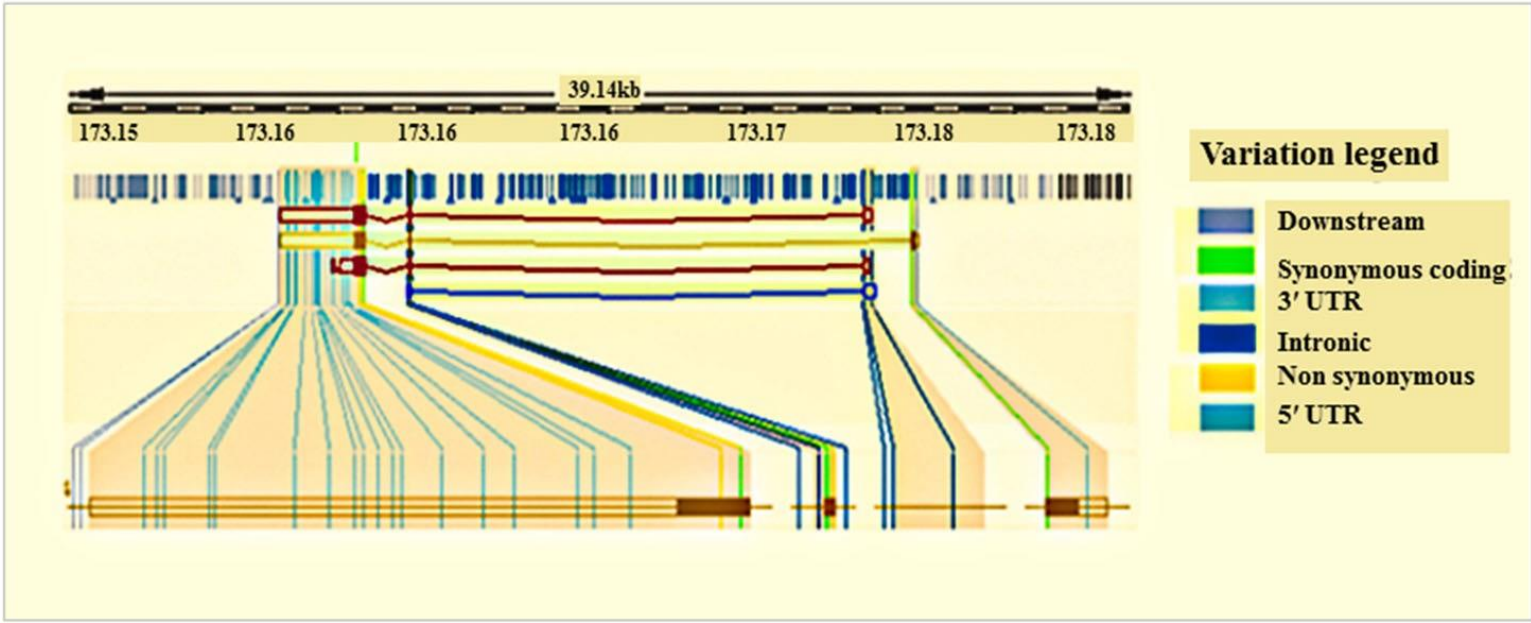
The first report of *TNFSF4* polymorphisms relating to gene expression appeared in 2005. QTL mapping in mice susceptible to atherosclerosis, a complex inflammatory disorder, revealed six point deletions and 2 SNPs in the proximal promoter region of *Tnfsf4*. These polymorphisms segregate with heart and aortic expression of Tnfsf4 mRNA 3.7 and 4.5 times higher, respectively, compared to controls. Genotyping studies of polymorphisms across the homologous *TNFSF4* region in humans were undertaken in Northern European atherosclerosis and myocardial Infarction (MI) - control cohorts. In both populations, the genotype of rs3850641 is associated with an increased risk of MI in females but not in males. The high degree of LD between the associated variant and adjacent genotyped SNPs translated into a risk haplotype associated in both study cohorts (Wang et al., 2005).

Figure 1.4 *TNFSF4 and neighbouring genes in a 588kb genetic interval on chromosome 1q25.1*



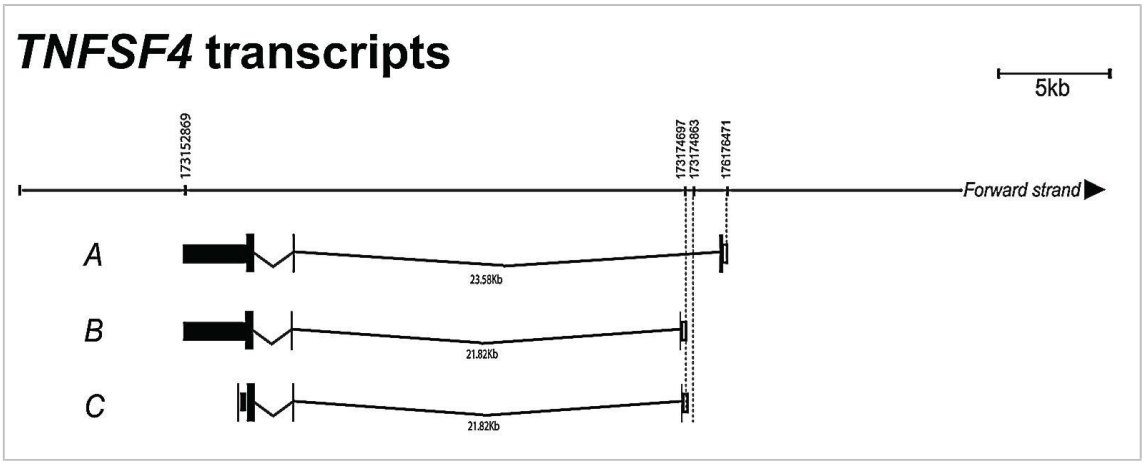
A plot depicting pair-wise LD relationships between SNP markers genotyped across 588kb of chromosome 1q25.1 in CEU Northern and Western European samples from HapMap phase III (data release phase 3/#3 May 2010, NCBI B36 Assembly).The upper section of this figure is annotated for the genes in this interval in relation to the LD plot below. The plot was generated in Haploview 4.2 using a standard algorithm for haplotype calling. The black triangles depict haplotype blocks and grey ticks, SNP location to scale. The multi-allelic correlation coefficient D' is a measure of linkage disequilibrium that ranges from 0 to 1, D' = 1 (deep red) indicates complete linkage disequilibrium (no evidence for historical recombination between blocks), and D' = 0 (white) represents zero LD (absent correlation between haplotype blocks).

Figure 1.5 Diagram illustrating the known variations in the *TNFSF4* gene



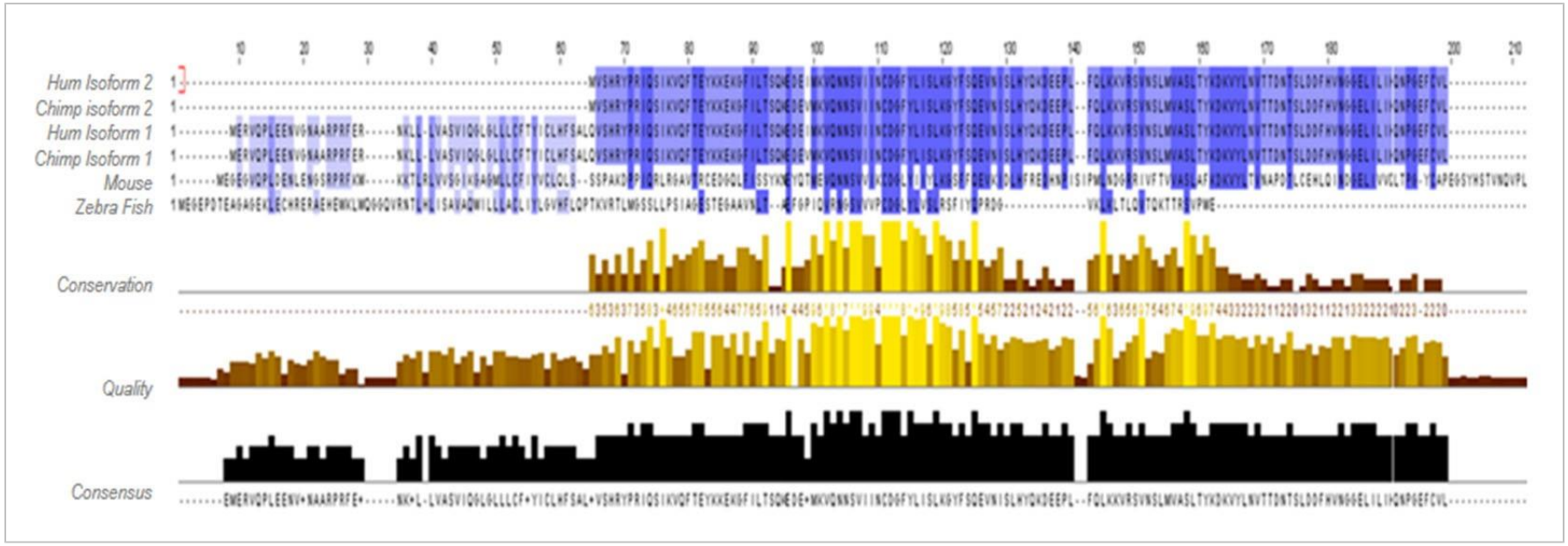
The image depicts known variants in the *TNFSF4* gene displayed as vertical lines against the four known transcripts of the gene. The coloured vertical lines in the top section of the diagram are coloured for their position in the gene, and adjacent 5′ and 3′ UTR regions. The variations in the bottom section are located in the *TNFSF4* coding sequence and are summarised by colour-coding for their effect on the translated protein. The synonymous (green) and non-synonymous (yellow) categories are informed by the Condel tool, used to provide a consensus prediction based on the SIFT and PolyPhen prediction scores. This variation diagram is generated in Ensembl version 65.

Figure 1.6 Translated splice-forms of human *TNFSF4*



The three translated spliceoforms (A-C) of the human *TNFSF4* gene are 3470, 3429 and 1240bp, respectively. Form A has been determined the most abundant form by 5´RACE-PCR and is translated into a protein of 188 amino-acid residues. Splice variants B and C are translated into an identical 133 residue protein. A fourth splice variant (not shown) is processed but does not have a known translated protein form.

Figure 1.7 Cross-mammalian conservation of TNFSF4



Multiple alignments of the TNFSF4 protein sequence for selected Eutherian mammals from Ensembl Release 65 were interrogated using the ClustalW2 tool (EBI-EMBL). The amino acids are numbered 1-210 and areas of similarity which may be associated with specific features that have been more highly conserved than other regions are coloured dark blue with scores of conservation, quality and consensus aligned below each amino acid.

1.6 Work which has led to these doctoral studies

At the time these doctoral studies were planned, the Vyse group had started to interrogate the association of variants spanning both *TNFRSF4* and *TNFSF4* in human SLE. These genes were chosen as candidates for a number of reasons: Chromosome 1q23-25 had been flagged in several linkage screens (Johanneson et al., 1999; Edberg et al., 2002) in lupus families, almost certainly because of the multiple immune-related genetic association signals across the interval. This flagged *TNFSF4* as a plausible candidate susceptibility gene. A parametric linkage screen of European and admixed Amerindian multi-case SLE families had also unambiguously confirmed genetic linkage of 1p36 (the interval containing *TNFRSF4*) and 1q25 (*TNFSF4*) with lupus (Johanneson et al., 2002). In addition to these early data, expression of TNFSF4 on a range of cells directly implicated in the adaptive immune response highlighted it as a candidate lupus susceptibility gene.

Using a candidate gene association study format, haplotype-tagging SNPs, already typed in CEU individuals as part of HapMap phase II, were selected for genotyping. These variants spanned a 220kb section of 1q25.1 encompassing the *TNFSF4* gene, 3′ UTR, 5′UTR and 5′ upstream region, up to the adjacent recombination hotspot. We found evidence that variants in the 5′ *TNFSF4* region were strongly associated with SLE in families and a SLE case-control cohort, both groups of northern European descent. Results presented in chapter 3 of this thesis suggest correlation of the associated risk haplotype with cell-surface expression (Cunninghame Graham et al., 2008) in this early study of *TNFSF4* in SLE. Using these same strategies, we illustrated absence of association of 1p36, and hence *TNFRSF4*, with SLE. Studying the recombination across the locus, we showed that the association arose from a 100kb haplotype ($P<1 \times 10^{-5}$, after permutation) in UK-Europeans and European-Americans.

The *TNFSF4* haplotype associated with risk of disease (*TNFSF4*$_{risk}$) is found at a frequency of 20% in European populations and is tagged by rs1234317-T, rs2205960-T, rs12039904-T and rs10912580-T. In the early aformentioned data from the Vyse group, conditioning on the contribution from each haplotype-tagging allele did not resolve the association signal. The most frequent haplotype at this locus was under-transmitted to European SLE families and under-represented in European cases. This haplotype was tagged by a single allele, rs844644-A. The increased association of *TNFSF4* 5′ risk alleles with disease has been replicated by GWAS in European and East Asian populations (Han et al., 2009; Yang et al., 2010), highlighting the genetic similarities at this locus in these ancestrally distinct populations.

A major obstacle in the identification of disease-specific causal variants at *TNFSF4* in the European and East Asian SLE cohorts has been the strong linkage disequilibrium (r^2>0.8) exhibited by genotyped *TNFSF4* alleles. This has resulted in a high frequency extended haplotype associated with risk of disease instead of delineating causal variations at the locus (Cunninghame Graham et al., 2008). It is probable that migration out of Africa involved many founder effects and bottlenecks to increase haplotype length in East Asian and European populations (Foster and Sharp, 2004). As illustrated earlier in this introduction, Hispanic and African-American populations are disproportionately affected by SLE (Molina et al., 1997) and health disparities in these groups show onset at a younger age (Fernandez et al., 2007).

1.7 Summary and study aims

The importance of non-MHC loci in genetic susceptibility to lupus is firmly established (Vyse et al., 1998; Wandstrat and Wakeland, 2001). Multiple studies of SLE families have found strong linkage with a genetic interval on chromosome 1q25, a region that harbours multiple genes involved in immune regulation. Using both a family-based and case-control study design, the Vyse group have shown association of the *TNFSF4* gene with risk of European SLE. A major obstacle in the definition of causal variation at this locus is the strong linkage disequilibrium (r^2>0.8) exhibited by genotyped *TNFSF4* alleles, which has resulted in a high frequency extended haplotype associated with risk of disease instead of delineating causal variations at the locus (Cunninghame Graham et al., 2008). I will use complementary and related strategies to explore the mechanism of disease preposition at the molecular and cellular level. I will attempt to define the molecular genetic basis of the disease association.

The specific questions that will be addressed in this thesis are as follows:

1. Does polymorphism at *TNFSF4* predispose to SLE in non-European populations?
2. Does the haplotype structure in non-European populations offer the potential to reduce the size of the *TNFSF4* region associated with SLE?
3. How does polymorphism at *TNFSF4* influence gene expression?
4. Are there additional novel variants which predispose to disease?

.

Study aims:

1. Use SNPs to fine map the *TNFSF4* locus in a large cohort of Europeans to achieve greater power to resolve the association with SLE.

2. Replicate the genotyping study undertaken in Europeans in aim 1. in five non-European populations, to define global *TNFSF4* association with lupus.

3. Evaluate *TNFSF4* for its utility in terms of dissecting disease pathogenesis: The functional relevance of identified *TNFSF4* risk-variants will be assessed for correlation with cell-surface expression.

4. Perform a targeting deep-sequencing study on the genomic region encompassing *TNFSF4* in SLE cases selected for their *TNFSF4* genotype: To identify additional polymorphisms, if any, for association mapping in SLE.

5. Use the novel variants from 4. to comprehensively define the variants unique to the $TNFSF4_{risk}$ and $TNFSF4_{non\text{-}risk}$ haplotypes for comprehensive haplotype construction.

6. Define the full spectrum of variants underlying the upstream *TNFSF4* association in lupus.

7. Define rare SLE-associated coding variants to inform future functional experiments to investigate pathogenic mechanism.

Chapter 2

Materials and methods

2.1 Baseline characteristics of study cohorts

The trans-ancestral mapping study presented in chapter 4 included over 17,900 SLE and control individuals of self-reported European, African-American (AA), Gullah, East Asian, Hispanic and Amerindian ancestry. All cases fulfilled four or more of the 1997 ACR revised criteria (Tan et al., 1982) for the classification of SLE and provided appropriate written informed consent. The cohorts are described in this section. As expected, SLE cases were predominantly women (82.33%).

Data presented for *TNFSF4* expression analysis additionally used samples from the BDA-Warren Repository held at the JDRF/WT DIL (Cambridge, UK) (Bain et al., 1990) and are described as cohort 7 in this section.

2.1.1 White European cases and controls (cohort 1)

The white European SLE cohort consisted of 3009 pooled samples of European SLE cases of European American and mainland European origin held under the Oklahoma Medical Research Foundation (OMRF) Institutional Review Board (IRB), together with 910 samples from the UK European SLE cohort held at Imperial College at the time the study was completed. After QC analyses, the final cohort used for the analyses is presented in Table 4.1 in chapter 4 of this thesis and consisted of 3432 cases and 3640 controls.

OMRF and USC Lupus Genetics Study cohort- white European SLE cases

White European SLE cases from America and mainland Europe were enrolled in the Lupus Genetics Study at the OMRF and additional collaborating European Institutions under the OMRF IRB in collaboration with Professor John B. Harley and the SLE genetics consortium (SLEGEN). Diagnosis was verified for all affected individuals through extensive medical record review and patient interview. Where possible self-reported ancestry was obtained on the basis of grandparental country of origin, otherwise parental ancestry was used. The cohort

included SLE individuals and probands from SLE pedigrees. Genomic DNA was extracted using standard methods from anti-coagulated blood samples and/or buccal swabs and/or mouthwash samples.

UK European SLE cases

This cohort represents a growing collection of SLE cases recruited through United Kingdom rheumatology clinics in London or by direct patient contact following publicity. The cohort is now held under the King's College London IRB after being held for 7 years under the Imperial College London IRB. Ethical approval was obtained under MREC/98/2/6 and all patients recorded their demographic variables by patient questionnaire. Clinical data on SLE manifestations in all subjects were obtained from medical record review.

Blood samples were collected from each participant and genomic DNA was isolated using a standard protocol for phenol-chloroform extraction from 40ml blood by technical staff and stored at 4°C. Clinical data and biological samples were collected at a single time-point at study enrolment. Disease phenotype and activity varied and samples were taken from individuals with acute disease in outpatients. 288 SLE probands from white UK European parent/proband trios recruited as above were included in addition to 622 individual cases.

Ex-paternity individuals identified through current and previous genetic mapping studies were excluded from the analyses and only individuals of self-reported White European ancestry included (Rhodes, B., 2008). This cohort was integrated with the larger cohort of individuals of white European ancestry (below) before genotyping.

European controls

The DNA samples from a total of 3491 genotyped European controls were provided to the organizing centre (OMRF) from eight separate investigators/centres. Of the total, 547 controls were contributed by the OMRF.

At the time of experiment, the controls were population-based and did not have SLE, no family history of SLE, and no other autoimmune illness. There is uncertainty whether controls provided by the other centres met all these criteria.

2.1.2 African-American cases and controls (cohort 2)

The African-American cohorts consisted of cases and controls from three main sources provided in collaboration with Prof. Robert P Kimberly and Dr Jeff Edberg from the University of Alabama (CASSLE and PROFILE cohorts). In addition, AA individuals were enrolled in the Lupus Genetics Cohort under the OMRF IRB in collaboration with Professor John B. Harley and the SLEGEN consortium. Collectively the cohort used for analyses comprised 1529 cases and 3577 controls after QC and details are presented in Table 4.1 of chapter 4.

PROFILE cohort

African-American patients from the multi-centre, multi-ethnic PROFILE cohort provided informed consent explicitly indicating their agreement to enrol for longitudinal data collection in this study. Detailed characteristics of recruitment and demographic variables have been published (Alarcon et al., 2002). Individuals were included if 16 years of age or older with disease duration less than 10 years from diagnosis to enrolment. African-American ancestry was defined by reporting all four grandparents to be of the same background. Genomic DNA was extracted from blood obtained after physical examination. Longitudinal data allowed for the examination of phenotypes and damage as per the SLICCC Damage Index (SDI), including time to, and nature of, renal involvement and disease activity.

CASSLE cohort

652 SLE cases and 926 age-matched African-Americans from the University of Alabama CASSLE cohort were included if 16 years of age or older with disease duration less than 10 years from diagnosis to enrolment. African-American

ancestry was defined by reporting all four grandparents to be of the same background. Additional SLE samples from the CASSLE are included in the multi-ethnic PROFILE cohort (above). Genomic DNA was extracted by standard procedures from whole blood and stored.

OMRF Lupus Genetics Study cohort- Africa American SLE cases

As for white European SLE cases enrolled under the same program described in section 2.1.1 of this chapter.

2.1.3 Gullah cases and controls (cohort 3)

As discussed in chapter 1, despite the high disease load in African-Americans, there is the perception that lupus is relatively rare in continental Africans, although AA samples can be used to investigate this; the significant genetic admixture (10-30%) which exists in AAs may confound attempts. A second informative group is the Gullah population of the Sea Islands of South Carolina, the Gullah have lower genetic admixture (<10%) due to geographical isolation and strong cultural heritage. Anthropologic studies indicate a direct ancestral link between the Gullah and Sierra Leonians. The SLEIGH (SLE in Gullah Health) cohort of 152 cases and 122 controls was provided in collaboration with Professor Gary Gilkeson of the Medical University of South Carolina.

Characteristics of recruitment and demographic variables of this cohort have been published (Kamen et al., 2008; Gilkeson et al., 2011). Inclusion criteria were age 2 years or more, self-identification as African-American Gullah from the Sea Islands region of South Carolina with all known ancestors to be Gullah, to meet 4/11 ACR criteria, the ability to understand English and be able to provide informed consent. Healthy AA-Gullah subjects recruited as controls were required to have no family history of autoimmune disease or known family members with SLE. Control subjects had to complete connective tissue screening questionnaires and screening examination for autoantibodies.

2.1.4 Amerindian and Hispanic cases and controls (cohort 4)

Latin Americans have been generically coined Mestizo (Hispanic), in many US studies on the basis of language, but actually constitute a markedly heterogeneous group of subjects with different cultural backgrounds but a common mother tongue, Spanish. Individual contributors to the Mestizo cohort can be found in Table 2.1 and general features of cohorts not already described above are listed in this table. An additional 34 cases and 7 controls from 6 additional contributors which were used are held under the OMRF IRB. The final cohort used after QC analyses are presented in Table 3.1 in chapter 3 of this thesis. The combined cohort of Amerindians and Hispanics consisted of 1348 cases and 717 controls.

Table 2.1 Contributors to the Amerindian and Hispanic cohort

Contributor	Institution	Cases, controls	Cohort	Details
Betty P Tsao	UCLA, California, USA	119, 21	UCLA lupus cohort	
Chaim O Jacob	USC, California, USA	479, 51	Lupus Genetics Study, USC	Jacob et al. 2009
John B Harley	OMRF, Oklahoma, USA	204, 138	Lupus family registry and repository	Sanchez et al, 2011
JM Ananya	Universidad del Rosario, Bogota, Colombia	164, 127	Colombian	Ananya et al. 2011
ME Alarcon Riquelme, Bernardo Pons Estel	Sanitorio Parque, Rosario, Argentina	193, 240	GLADEL	Pons Estel et al. 2011
ME Alarcon Riquelme, Ignacio Garcia de la Torre	University of Guadalajara, Guadalajara, Mexico	101, 64	GENLES	Sanchez et al. 2011
Robert P Kimberly, Jeff Edberg, EE Brown	UAB, Alabama, US	235, 148	PROFILE	Alarcon et al, 2002

GLADEL (Grupo Latinoamericano de Estudio del Lupus) cohort

Clinical, laboratory and prognostic variables were analysed in this Latin American cohort samples from many centres in 9 Latin American countries for which data have been published (Pons-Estel et al., 2004). Each centre incorporated a maximum of 30 randomly selected patients. Disease activity using both SLEDAI (Bombardier et al., 1992) and MEX-SLEDAI (Guzman et al., 1992) was measured in all patients at the time of entry and every 6 months thereafter. All data was collected and held using the ARTHROS 6.0 database which has a lack of language barriers since all elements are coded and allowed English-speaking investigators at the OMRF to collect it.

Lupus family registry and repository Mestizo SLE cases of Mexican ancestry

Mestizo individuals of mainly Mexican ancestry were recruited with the same criteria as for all other samples used in the Lupus genetics Study cohorts described above and held in the OMRF family registry and repository.

Colombian lupus cohort

Colombian patients under approval of the local ethics commitee were recruited at multiple clinics in Medellin, Colombia. All patients fulfilled the minimium (ACR) criteria for the classification of SLE. Controls were unrelated to patients, without inflammatory or autoimmune disease, matched to patients by age (±5 years), sex, and ancestry. Clinical and laboratory variables were evaluated by medical exam, severity and damage measured using the SDI (Correa et al., 2003).

GENLES

This cohort represents a growing collection of SLE cases recruited throughout Latin America known as GENLES. The 101 SLE cases and 64 controls used in the analyses presented in chapter 3 were collected throughout Mexico (specifically

from the cities of Guadalajara, Morelia, Culiacán and Mexico City) (Sanchez et al., 2010).

2.1.5 East Asian SLE-control (cohort 5)

This cohort consists of a heterogeneous group of East Asian subjects with most cases from Korea. Individual contributors and general features of the cohort can be found in Table 2.2. Population-based control samples were supplied by most SLE case contributors. At the time these doctoral studies were planned, the controls had not been diagnosed with SLE or other autoimmune illness and had no family history of SLE illness. The final cohort used for analyses after QC is presented in Table 4.1 in chapter 4 of this thesis.

Table 2.2 Contributors to the East Asian cohort

Contributor	Cases, controls	Origin	Cohort	Details
Susan Boackle	13,0	Asia	Colorado/ Denver	
SC Bae	648, 753	Korea	Hanyang Lupus Cohort	Chun et al. 2005
CO Jacob	90	Asia	Lupus Genetics Study Cohort, USC	Jacob et al. 2009
Judith James	2, 73	Korea	Lupus Genetics Study Cohort, OMRF	Harley et al, 2008
Anne Stevens	13,0	Asia	Lupus Research Institute cohort	Liao et al. 2011
BP Tsao/Nan Shen	571, 522 40, 8 113, 247 23, 8 255, 259 29,0 111,0	Total Asia China Japan Korea Singapore Taiwan	Shanghai Institute of Rheumatology and UCLA. Asian Lupus cohort	Lessard et al. 2011

Hanyang Lupus Cohort

All patients in this growing collection of SLE patients gave informed consent for enrolment in this cohort. All of the patients were Korean in ethnicity. The age, duration of disease and duration of follow-up for a large proportion of patients is

published (Chun and Bae, 2005). In 2005, when these features were published, the patients were aged 36.1+12.1 (mean+SD; range 8–74) years old, their disease duration and follow-up duration were 5.6+3.0 (range 0.4–19.9) and 4.7+2.8 (range 0.4–14.5) years respectively. The female-to-male ratio at the time of this study was 13.6 (434/32).

Shanghai Institute of Rheumatology, Asian Lupus

SLE cases from multiple Asian countries were recruited in this cohort, the Chinese subjects from medical centres in Zhejiang, Shangdong and Liaoning provinces. Written informed consent followed by medical record review confirmed patient eligibility and clinical variables were collected at the time of diagnosis. Features of the disease were recorded by questionnaire. The clinical and immunological features of the SLE patients have been in published (Liao et al., 2011), the controls are area-matched unrelated healthy individuals visiting hospitals.

Phenotypes

Clinical data on SLE manifestations in all subjects used were obtained from medical record review performed at individual institutions, collected and processed at the OMRF, with additional phenotypic information from KCL, MUSC (Gullah) and UAB (PROFILE and CASSLE).

2.1.6 Wellcome Trust Case Control Consortium (WTCCC) Controls (cohort 6)

This collection comprises a common set of nationally-ascertained controls from Great Britain. The samples have two major sources; the 1958 Birth Control Cohort and UK blood donors. This cohort has been used extensively as a common cohort to catalogue human genetic variation in common inflammatory and autoimmune disorders and details of the cohort are described at (URL: https://www.wtccc.org.uk/index.shtml). Fifty individuals

were randomly selected from the 1958 Birth Control Cohort for genotyping of SNPs selected for the European cohort (cohort one) to increase the power of imputation analysis described in section 2.9 of this chapter.

2.1.7 British Diabetic Association (BDA)- Warren Repository (cohort 7)

In 1989, the BDA, in conjunction with a bequest from Alec and Beryl Warren, initiated a major genetic resource of DNA and EBV-transformed cell lines from multiplex type 1 diabetes pedigrees. The efforts of clinical staff, patient groups and publicity aided sample recruitment throughout the UK and Ireland. Contact was made by letter; informed consent was obtained and venesection arranged. To ensure ancestral homogeneity all individuals had four grandparents born within the British Isles. Inclusion criteria included at least one live parent with or without diabetes. Thereby an extensive, growing resource was founded and made available to research groups by material transfer agreement (MTA); at the time these doctoral studies were planned the collection included 3276 samples. Detailed information on Warren collection samples has been published (Bain et al., 1990). Lymphoblasoid cell lines and DNA from this collection was provided in collaboration with Prof. John Todd and colleagues, JDRF/WT DIL laboratory, Cambridge Institute of Medical Research, Cambridge, UK.

2.2 SNP selection

2.2.1 *SNPs selected to resolve the TNFSF4 association with SLE*

Haplotype tag SNPs and proxy variants capturing all common haplotypes were selected for genotyping. This meant we did not type all markers in all groups as marker selection was dictated by *TNFSF4* locus architecture and additional SNPs found to be associated in our European association study (Cunninghame Graham et al., 2008).

At the time these studies were planned, the association of *TNFSF4* SNPs with European lupus was yet to be published; no other pertinent data were in the public domain. Therefore empirical genotypes for HapMap II reference populations dictated *TNFSF4* locus architecture and tag SNP selection.

The selection of SNPs for typing followed a similar pattern for the European, African-American and Gullah and East Asian cohorts, using the northern/western European (CEU), west African (YRI) and East Asian (CHB/JPT) HapMap II panels (Barrett et al., 2005; de Bakker et al., 2006), respectively. The distribution of the correlation (r^2) values between the allelic tests based on the tag SNPs and the untyped variants was assessed using the Tagger facility of Haploview and an r^2 threshold of 0.8. The effective coverage of these tag SNPs in comparison to the entire set of locus-specific polymorphic variants were evaluated and markers which best defined each haplotype block in each cohort were selected. Additional SNPs found to be associated in our European association study, at the time not published, (Cunninghame Graham et al., 2008) were also typed in all cohorts.

Our Mestizo Hispanic and Amerindian samples were collected from regions with decreased African and increased European admixture (Sanchez et al., 2010). These data suggest the largest proportion of source ancestry for these individuals is Amerindian or Southern European, and so they are not specifically represented by the HapMap phase II datasets. It was our view that markers selected for the European cohort spiked with the best tag SNPs selected for each of the other cohorts would best represent the common *TNFSF4* haplotypes associated with SLE in Amerindians.

In all, 125 different SNPs in a 200kb region (chromosome 1q25.1, 171,400,000-171,600,000, NCBI build 36.3) encompassing the *TNFSF4* gene and 5′ region were selected for fine-mapping the *TNFSF4* locus.

2.2.2 SNPs selected to address admixture

A panel of 347 genome-wide SNPs as used by Halder and colleagues (Halder et al., 2008) were used to correct for major ancestry across all cohorts tested. Additionally, 20 1q25-specific ancestry markers were genotyped to correct for local two-way admixture between Europeans and West Africans. These markers were selected for their large differences in allele frequencies between West Africans and Europeans and were not in LD in the ancestral populations.

2.2.3 SNPs selected to discriminate TNFSF4risk and TNFSF4$_{non\text{-}risk}$ haplotypes for gene expression studies

Four SNPs- rs3850641, rs1234314, rs3861953 and rs2205960- were selected to discriminate between *TNFSF4*$_{risk}$ and *TNFSF4*$_{non\text{-}risk}$ homozygote individuals selected from the Warren Collection held at the JDRF/WT DIL laboratories.

2.3 Genotyping

For the trans-ancestral mapping data presented in chapter 4, genotyping was performed in two independent experiments using an Illumina Golden Gate custom genotyping assay for a first round of genotyping of the African-American and European samples followed by fine-mapping of all cohorts on the basis of these data on the Illumina iSelect platform. All genotyping was undertaken at OMRF for the combinations of haplotype tag SNPs and proxy variants described in 2.2.1 with the aim of capturing all common haplotypes at *TNFSF4*. The *TNFSF4* locus was genotyped as part of a larger study of genetic loci in SLE, this significantly reduced cost per genotype.

Technical staff at the JDRF/WT DIL laboratory (CIMR, Cambridge, UK) genotyped the *TNFSF4* variants described in 2.2.3 by Taqman assay (Applied Biosystems, Carlsbad, California, USA). Theoretical aspects of Taqman

genotyping and protocols are described at (URL: http://www.appliedbiosystems.com/absite/us/en/home/applications-technologies/real-time-pcr/taqman-probe-based-gene-expression-analysis/taqman-gene-expression-assay-selection-guide.html).
Genotyping enabled interrogation of the Warren collection for individuals relevant to the experiments planned for which data are found in chapter 3 of this thesis.

2.3.1 Design aspects of custom genotyping assays

Genotyping was carried out in two stages as described, custom SNP assay panels were designed using the Illumina online Assay Design Tool (ADT) (URL: http://www.illumina.com/support/array/array_software/assay_design_tool.ilmn). Input of SNP loci generated an output file which predicted success information including the validation status of each allele by at least two independent methods and the design success in the context of previous successful Illumina genotyping. The output is used to refine the assay panel; three SNPs were replaced on the basis of low ADT score, increasing the likelihood of successful genotyping. The iSelect platform was selected because the historically high call rates give accurate detection of polymorphisms and the platform easily lends itself to high multiplexing for the SNPs selected in section 2.2.

2.3.2 Theoretical aspects of GoldenGate chemistry

Briefly, the workflow to generate genotypes typically starts with allele-specific hybridization followed by DNA extension and ligation. A universal PCR amplification step preceded hybridisation of product to BeadChip and autocalling of genotypes. A more detailed description is found at URL: http://www.illumina.com/technology/goldengate_genotyping_assay.ilmn

2.3.3 Theoretical aspects of iSelect genotyping

Markers were interrogated using 50-mer probes which selectively hybridized to

the selected SNP loci but which stopped one base before the marker. Specificity was conferred by enzymatic single-base extension to incorporate a labelled nucleotide. Dual-colour florescent staining allowed the labelled nucleotides to be detected by Illumina's iScan imaging system, to identify both colour and signal intensity, for which a description can be found at URL: (http://www.illumina.com/technology/infinium_hd_assay.ilmn).

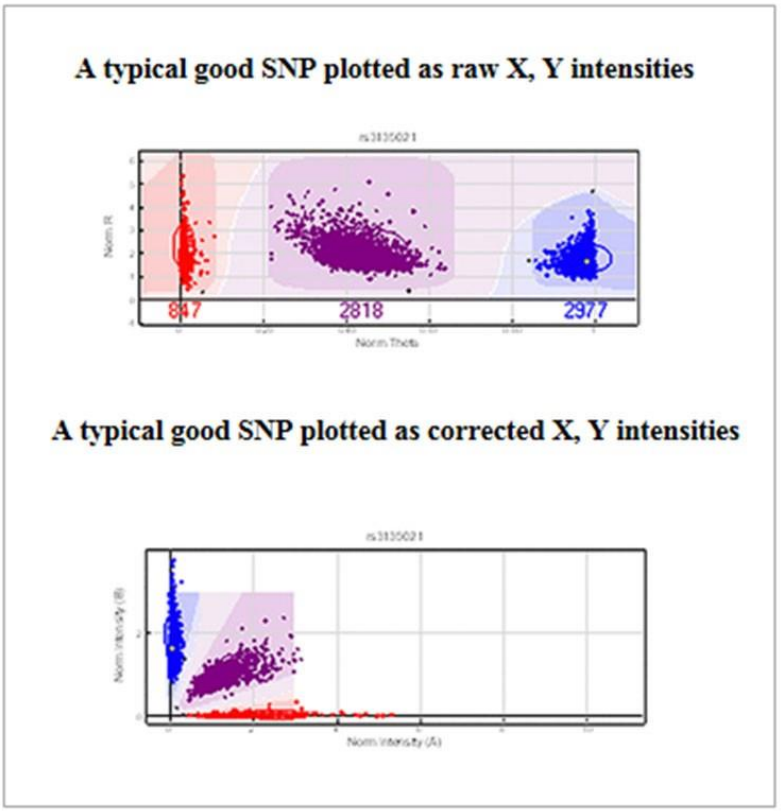
2.3.4 Cluster profiles

SNPs genotypes in their raw form comprised three different plots of the fluorescent intensity and were used to evaluate the quality of the genotype calls. Two fluorescent dyes were used to measure the presence of an allele (X and Y). The allele specific intensities were normalized using a proprietary algorithm in the Illumina Beadstudio software (URL: http://www.illumina.com/Documents/products/datasheets/datasheet_beadstudio.pdf). Normalized allele intensities were transformed to a combined SNP intensity and an allelic intensity ratio. Figure 2.1 illustrates raw and corrected plots for a well genotyped SNP. High quality SNPs should produce three clusters representing the homozygous XX, YY and XY genotypes. Poor separation between clusters, multiple clusters (caused by adjacent polymorphisms) or diffuse clusters resulted in poor accuracy in genotype calls.

2.4 Data storage and management

Genotypes and data on clinical variables and disease manifestations were stored, formatted and indexed in BCSNPmax version 2.5.5 (Biocomputing Platforms Ltd, Espoo, Finland). The BC suite integrated many of the analyses programs used for data presented in chapter 4. Data analysis was undertaken using this system in our local cluster environment.

Figure 2.1 Raw (top) and corrected (below) plots for a well genotyped SNP. Diagram courtesy of Dr Ken Kaufman, OMRF



2.5 Quality control (QC) analyses

2.5.1 QC filtering of individuals

QC of individuals was undertaken before QC of markers to reduce SNP fallout prior to association analyses. Samples with greater than 10% missing genotypes were excluded for poor DNA quality. Individuals with a large proportion of heterozygous genotypes compared to the mean for each cohort were removed as likely contaminated. Individuals were also excluded for low and high heterozygosity. The boundaries for low and high heterozygosity depended on the population, with the highest in African-Americans and lowest in Europeans. Unknown familial relationships due to identity by descent (IBD) were detected between pairs of individuals using the pi-hat approach in PLINK to remove second-degree relatives up to duplicates.

Population stratification bias and effects due to admixture were addressed by genotyping 347 genome-wide SNPs, on the Illumina iSelect platform, as used by Halder and colleagues (Halder et al., 2008). The authors had selected the panel of autosomal AIMs to distinguish individual biogeographical ancestry and admixture proportions for the same four continental ancestral populations used in the study presented in chapter 4 of this thesis. 20 Additional 1q25-specific ancestry markers were interrogated by Illumina GoldenGate custom array to correct for two-way admixture between Europeans and West Africans. Within each cohort the Eigenstrat program, as described in chapter 1, was used for principal components (PC) analysis and global ancestry estimates were additionally inferred by a combined Bayesian and sampling-theory approach (Admixmap). The African-American data was spiked with HapMap phase III West African (Yoruba, YRI), Southern European (Tuscan, TSI) and Northern/Western European (CEU) genotypes to cross-compare two-way admixed AAs with their source populations (chapter 4, Figure 4.2).

2.5.2 QC filtering of SNPs

Following filtering for duplicates, first-degree relatives, assessment of HWE, missingness and major ancestry, the dataset comprised 111 *TNFSF4* SNPs and 294 AIMs and 15600 samples (Details of individual cohorts are found in Table 4.1). Markers with less than 90% genotyping efficiency were excluded from the analysis. The relationship between genotype and allele frequency was evaluated by Hardy-Weinberg Equilibrium (HWE) in control samples of each cohort, to evaluate random mating in the absence of selection. We included markers which deviated up to $P<0.01$ away from HWE as assessed using Pearson's chi-squared test. The null hypothesis, that the distribution of allele frequencies is consistent with established allele frequencies per population, was tested. A single marker, rs1234313 had a HWE value which deviated in East Asian controls ($_{HWE}P=10^{-5}$), rs1234313 was associated with SLE in multiple populations tested in this study, and a previous reported association of this marker with SLE influenced the decision to include it. All other genotyped markers were well within the selected deviation parameters.

2.6 Statistical methods I

2.6.1 Imputation methods

Imputation of the genomic region from 171,385,000 to 171,600,000 (NCBI build 36.3) on chromosome 1q25.1 was performed using IMPUTE2 and combinations of HapMap phase III populations with second reference genotypes dictated by population (described in chapter 4, Table 4.1). Imputation was used to fill missing gaps in the genotyping data and impute markers with MAF>3% missing between datasets to examine structure of common haplotypes across the populations. Imputed SNPs were included in downstream analysis if SNP certainty scores were greater than 0.8 and an Impute info threshold of 0.7 or above. These criteria successfully filtered out all but the best-imputed SNPs. The final datasets comprised Europeans (112 SNPs in 3432 cases, 3640 controls), East Asians (100 SNPs in 1500, 1396), African-Americans (121 SNPs in 1529, 2048) and Hispanics (51 SNPs in 1348, 717, not imputed).

2.6.2 Inference of recombination

FastPHASE v1.2 (Scheet and Stephens 2006) was used to infer missing genotypes and haplotypic phase from unphased *TNFSF4* SNP genotypes from 6272 unrelated control chromosomes (1568 from each population), randomly chosen after QC filtering. FastPHASE incorporates a Hidden Markov Model which allows flexible clustering of SNPs spanning the locus. *Rhomap* from the LDhat2.0 package was used to estimate population scale recombination rates in the presence of hotspots using pre-computed maximum likelihood tables in the analysis (Scheet and Stephens., 2006). Using the approach of Auton and colleagues (Auton and McVean, 2007; Auton., 2007), *rhomap* was run for a total of 1,100,000 iterations including a burn-in of 100,000 iterations, the chain was sampled every 100 iterations after the burn-in. Each simulation incorporated 196 chromosomes meaning a total of eight simulations were completed per group and the mean average recombination calculated between each pair of markers at the *TNFSF4* locus. Simulations were executed in their entirety on three independent

occasions and additional simulations were undertaken in each group by varying the burn-in and chain sampling parameters to ensure there were no irregularities. These analyses were extended to infer recombination in phased chromosomes from African-American *TNFSF4*$_{risk}$ and *TNFSF4*$_{non\text{-}risk}$ homozygote individuals: Figure 4.5 depicts the data in chapter 4 of this thesis.

2.7 Statistical methods II

2.7.1 Single marker association analyses

After QC filtering, single marker association and conditional data were generated using a case-control format and the continuous covariate function in SNPTEST v2 under the additive model (Marchini., 2010). A frequentist statistical paradigm and a probabilistic method was used to treat genotype uncertainty. A logistic regression model which was additive on the log-odd scale was used to evaluate *TNFSF4* variants. Under this model, the score test, an asymptotic test of hypothesis, was used to test association of the variants for the binary phenotype (case, control or phenotype-control), under the null hypothesis. For non-imputed variants and the high certainty imputed SNPs with info>0.7 that were included in analysis post QC, the test statistic reduced to the Cochran-Armitage trend test statistic. The score test was presumed to produce a sensible result since the validity of the quadratic function (of the log likelihood curve) was not undermined by small sample size, low allele frequency or increasing genotype uncertainty. The trend test exploited the suspected effect direction to increase power to detect association.

To preserve the type 1 error, the variance of the score test was adjusted using genomic control to control for inflation. GC was calculated on null loci to estimate variance. Association was computed at each of the null SNPs, and λ calculated as the empirical median, divided by its expectation under the χ^2 distribution (Balding., 2006). The association was then computed for candidate SNPs, where they reached $\lambda > 1$, the test statistics were divided by λ, testing required 2 df. The quantile of the score test statistic was interpreted by calculating

p-values. An arbitrary locus-wide significance level for rejecting the null hypothesis was set at P=5 x10^{-5}. Odds ratios (OR) with 95% confidence intervals (95% CI) were taken from the exponent of the beta coefficient of the logistic regression model together with the standard errors. Significance of association of corrected p-values were based on permutation testing (5000 permutations). Data are represented as nominal uncorrected p-values and permuted (*Pp*) values.

Per SNP significance level α' should satisfy $\alpha = 1 - (1 - \alpha')$, leading to the Bonforroni correction n $\alpha' \approx \alpha$ / n for independent variants tested under a statistical paradigm. However, this correction was judged conservative for the genotyped variants at *TNFSF4*, many of which exhibited high LD (r^2>0.7). Instead, the type-1 error was approximated by a permutation procedure. Case-control status was randomised x5000 whilst maintaining the LD structure of variants for each dataset, to satisfy the null hypothesis, in order to estimate the false-positive rate (Balding., 2006).

Technical details of the aforementioned score test used in SNPTEST v2 are found at URL:http://mathgen.stats.ox.ac.uk/genetics_software/snptest/snptest.v2.pdf

2.7.2 Meta-analysis

A logistic regression model fitted with an interaction term (effect) in the R statistical package was used to investigate cross-study heterogeneity. P-values for individual associated SNPs were generated using the likelihood-ratio test. Rs1234314, rs1234317, rs2205960, rs12039904, and rs10912580 were selected as representatives for this test. I implemented a fixed effects meta-analysis method combining the association results for African-Americans, East Asians, Europeans and Hispanics to more powerfully estimate the true effect size, the results of these analyses are described in Table 4.6, chapter 4. The average effect size across all datasets was computed using inverse variance weighting of each

study. SNPs were organised into two categories (*TNFSF4* gene or 5′ region) and are highly correlated with one another ($r^2 > 0.7$) within each group. Associated SNPs from African-American cohort were tested for heterogeneity, and included in the meta-analysis where the associated allele was the same.

2.7.3 Haplotype bifurcation

The Long Range Haplotype (LRH) test was used to investigate common alleles with long-range linkage disequilibrium (LD): I was able to represent the breakdown of the risk and non-risk haplotypes. $TNFSF4_{\text{risk and}}$ $TNFSF4_{\text{non-risk}}$ were anchored by a core associated marker, rs1234314, in all groups and conveniently positioned at the boundary of the *TNFSF4* gene and 5′ region, also at the boundary of two haplotype blocks. Haplotype bifurcation diagrams were then generated in the program Sweep™.

2.7.4 Haplotype association and conditional regression

Haplotypes in the *TNFSF4* gene and 5′ region were constructed in Haploview 4.2 using a custom algorithm, based on the r^2 measure of linkage disequilibrium (LD). Markers and haplotypes with frequencies greater than 5% and 4% respectively, were included in the analyses. Haplotypes were anchored using tag SNP genotype data and boundaries were inferred using recombination data. SLE case-control association and step-wise conditional logistic regression data for each haplotype was generated in PLINK, as were OR (95% CI) these are represented as nominal uncorrected p-values and x5000 permuted (P_p) p-values.

2.7.5 Sub-phenotype association

Searching for *TNFSF4* alleles linked to specific clinical manifestations of lupus may prove informative with regards to mechanism and so better resolve causal alleles because of greater genetic homogeneity compared to the disease per se:

Phenotypes amenable to study, and relevant to the biologic function of *TNFSF4* in lupus, are described in section 1.1.2 of chapter 1. *TNFSF4* variants were tested for association using the same methods described in section 2.7.1 against interquartile age at diagnosis using a case-only format. The variants were also tested against leukopenia and lymphopenia, anti-La, anti-Ro and anti-Sm autoantibody subsets, which are associated with SLE, together with renal disease, using both case-only and phenotype-control formats. A covariate for the most associated marker per aforementioned phenotype was included for each population to investigate independent effects. The SNPTEST v2 program was used for these tests.

Investigating the correlation of *TNFSF4* genotype with expression

In order to better understand how *TNFSF4* acts as a susceptibility gene in SLE, the upstream

The $TNFSF4_{risk}$ and $TNFSF4_{non\text{-}risk}$ haplotypes were investigated for their influence on *TNFSF4* expression in EBV lymphoblastoid cell lines (LCLs) (provided by the JDRF/WT DIL laboratory, Cambridge, UK) and in peripheral blood mononuclear cells (PBMC) from our UK-European SLE collection described in section 2.1 of this chapter. Haplotypes were defined by genotyping locus-specific SNPs which captured common, informative haplotypes in multiple Northern European parental-proband and SLE-control cohorts (Cunninghame Graham et al., 2008). In excess of 1,000 LCLs were genotyped at the tag SNPs rs2205960, rs1234314 and rs7514229. LCLs were selected if they were risk or non-risk haplotype homozygotes on the basis of this genotyping. Re-sequencing of the risk-haplotype tagging SNPs rs10912580, rs12039904 and rs1234317 and non-risk haplotype-tagging rs844644 confirmed homozygosity of the smaller subset of 12 JDRF/WT DIL samples selected for expression analysis.

2.8 Preparation of genomic DNA

Genomic DNA was isolated from 40ml anti coagulated whole blood by a standard phenol-chloroform extraction. Lymphocytes were separated from anti-coagulated whole blood by centrifugation through Histopaque-1077 (Sigma-Aldrich) using a protocol described on the insert of and using **ACCUSPIN™** tubes (Sigma-Aldrich).

2.9 Selection of samples for expression analysis

Expression analysis was performed on peripheral blood mononuclear cells (PBMC) taken from our UK- European SLE cohort and on EBV-transformed lymphoblastoid cell lines (LCL-cells) from the Warren collection (provided by Prof. John Todd and colleagues, JDRF/WT DIL laboratory). These cells were genotyped using the risk- haplotype-tagging SNPs rs2205960, rs1234314 and rs7514229 by standard Taqman assay (Applied Biosystems, Carlsbad, California, USA). The phase of the upstream haplotypes, for SNPs rs2205960 and rs1234314, was determined using PHASE v2 (Scheet and Stephens, 2006). In order to control for potential variation in expression at the 3′ end of *TNFSF4*, samples were only included in the expression study if they were also homozygous for SNP rs7514229 in the 3′ UTR of the gene. In LCL-cells, samples were selected which were homozygous for tag SNPs carried by the upstream risk haplotype (n=3) and those which were homozygous for the tagging SNPs carried by the under-transmitted upstream haplotype 3 (LCL-under) (n=3).

2.10 *In vitro* activation of PBMCs and LCL-cells

Peripheral blood lymphocytes or LCL-cells were suspended in complete RPMI medium (RPMI 1640 medium (Invitrogen, Paisley, UK) supplemented with 10% Foetal Bovine serum (Invitrogen, Paisley, UK), 10,000 U/ml Penicillin (Invitrogen, Paisley, UK), 10,000 µg/ml Streptomycin (Invitrogen, Paisley, UK) and 200 mM L-Glutamine supplement (Invitrogen, Paisley, UK)) and grown in

suspension at a concentration of $3x10^6$ cells/ml. Cells were stimulated with 10ng/ml rCD40L (Axxora, Nottingham, UK), 20ng/ml CD40L enhancer (Axxora, Nottingham, UK) and 2ug/ml goat anti-human anti-IgD polyclonal antibody (Serotec, Oxford, UK). 2 x 10^6/ml cells were frozen in Trizol® for RNA extraction and $0.5x10^6$ cells/ml re-suspended in FACS staining buffer and then stained with fluor-conjugated antibodies for FACS analysis**.**

2.11 FACS analysis

PBLs or LCL-cells were stained in FACS staining buffer using FITC- conjugated anti-human CD86 mAb (MCA1118F, Serotec, Oxford, UK) as a marker of B-cell activation, in combination with phycoerythrin-conjugated anti-human TNFSF4 mAb (ANC10G1, Axxora, Nottingham, UK).

Cells were size-gated and analysed for expression of CD86 and TNFSF4 and designated negative for *TNFSF4* if expression fell within the background staining compared to a mouse IgG1 negative control mAb (MOPC 31C, Ancell). All analyses of FACS data were carried out on the FACScalibur cell sorter using Cellquest software (Becton Dickinson, Franklin Lakes, USA) and cell plots generated using publically-available WinMDI software (Joe Trotter, Scripps Institute).

2.11.1 Statistical analysis of FACS data

The Mann-Whitney test was used to compare the differences between the numbers of *TNFSF4*-positive cells carrying the $TNFSF4_{risk}$ and $TNFSF4_{non\text{-}risk}$ haplotypes. Further tests were undertaken for the Epstein-Barr virus (EBV)-transformed lymphoblastoid cell lines (LCL cells) and peripheral blood lymphocytes (PBLs).

2.12 NGS sequencing of *TNFSF4* gene and 5′ region

To search for new *TNFSF4* alleles associated with SLE, the Roche-454 Titanium platform (Rothberg and Leamon, 2008), (www.my454.com) was used to sequence a 118kb section of chromosome 1q25.1 encompassing the *TNFSF4* gene and risk-associated 5′ region up to the upstream boundary of this haplotype in 71 individuals. Step-wise conditional regression of the risk-associated alleles at this locus indicate the presence of an independent non-risk signal located in this section of the *TNFSF4* 5′ region tagged by a single non-risk marker, rs844645. Defining the boundary of the association signal in Northern Europeans by mapping genotyped *TNFSF4* and longer range SNPs at adjacent loci allowed selection of individuals who possess two copies of the risk or non-risk SLE-associated *TNFSF4* haplotype. Data from our previous case-control and family-based analyses of *TNFSF4* in UK European SLE are used for haplotype definition (Cunninghame Graham et al., 2008).

2.12.1 Long-range PCR amplification

DNA stored at 2-8°C for selected UK-European SLE individuals was quantified by agarose gel electrophoresis to ensure high molecular weight. PCR primers were designed according to dictates required for long-range amplification using Primer3 (v. 0.4.0) software (Rozen and Skaletsky, 2000). Primer sequences are available on request. Enrichment of the *TNFSF4* locus for NGS was accomplished by long-range PCR in a tile-path which spanned the locus to prevent gaps in sequencing coverage and with amplicons of a size range of 8-14kb. High fidelity PCR was undertaken using the SequalPrep™ Long PCR Kit with enhancer B (Invitrogen, Carlsbad, UK) according to manufacturer's cycling times, adjustments to the extension time at 1min were made for every 1kb of additional sequence amplified.

2.12.2 Purification and pooling of PCR products

High-throughput ultrafiltration-based purification of PCR products >100bp (MinElute 96 UF PCR Purification, Invitrogen) preceded picogreen quantification (Quant-iT™, Invitrogen) to determine molarity. Additional quantification of the products of PCR by agarose gel electrophoresis was undertaken to ensure amplicon fidelity. PCR products were pooled in molar amounts on a per individual basis up to a final concentration of 2ug. These steps preceded hydrodynamic shearing of DNA to a 500-800bp size range by brief pulses of sonication at 4°C (Bioruptor® UCD-200, Diagenone). Further purification by solid-phase reversible immobilisation (SPRI) beads removed sheared DNA <100bp (Dhanya et al., 2008) (Agencourt) and allowed concentration of the volume in EB buffer. All but 400ng of sheared pooled product per sample was stored in liquid nitrogen.

2.12.3 Parallel-tagging and library preparation

The parallel-tagged sequencing (PTS) strategy of Meyer and colleagues (Meyer et al., 2008) was adopted to facilitate processing of all samples in parallel whilst using only half a Roche-454 Titanium chip. The barcoding adaptors comprise single self-hybridised palindromic oligomers 8 nucleotides long; they carry a *Srf*I restriction site in the middle (GCCCGGGC). *Srf*1 cuts approximately every 150kb in mammalian genomes. The sequence tag may start with either an A or a T followed by six freely chosen nucleotides, and ends in a C or G. Homopolymers are not allowed within the tag sequence. For example, if the oligomer is TCTCTGTG its reverse complement is CACAGAGA, so the adapter in full is GCCCGGGCTCTCTGTG-Sequence-CACAGAGAGCCCGGGC, half of the tag is cut off by SrfI so the sequence reads GGGCTCTCTGTG-Sequence-CACAGAGAGCCC

The PTS method is detailed in a Natural Protocols paper from Meyer and colleagues (Meyer et al., 2008). The order of the PTS and library preparation

stages can be summarised: Multichannel reaction set-up was followed by blunt-end repair and ligation of sample-specific self-hybridising barcoding adapters to both ends of the molecule. Nicks introduced during adapter ligation were repaired by a strand-displacing polymerase, *Bst*, samples were re-quantified by picogreen assay and pooled into equimolar ratios and any unligated molecule ends were excluded by dephosphorylation and *Srf*I restriction enzyme digestion. Universal 454 primers were blunt-ligated to the pooled template and the sample sent to University of Liverpool, Advance Genomic Facility for generic library preparation and 454 sequencing.

2.12.4 De-tagging sample-specific sequencing reads

Novobarcode (www.novocraft.com/userfiles/file/NovoBarcode.pdf) was used to de-multiplex barcoding adapters embedded in 5' and 3' ends of sequencing reads and group them in per-individual FASTQ files. Reads with low quality tag alignments were written to a catch-all file with the tag sequence intact.

2.12.5 Assessment of false assignment rate

To estimate the reliability of PTS in this experiment, the false assignment frequency due to sequencing error and cross-contamination was calculated using the equation:

$$\frac{F}{T} \times \frac{N}{A - N}$$

F = Number of sequences carrying tags from unused barcoding adapters

T = Total number of sequence reads obtained in the experiment

N = Total number of barcoded samples that were sequenced in parallel

A = Total number of barcoding adapters within the chosen category that have actually been synthesized

The barcoding adapters that were generated and used are oligomers with a minimum of three substitutions difference between each other, to minimise the false assignment rate

2.12.6 Generation of variant profiles

Variant profiles were generated using an in-house variant calling pipeline assembled by Michael Simpson at Kings College London (m.simpson@kcl.ac.uk). Briefly, sequence reads were aligned to the reference genome (hg18) with Novoalign (Novocraft Technologies Sdn Bhd) and anomalous reads (duplicates and those with multiple mapping coordinates) were excluded from downstream analysis. Depth and breadth of sequence coverage was calculated using custom scripts and the BedTools package (Quinlan and Hall, 2010) and visualised using the integrated genomes viewer (IGV). Single nucleotide substitutions and small insertion deletions (indels) were identified and quality filtered within the SamTools software package (Li and Durbin, 2010) and using in-house software tools, these variants were visualised using the IGV(Robinson et al., 2011). Annotation of variants with respect to the two most abundant transcripts of the *TNFSF4* gene was accomplished using the Variant Classifier tool (Li and Durbin, 2010).

2.12.7 Identification of novel variants

Novel variants were identified after converting their coordinates from UCSC hg18 to hg19 (based on the February 2009 high coverage assembly GRCh37) using the UCSC LiftOver tool (URL: http://genome.ucsc.edu/cgi-bin/hgLiftOver). This enabled the screening of identified variants against SNPs and structural variations found in Ensembl genome browser 64, dbSNP131, HapMap data release 28, 1000Genomes high coverage trios, 1000G high coverage exons and 1000G low coverage data. Novel exonic variants identified in the *TNFSF4* gene were also probed against those identified in 350 control exomes sequenced and analysed by the method described above at Kings College London.

In addition novel variants were probed against novel *TNFSF4* exonic variants found in the first data freeze of 2500 European and African-American control exomes and contained within the Exome Variant Server from the NHLBI Exome Sequencing Project (ESP) (URL: http://evs.gs.washington.edu/EVS/).

2.12.8 Jaspar: *sequence-based approach using curated binding profiles*

On identification of causal variants, the encompassing DNA sequence was examined for interaction with regulatory proteins including transcription factors (TFs). The alleles of associated variants identified were investigated for their impact on binding affinity of the motif for target proteins. In addition, the same SELEX binding data and position weight matrix (PWM) profiles (curated and stored in the Jaspar core database) were used to investigate DNA sequence motifs for degeneracy (Portales-Casamar et al., 2010).

2.12.9 Polyphen-2: sequence and structure-based approach

The prediction algorithm Polyphen-2 (Adzhubei et al., 2010) was applied to all newly identified coding region variants to evaluate the effect of non-synonymous SNPs (nsSNPs) on the TNFSF4 protein sequence. The *in silico* predictions were intended to guide future experiments. There was the possibility that protein structure would be affected by variants which influence disease susceptibility; thus both orthologs and paralogs were used in multiple sequence alignment (MSA) by the Polyphen-2 tool (Adzhubei et al., 2010). Homologs of the *TNFSF4* sequence were identified and aligned; the amino acid sequence was refined and clustered with regards to accuracy. The ancestral allele was compared with its replacement at each locus for several features including hypermutability and the relative fit of the replacement allele with respect to the adjacent alleles. A naive Bayes classifier predicted the functional significance of the replacement allele. The HumDiv dataset of 3,155 damaging alleles was used in the first *in silico* prediction, the HumVar (Capriotti et al., 2006) dataset of 13,032 human disease-causing mutations from UniProt (URL: http://www.uniprot.org/) were used in the second.

Chapter 3

Evaluating *TNFSF4* expression

For the most part, expression of *TNFSF4* on cells that control immune functionality is inducible: The gene must respond to a stimulus prior to functioning as a ligand at the beginning of a signalling cascade. This suggests temporal regulation of *TNFSF4* during the immune response. Evaluating *TNFSF4* for its utility in terms of dissecting disease pathogenesis, a simple and direct rationale is that risk-associated polymorphisms modulate SLE pathogenesis by causing aberrant expression of *TNFSF4*. The functional relevance of identified *TNFSF4* risk-variants were assessed for correlation with cell surface expression and the data presented in this chapter.

3.1 Evaluating TNFSF4 cell-surface expression- study aims

Expression of inducible immune-related genes can be used as a broad measure of the activity of the immune system. Conversely, the dynamics of gene expression might be perturbed by the disease process, and so influence severity. The aim of the preliminary studies presented in this chapter is to quantify expression of *TNFSF4* in $TNFSF4_{risk}$ and $TNFSF4_{non\text{-}risk}$ EBV-transformed cell lines. Quantifying the expression in non-SLE samples selected for their genotype will better clarify the role of *TNFSF4* as a susceptibility gene or as a modulator of disease severity. Warren repository samples held at JDRF/DIL were used for these preliminary studies. Evaluating the relationship between *TNFSF4* genetic variation and expression in cell-lines will be used to guide future larger scale experiments in genotype-relevant controls and SLE individuals.

Polymorphisms in genes at the start of immune-system signalling cascades can have a disproportionate effect on phenotype due to the amplification of minor affects (Sackton and Clark, 2009). The global perturbation of the immune system in SLE means it is difficult to separate the proportion of this aberrance due to genetic variance from that due to the disease process. Polychromatic fluorescent staining of peripheral blood mononuclear cells and Warren EBV-LCL cells was used to evaluate activated CD86+TNFSF4+ double-positive cells in the aforementioned risk and non-risk subgroups. The correlation between *TNFSF4* risk and non-risk genotype and *TNFSF4* cell-surface expression was evaluated for these cells. An important objective of this study was to determine differences in protein expression between cell-lines and lupus patients and evaluate the role of *TNFSF4* as a causal gene or modifier of disease progression.

Results

3.2 TNFSF4 cell-surface expression on LCLs and PBMCs in *TNFSF4*$_{risk}$ and *TNFSF4*$_{non\text{-}risk}$ homozygotes

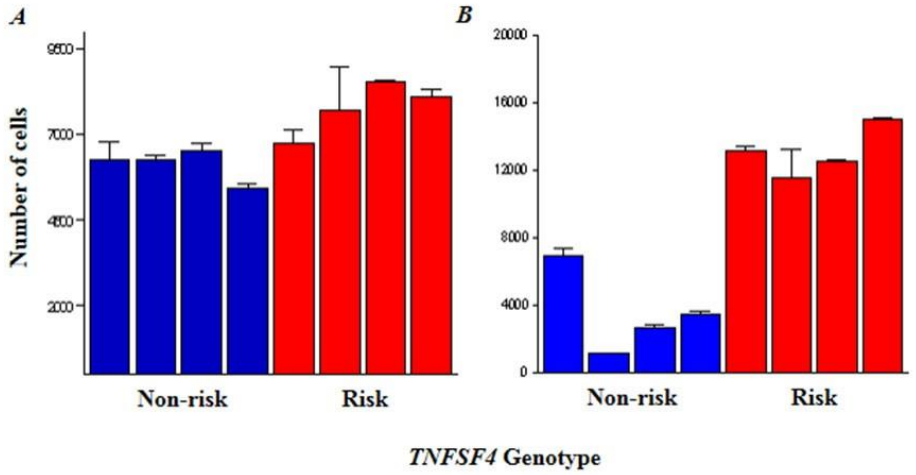
In order to better understand how *TNFSF4* acts as a susceptibility gene in SLE, I evaluated the upstream *TNFSF4* risk and non-risk haplotypes for influence on TNFSF4 expression in EVB lymphoblastoid cell lines (LCLs) (provided in collaboration the JDRF/WT DIL laboratory, Cambridge, UK) and in peripheral blood mononuclear cells (PBMC) from our UK-European SLE collection. Details of the European cohort are described in chapter 2 (2.1.1) of this thesis. Haplotypes were defined by genotyping locus-specific SNPs in multiple Northern European parental-proband and SLE-control cohorts (Cunninghame Graham et al., 2008). In excess of 1,000 LCLs were genotyped at the haplotype tagging SNPs rs2205960, rs1234314 and rs7514229 at the JDRF/WT DIL laboratories. LCLs were selected for *TNFSF4* risk or non-risk haplotype homozygosity on the basis of this genotyping. I re-sequenced the risk-haplotype tagging SNPs rs10912580, rs12039904 and rs1234317 and non-risk haplotype-tagging rs844644 to confirm homozygosity over a longer span of the *TNFSF4* locus in a smaller subset of 12 JDRF/DIL samples selected for expression analysis.

A time series was used to determine optimum activation of *TNFSF4* in LCLs and PBMCs after CD40L and anti-IgD stimulation and cell-surface expression was found to be optimal at 48 hours post-treatment. A slight increase in cell-surface expression of *TNFSF4* protein in risk relative to non-risk homozygote LCLs (**Figure 3.1A**) was found but this was not statistically significant (Mann-Whitney $P>0.05$). The trend in cell surface *TNFSF4* expression was found with borderline statistical significance (Mann-Whitney $P=0.02$) for PBMCs from UK SLE individuals selected for the same risk and non-risk *TNFSF4* alleles (**Figure 3.1B**)

Investigating the geometric mean (Gmean) fluorescent intensity of activated LCL cells found similar intensities in the expression of CD86 between the two genotype subgroups, but a higher intensity of TNFSF4 in the risk homozygote groups (**Figure 3.2**).

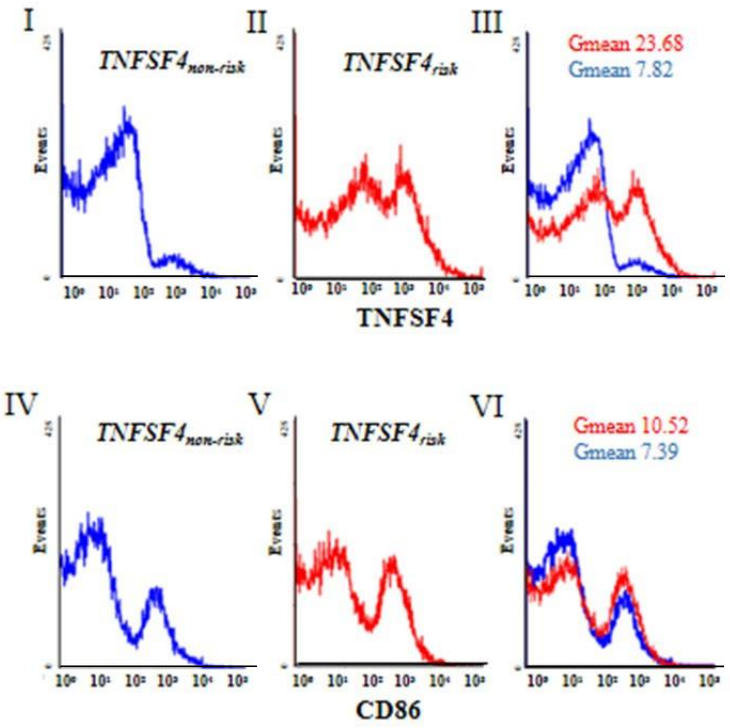
Investigating the geometric mean (Gmean) fluorescent intensity of activated PBMCs also found similar intensities in the expression of CD86 between the two genotype subgroups, but a higher intensity of TNFSF4 in the risk homozygote groups (**Figure 3.3**). Double positive analysis of TNFSF4 and CD86 expression in stimulated and unstimulated PBMCs from UK SLE risk and non-risk homozygote probands found the percentage of CD86+TNFSF4+ cells within the activated cell fraction of PBMCs to be increased in the risk group (**Figure 3.4**).

Figure 3.1 Cell-surface expression of TNFSF4 on LCL-cells and peripheral blood cells



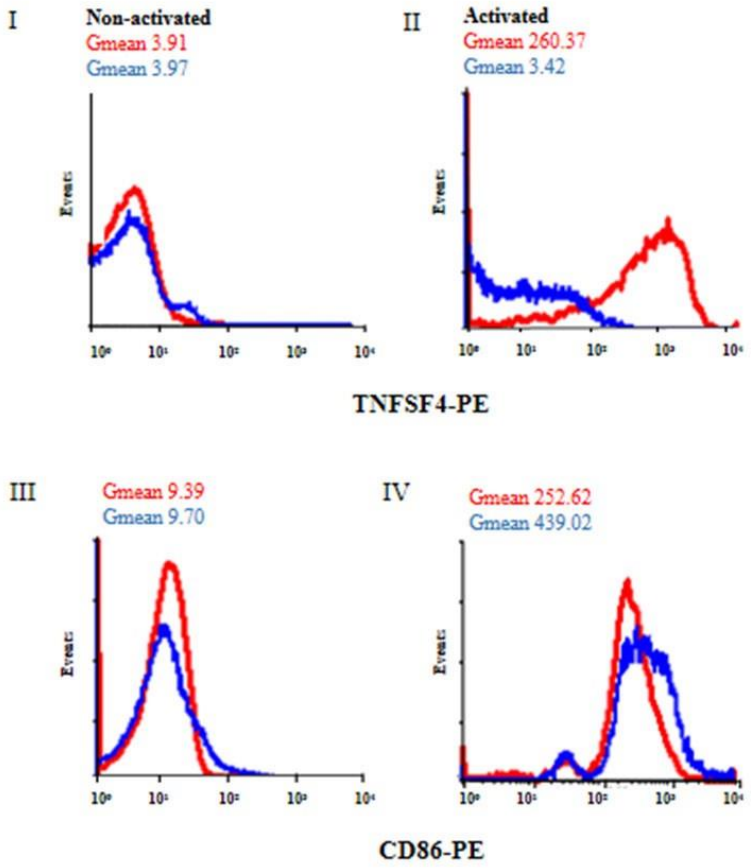
A. Numbers of CD40L and anti-IgD-stimulated cells in eight different homozygous cell lines (two $TNFSF4_{\text{non-risk}}$ and three $TNFSF4_{\text{risk}}$). Each bar represents the mean of two independent replicates. *B*. Numbers of TNFSF4+ PBLs taken from eight SLE-affected probands (four homozygotes for each of the risk- and non-risk haplotypes), 48hrs after CD40L- anti-IgD stimulation. Each bar represents the mean of two independent replicates.

Figure 3.2 Cell-surface expression of CD86 and TNFSF4 in EBV-LCL cells



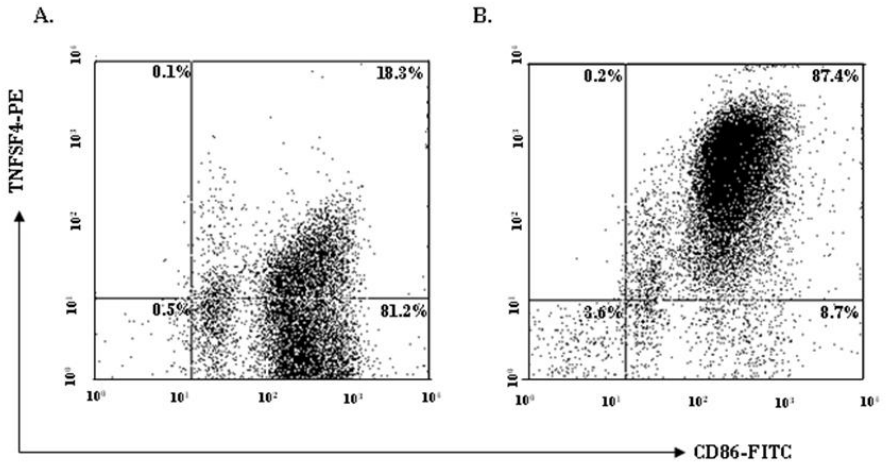
Histograms of CD40L anti-IgD-stimulated EBV-LCL cells showing expression of TNFSF4 ((I), (II) and (III)) and CD86 ((IV), (V), and (VI)) for representative samples homozygous for the non-risk haplotype (blue) (*TNFSF4*$_{\text{non-risk}}$, (I and IV)) and for the risk haplotype (red) (*TNFSF4*$_{\text{risk}}$, (II and V), respectively. Geometric mean (Gmean) values of the fluorescent intensity are shown for each marker. Data presented are from one of two experiments with similar results on the same cell lines.

Figure 3.3 Influence of stimulation on TNFSF4 and CD86 expression by PBMCs



TNFSF4 (I and II) and CD86 expression (III and IV) in CD40L/anti-IgD–stimulated and unstimulated PBMCs. These cells were taken from UK SLE probands that were homozygous for the non-risk haplotype (shown in blue) and for the over-transmitted risk haplotype (shown in red). The Geometric mean (Gmean) values presented are shown for the cell-surface markers in two individuals.

Figure 3.4 Representative FACS dot plots of stimulated PBMCs



Cells were taken from probands expressing two copies of (A.) the non-risk haplotype and (B.) the risk haplotype. The percentage of CD86+TNFSF4+ cells within the activated cell population is indicated in the upper left quadrant. Cells were designated TNFSF4-ve if TNFSF4 expression fell within background staining compared to a mouse IgG1negative control mAb (MOPC 31C, Ancell).

3.3 Discussion

3.3.1 Summary of findings

To better understand how *TNFSF4* may act as a disease susceptibility gene in SLE, $TNFSF4_{RISK}$ and $TNFSF4_{NON\text{-}RISK}$ homozygote lymphoblastoid cell lines (LCL cells) and peripheral blood mononuclear cells (PBMCs) from the UK SLE cohort were investigated for their cell-surface TNFSF4 expression. The risk haplotype correlated with increased expression of both cell-surface TNFSF4 and *TNFSF4* transcript.

3.3.2 Results in the context of published work

I hypothesized that variation in the upstream region of the gene increased the expression of TNFSF4, and, through TNFRSF4, increased co-stimulation for CD4+ T-cells and/or further activated the TNFSF4-expressing (Stuber et al., 1995) APCs. This increased expression of TNFSF4 may act by destabilizing peripheral tolerance through inhibiting the generation of IL-10–producing CD4+ type 1 regulatory T-cells (Ito et al., 2006). Notably, *TNFSF4* has also been associated with susceptibility to atherosclerosis (Wang et al., 2005), and individuals with SLE are prone to accelerated arterial disease. The role of *TNFSF4* in the pathogenesis of SLE highlights the importance of the role of the T-cell–APC interaction in this disease, a conclusion supported by the genetic influence, albeit modest, arising from the CTLA4-ICOS locus (Cunninghame Graham et al., 2006).

3.4 Limitations to the data

The primary aim of the data presented in this chapter was to evaluate the trend between SLE- associated *TNFSF4* genetic variants and expression of the gene product in cell lines from individuals without lupus and in PBMCs from individuals with disease. These data were intended as a pilot study, using few individuals and therefore interpretations are limited in their scope. Although the

trend between genotype and expression appears robust there are a number of sources of potential error which allow only very limited interpretation of these data.

3.4.1 Cell lines

The data presented in this chapter find increased cell-surface expression of the *TNFSF4* gene *in ex vivo* peripheral cells compared to EBV- cell lines: Both cell types were selected for the same genotypes. For the latter group, there was no difference in gene expression between *TNFSF4* risk and non-risk homozygotes (Figure 3.1*A*). Although the cell lines were a significant tool in the research presented in this chapter, they would have undergone significant mutations during EBV-transformation, limiting their fidelity. Clonality due to continuous passaging of the cell-lines may affect expression of their cell-surface proteins, so limiting biological relevance with regards to TNFSF4 cell surface-expression.

3.4.2 Variation in disease activity

Adjustments were also not made for disease activity, which was variable, or for inflammatory disorders. There was no global assessment of disease activity, or acute clinical variables, at the time data was collected, which would influence results. Disease activity can be assessed by a rheumatologist using a standardized disease activity index, SLEDAI (Bombardier et al., 1992). These data could have been used as a covariate in a standard regression model, to adjust data.

3.4.3 Use of multiethnic SLE cohort

Samples used from SLE individuals in this experiment were from a diverse mix of South Asian, UK European and Afro-Caribbean patients who regularly attended the West London Rheumatology Clinic, Hammersmith Hospital, UK. Chapter 4 of this thesis describes the ancestry differences at the *TNFSF4* locus in detail. With regards to disease activity, this is a potential source of random error.

At the time of the experiment, only a limited number of samples with the relevant genotype were available.

3.4.4 Absence of longitudinal data

Correlation research such as undertaken in this chapter, in individuals with SLE, which is a relapse-remitting condition, requires repeat measurements of expression over a time course to more accurately reflect observed changes. Although we attempted these additional studies, they were not undertaken due to difficulties in recruiting patients at the same time.

3.4.5 TNFSF4 peak expression

Samples were harvested 48hours post stimulation; this time point was selected for peak expression of cell-surface *TNFSF4* and determined by flow cytometry. During optimisation of the protocol, cells expressed a proportion of TNFSF4 at 24hours. The abundance of transcript was not quantified in these cells and the peak transcript level is probably not coincident with that of peak protein, but in all likelihood precedes it. Real-time PCR data which accurately reflects transcript expression changes temporally in each subgroup is required so that we are better informed with respect to pathogenic mechanism. Bias in base composition between $TNFSF4_{\text{risk}}$ and $TNFSF4_{\text{non-risk}}$ transcripts could lead to efficient degradation of $TNFSF4_{\text{non-risk.}}$

Chapter 4

Trans-ancestral mapping of *TNFSF4* in SLE

SLE segregates with race and geographical location: Epidemiological studies for this trait suggest increased disease burden in non-Europeans. However, Northern and Western European SLE cohorts have been used most extensively in studies that have directed research (Simard and Costenbader, 2007). Since these doctoral studies were planned, GWAS has identified or confirmed the association of immunologically relevant SLE loci, including *TNFSF4*, with disease (Harley et al., 2008; Hom et al., 2008; Han et al., 2009; Yang et al., 2010). To date, the published GWA studies have confirmed association of an extended haplotype consisting of risk-associated variants at *TNFSF4* in European and East Asian SLE cohorts. Pathological variables associated with severe disease are less common in Europeans, the population most likely to attain remission (Korbet et al., 2007). Thus, extensive research is required to clarify the role of genetic risk, socioeconomic status and quality of care in non-Europeans with SLE, particularly forms of the disease associated with severe phenotypic manifestations.

4.1 Trans-ancestral mapping experiment- study aims

4.1.1 Re-evaluation of TNFSF4 association in Europeans

Although data documenting the *TNFSF4* association with lupus were not published at the time these doctoral studies were planned, it was understood that in Europeans strong LD across the locus was an obstacle in the delineation of *TNFSF4 c*ausal variation. To this end, I re-evaluated the disease association in the largest European SLE-control cohort available at the time. The aim of using this cohort was to resolve the association signal owing to *TNFSF4* in European SLE with increased power.

4.1.2 Evaluation of TNFSF4 in non-Europeans

A second aim of the work presented in this chapter was to evaluate the *TNFSF4* locus in multiple non-European SLE cohorts: The six cohorts tested for association included African-Americans, African-American Gullah, East Asians and SLE individuals of Amerindian descent. These groups are disproportionately affected by SLE and health disparities in these groups show onset at an earlier age. By investigating multiple SLE-control cohorts, I wished to establish whether *TNFSF4* risk in SLE is population-specific or global. The locus was fine-mapped in each group; the density of variants genotyped in this study greater than double those in the original European study (Cunninghame Graham et al., 2008). Selected combinations of haplotype-tagging and proxy SNPs were tested in each group to capture the majority of common SNPs by imputation.

4.1.3 Inference of recombination rate

An estimation of recombination rate across different ancestral groups is included in this chapter: The map available from HapMap phase II is population-averaged, the 1000 Genomes map at this locus based predominantly on low coverage, sex-averaged data. These maps give different recombination patterns at *TNFSF4* compared to the deCODE maps for the same ancestry. The European deCODE

sex-averaged and female-only recombination maps (URL: http://www.decode.com/addendum/), are based on 15,257 and 8,850 directly observed recombinations respectively. These maps have a resolution effective down to 10kb and I compared the standardised recombination rate for deCODE females and deCODE males to the HapMap phase II combined map (Appendix D). For the 200kb region at chromosome 1q25.1 including *TNFSF4*, I found recombination differences between the maps. Therefore, to more accurately decipher genetic architecture at *TNFSF4*, I attempted to resolve the fine-scale recombination rate in large numbers of control chromosomes matched for four SLE cohorts tested for association. The populations comprised African-American, East Asian, European and Hispanic control individuals. The recombination rate and presence of hotspots was also evaluated in $TNFSF4_{\text{risk}}$ vs. $TNFSF4_{\text{non-risk}}$ homozygote individuals.

4.1.4 Evaluate haplotypic association

The haplotypic association with SLE was investigated in African-American, East Asian, European and Hispanic cohorts and data presented in this chapter: Strong pair-wise LD between SNPs meant a 100kb haplotype upstream of *TNFSF4* correlated with risk of disease in the original European study. African populations tend to have shorter haplotypes because they are often subdivisions of the larger haplotypes found in non-Africans and so can be correlated to these (Daly et al., 2001). In Hispanics and African-Americans the genetic component attributable to the West African ancestral population *TNFSF4* would equate to a faster decay of LD, much greater in African-Americans, with component estimates upwards of 80%, compared to Hispanics with ranges of 4-11% (Price et al., 2007; Winkler et al., 2010). Performing high-resolution trans-ancestral association mapping with tag SNPs and proxy variants, I aimed to anchored haplotypes in ancestral and admixed populations. A principal components (PC)-based strategy was used to adjust for major ancestry using a set of genome-wide ancestry informative markers. Higher numbers of diverse haplotypes are predicted due to differences in genetic architecture at *TNFSF4*. Recombinant

haplotypes, unique to African-American and Hispanic individuals, were used to resolve the association with disease.

4.1.5 Evaluate TNFSF4 association with sub-phenotypes

Data on clinical variables and phenotypic manifestations for all populations surveyed in this chapter were collated. Analysing *TNFSF4* alleles in phenotypic subsets of SLE cases, I aimed to enrich for risk variants with increased effect size, in the hope that associations would prove informative for causal mechanism as these SLE sub-groups are less heterogeneous than SLE per se.

Results

4.2 QC filtering and population demographics

To delineate causal variation at *TNFSF4,* SNPs in a 200kb section of chromosome 1q25 encompassing *TNFSF4* (23.6kb) and 150kb of the 5′ region were genotyped. Population stratification bias and effects due to admixture were addressed using the approach of Namjou and colleagues (Namjou et al., 2009). Genotyping 347 genome-wide SNPs, as selected by Halder and colleagues (Halder et al., 2008) allowed for correction of major ancestry in each population. A PCA-based approach was used to do so (**Figure 4.1, Figure 4.2, Figure 4.3).** As outlined in chapter 2, SNPs and individuals that failed quality control were filtered. Pre- imputation, the cohorts were; African-Americans (88 SNPs in 1529 cases, 2048 controls), African-American Gullah (51 in 152, 222) East Asians (65 in 1500, 1396), Europeans (89 in 3432, 3076) and Hispanics (51 in 1348, 717). A detailed description of the component sample sets is presented in **Table 4.1**.

Table 4.1 **Population demographics and imputation reference data for SLE-control cohorts post QC filtering**

	European			East Asian			Hispanic			AA-Gullah		
	Case	Control	ALL	Case	Control	ALL	Case	Control	ALL	Case	Control	ALL
Males	344	1151	1495	167	225	392	119	73	192	136	593	729
Females	3088	2489	5577	1333	1171	2507	1229	644	1872	1541	1341	2882
Unknown										3	236	239
TOTAL	3432	3640	7072	1500	1396	2896	1348	717	2065	1680	2170	3850
SNPS, TYPED	89			65			51			88		
SNPS, ALL	244			450			460			393		
Imputation reference 1	*CEU, *TSI			*CHD, *CHB			-			*YRI		
Imputation reference 2	WTCCC controls											

Numbers after filtering for duplicates, FDRs, HWE, missingness and major ancestry. post SNPs with INFO scores <0.7 excluded, SNPS with HWE<0.01 excluded. *HapMap phase III haplotypes, details of individual populations are found in abbreviations

4.3 Phasing and imputation

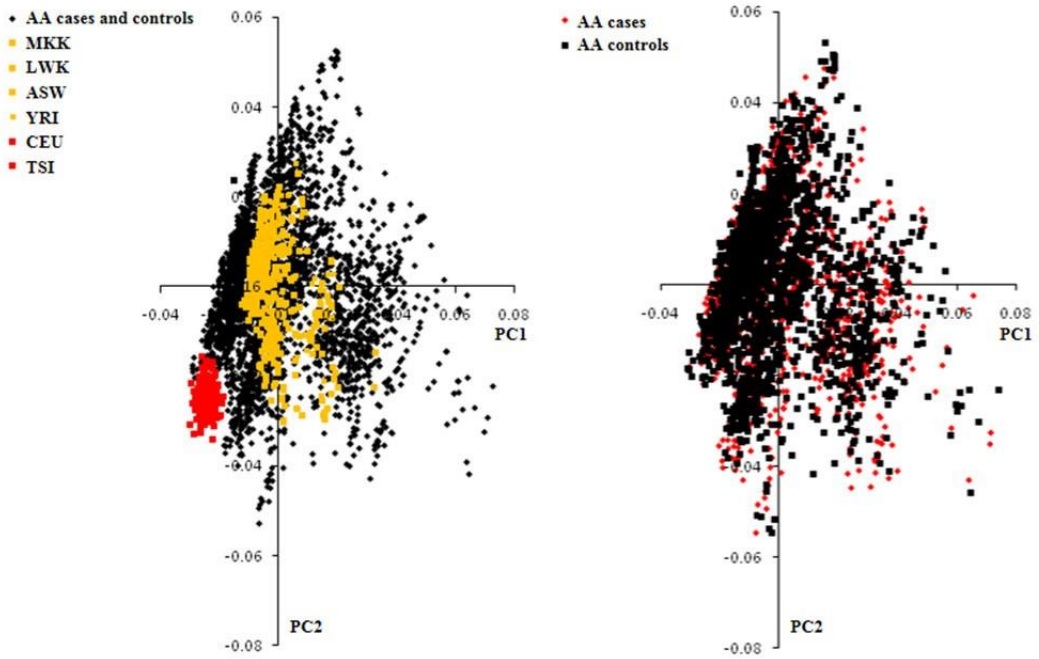
To directly compare genotyped SNPs, the phased chromosomes of HapMap phase III samples and the IMPUTE 2.1 algorithm were used, together with the second reference sets defined in **Table 4.1.** Missing genotype data and SNPs with a MAF>0.03 were imputed. As described in methods, filtering to include only genotyped and high quality imputed SNPs resulted in final datasets for Europeans (112 SNPs in 3432 cases, 3640 controls), East Asians (100 SNPs in 1500, 1396) and African-Americans (121 SNPs in 1529, 2048). The Hispanic/Amerindian cohort was not imputed.

4.4 Inference of fine-scale map of recombination rate

An accurate map of the recombination rate facilitates mapping multiple independent contributors to disease risk at a single locus. The European sex-averaged and female-only recombination maps generated by deCODE (URL: http://www.decode.com/addendum/) (Kong et al., 2010), are based on 15,257 and 8,850 directly observed recombinations, respectively. These maps have a resolution effective down to 10kb; Comparing the deCODE map to the HapMap phase II (The International HapMap Project, 2007) combined map, there were clear differences in the recombination pattern at the *TNFSF4* locus (Appendix D). The differences in recombination were greater between the deCODE female-only and HapMap maps. The SLE cases used for association testing presented in this chapter were predominantly female (82.33%). The recombination rate between the two maps differed in the 5′ region spanning the $TNFSF4_{risk}$ haplotype. Thus, I estimated background recombination rates in African-Americans, East Asians, Europeans and Hispanics using *rhomap***,** a Bayesian composite-likelihood method.

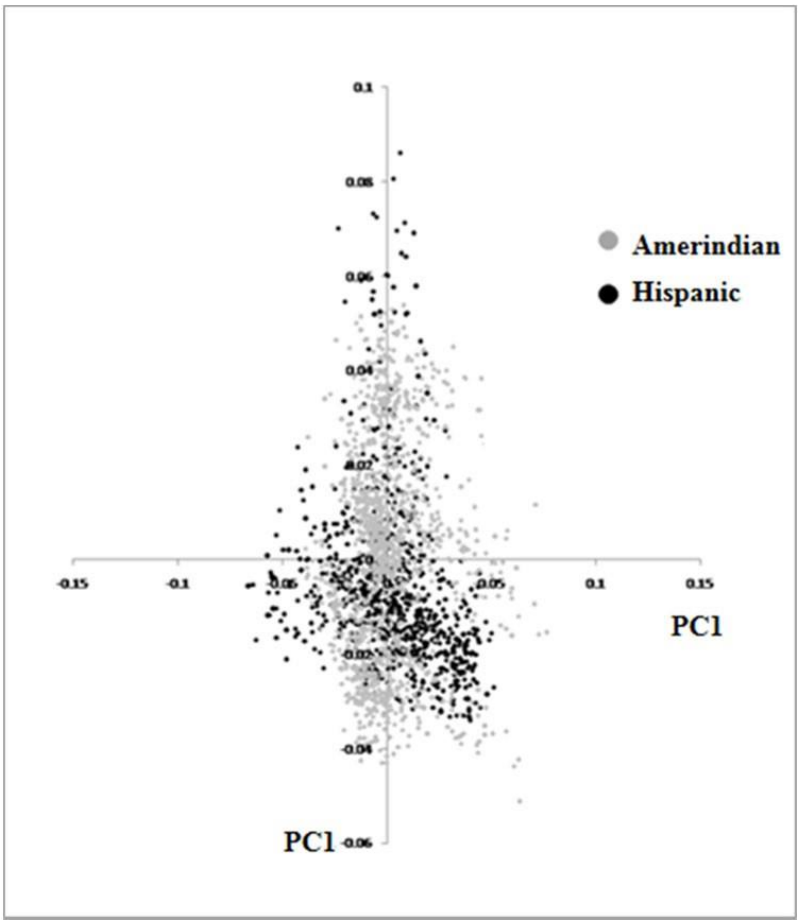
The inclusion of a hotspot model allowed sampling of hotspots from the Markov chain and inference of mean posterior hotspot densities from a threshold upwards of 0.25, giving a detection power of 50% and a false-discovery rate of 4% (Auton and McVean, 2007). In Asians, Europeans and Hispanics the bulk of the recombination occured in less than 5% of sequence (**Figure 4.4**). An exception to this pattern was found in the African-American cohort, with increased recombination rate and higher density and proportion of hotspots across the locus (**Figure 4.5**). In all populations, peak recombination was at the 5′ boundary of the *TNFSF4* gene and approximately 120kb into the 5′ region. The recombination extended 30kb further from the *TNFSF4* gene boundary into the 5′ region in African-Americans compared to negligible recombination in this region for the other populations (**Figure 4.4**). This is compatible with increased complexity of the genomic region in African-Americans.

Figure 4.1 A plot of PC1 vs. PC2 for the African-American population



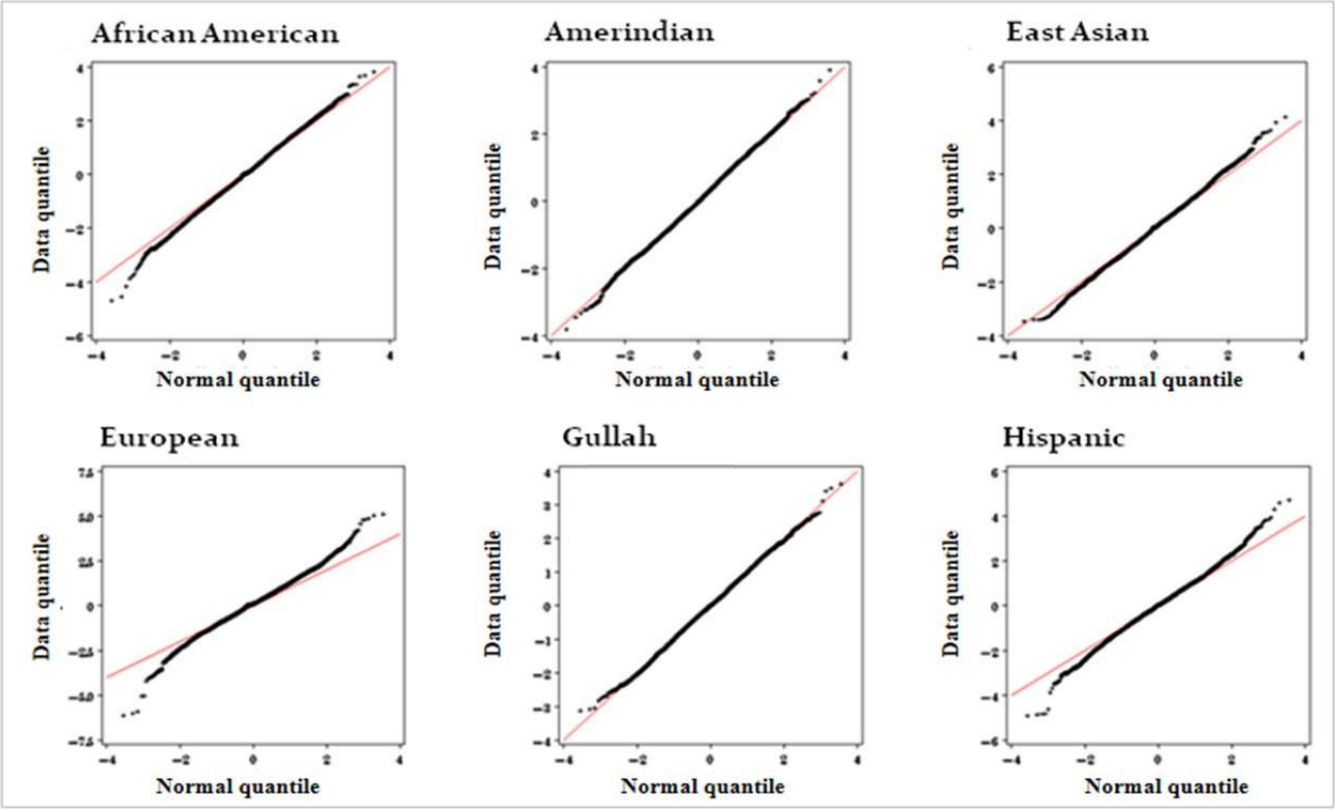
Left, Principal component (PC) 1 versus PC2 analyses of four HapMap phase III African (yellow) and two HapMap III European (red) populations and our African-American SLE-control cohort (black). *Right.* Population stratification between African-American cases (red) and controls (black) was minimised by principal components analysis using 347 major ancestry informative markers. This figure depicts the most profound ancestry differences along continuous axis of variation between cases and controls after QC filtering of the AA cohort.

Figure 4. 2 Plot of PC1 vs. PC2 for the Amerindian (grey) and Hispanic (black) cohorts using genome-wide AIM markers



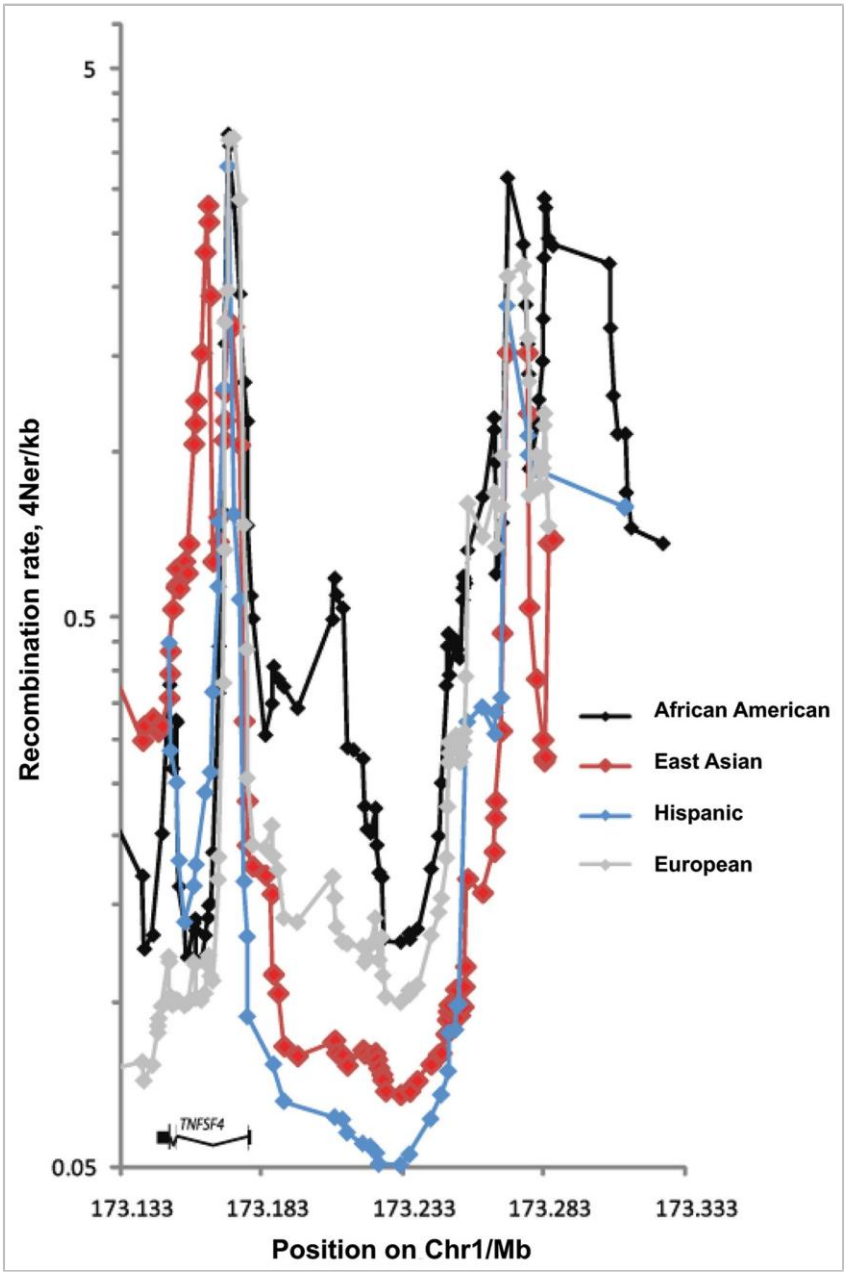
Population stratification between Amerindians (Grey) and Hispanic (black) cohorts was minimised by principal components analysis using 347 major ancestry informative markers. This figure depicts the most profound ancestry differences along continuous axis of variation after QC filtering of the cohorts.

Figure 4.3 Quantile-quantile (QQ) plots for p-values for each dataset



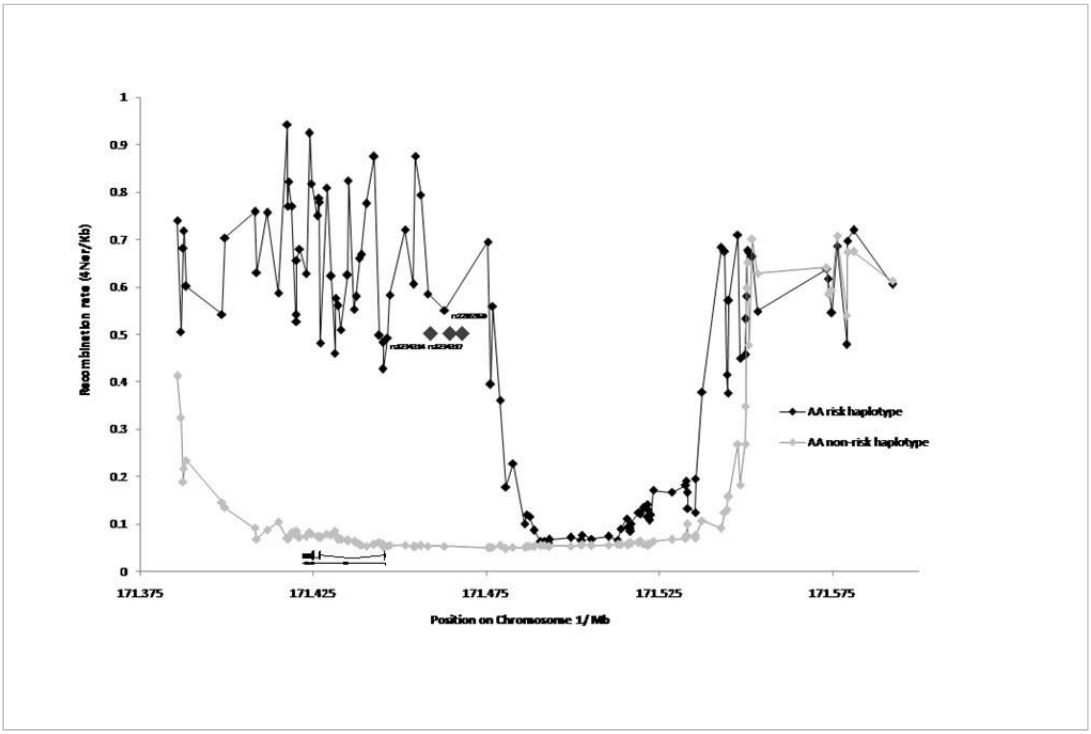
The figure illustrates the distribution of –log10 (expected p-value) on the x-axis, and –log10 (observed p-value) on the y-axis. Close adhesion of p- values to the red line, corresponding to the null hypothesis, was visualised for all but the European dataset (λ=1.3). Data were plotted for 347 ancestry informative markers and adjusted for values of PC1, PC2 and PC3 to confirm the effectiveness of the AIM panel. The log scale emphasised the smallest p-values. AIM SNPs were genotyped in the cohorts of African-American, Amerindian, East Asian, European, Gullah and Hispanic individuals (courtesy of Professor Carl Langefeld, Wake Forest, US).

Figure 4.4 Fine scale maps of recombination rate inferred for four populations



1568 randomly assigned control phased chromosomes from East Asian, European, Hispanic and African-American populations were tested using *rhomap* from the LDHAT2.0 package to infer the fine-scale map of recombination rate (4Ner/Kb). 200kb of chromosome 1q25.1 encompassing *TNFSF4* gene, and extended 5′ and 3′ regions were tested. R*homap* was run for a total of 1,100,000 rjMCMC iterations including a burn-in of 100,000 iterations and sampled the burn-in every100 iterations.

Figure 4.5 Comparison of recombination at *TNFSF4* in African-American $TNFSF4_{risk}$ and $TNFSF4_{non\text{-}risk}$ individuals



Phased chromosomes from African-American SLE individuals homozygous for $TNFSF4_{\text{risk}}$ (n=10) and $TNFSF4_{\text{non-risk}}$ (n=10) were tested using *rhomap* from the LDHAT2.0 package to infer the fine-scale map of recombination rate (4Ner/kb) across 200kb of chromosome 1q25.1 encompassing *TNFSF4* gene, and extended 5′ and 3′ regions. Individuals were identified as homozygous for $TNFSF4_{\text{risk}}$ or $TNFSF4_{\text{non-risk}}$ if they had two copies of AGTTCT (risk) or ACTTCT, (non-risk). For each simulation, *rhomap* was run for a total of 1,100,000 rjMCMC iterations including a burn-in of 100000 iterations, sampling the chain after every 100. Grey diamonds indicate the location to scale of SNPs significantly associated with risk of SLE in this cohort, the *TNFSF4* gene is also located to scale under the graph.

Figure 4.6 Single marker associations of SNPs at *TNFSF4* locus in A. East Asian, B. European, C. Hispanic, D. African-American SLE populations

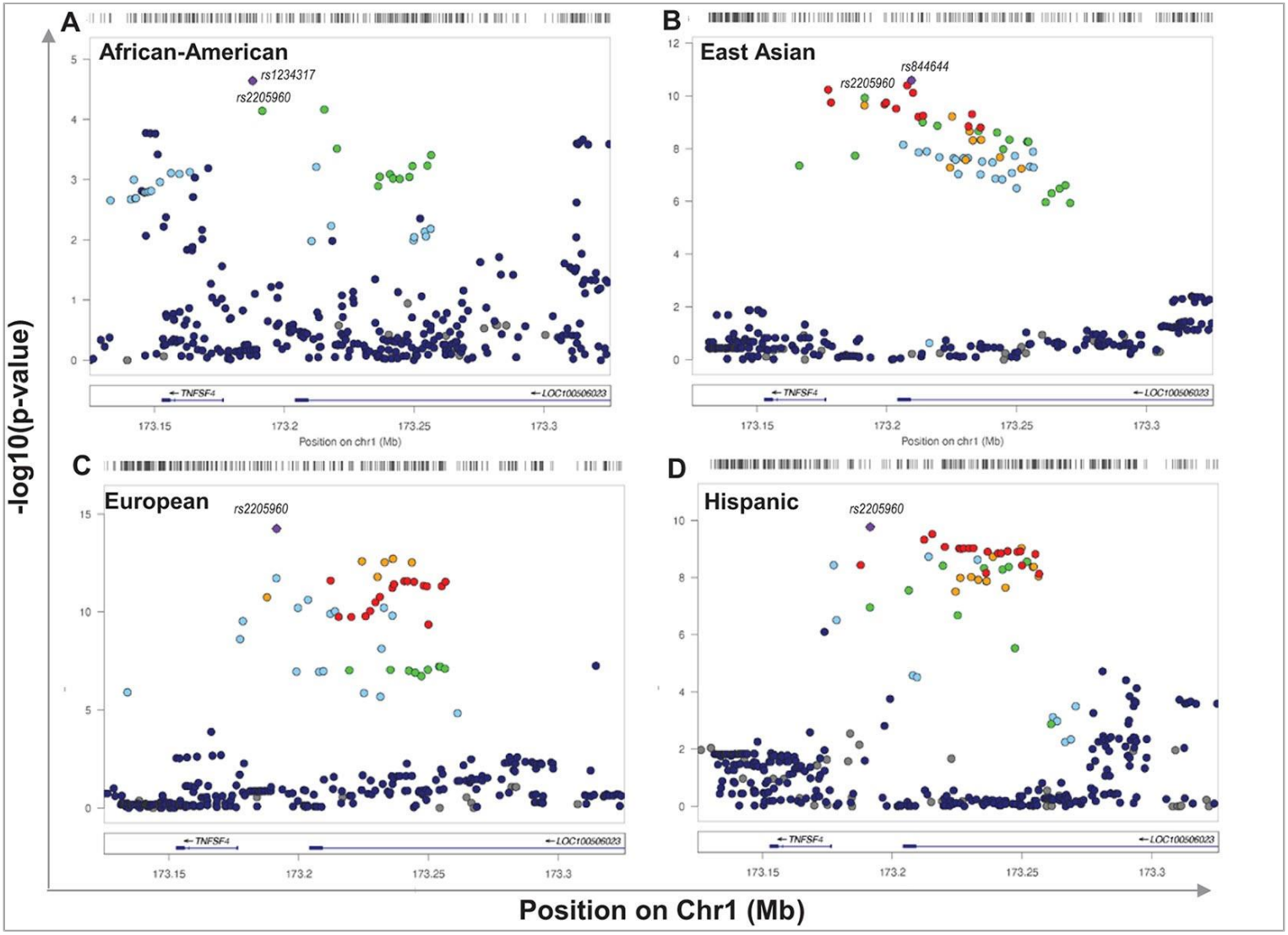


Figure 4.6 Single marker associations of SNPs at *TNFSF4* locus in A. East Asian, B. European, C. Hispanic, D. African-American SLE-control populations. This plot illustrates the strength of the association (-log10uncorrectedP) of markers in a 200kb segment of chromosome 1q25 encompassing *TNFSF4* with SLE versus chromosomal position (kb) in East Asians, Europeans, Hispanics and African-American populations. SNP names on each regional plot indicate the most associated SNPs in each group. SNPs are colour-matched for LD.

4.5 Single marker association of 5′ *TNFSF4* SNPs with SLE

A logistic regression model which was additive on the log-odd scale was used to evaluate *TNFSF4* variants. Under this model, the score test, an asymptotic test of hypothesis, was used to test association of the variants for the binary phenotype (case, control) under the null hypothesis. For non-imputed variants and the high certainty imputed SNPs with info>0.7 that were included in analysis post QC, the test statistic was reduced to the Cochran-Armitage Trend test statistic.

The association data presented is for markers with 90% of the alleles genotyped; imputation was used to compare haplotypes in this study. There was at least a base line of mutual dependence from the LD correlation coefficient r^2 scores between genotyped *TNFSF4* SNPs in each population and so they were permuted 5000 times to correct for multiple testing. The quantile of the score test statistic was interpreted by calculating p-values. An arbitrary locus-wide significance level for rejecting the null hypothesis was set at $P=5 \times 10^{-5}$. Odds ratios (OR) with 95% confidence intervals (95% CI) were taken from the exponent of the beta coefficient and standard error of the logistic regression model. Significance of association of corrected p-values was based on permutation testing (5000 permutations).

Data are represented as nominal uncorrected p-values and permuted (*Pp*) values for SNPs (**Table 4.2, Table 4.3**) and haplotypes. In East Asians, Europeans and Hispanics many strong associations ($P\ 10^{-8} < 10^{-16}$) at *TNFSF4* were detected: Multiple susceptibility alleles in the *TNFSF4* 5′ region were overrepresented in SLE cases (**Table 4.2, Figure 4.6**). In terms of single markers, best evidence of association with disease in Europeans was observed with rs2205960-T, 10kb 5′ from the *TNFSF4* gene ($P=4.6 \times 10^{-15}$, $P_p < 10^{-4}$, OR=1.34 (95%CI 1.25-1.44)).

The T allele of rs2205960 was also best-associated with Hispanic SLE ($P=7.24x10^{-11}$, $P_p<10^{-4,}$ OR=1.65 (95% 1.42-1.91)). Allele frequencies from the 1000 Genomes project cohorts for associated SNPs are found in Appendix C.

In Europeans, an additional 15 SNPs reached genome-wide significance ($P<10^{-8}$), 11 of the risk alleles also reach this level of significance in the East Asian and Hispanic cohorts (**Table 4.2**). Several 5′ risk alleles associated with disease in East Asians, Europeans and Hispanics were also associated in African-Americans and the association replicated in a small cohort of AA-Gullah (**Table 4.3**), underpinning this gene as a global SLE susceptibility gene. In African-Americans, the best evidence for the 5′ association with disease came from rs1234317-T ($P=1.45 \times 10^{-5}$, $P_p=0.01$, OR=1.39 (95%CI 1.19-1.62)) and rs2205960-T ($P=3.79 \times 10^{-5}$, $P_p=0.05$, OR=1.48 (95% CI 1.22-1.79)). Association p-values for these variants were greater than the arbitrary level of locus-wide significance assigned to African-derived populations ($P=5x10^{-5}$).

There was a trend for under representation of rs844644-A, located 30kb 5′ to the *TNFSF4* gene, in European ($P=2.55x10^{-7}$, $P_p<10^{-4}$, OR=0.85 (95% CI 0.8-0.9)), East Asian ($P=8.61x10^{-12}$, $P_p<10^{-4}$, OR=0.7 (95% CI 0.64-0.78)) and Hispanic (P=3.75x10-5, $Pp=8 \times 10^{-3}$, OR=0.74 (95% CI 0.64-0.86)) cases, consistent with our previous findings for this allele in two Northern European SLE cohorts(Cunninghame Graham et al., 2008). There was no evidence of association of either allele of this SNP in the African-American cohort (*P=0.69*, OR=1.03(95%CI=0.9-1.16)).

Table 4.2 **Single marker association results for East Asian (As), European (Eur) and Hispanic (Hisp) SLE-control cohorts**

Marker	Coordinate	[a]Location	A1	F_A/F_U			A2	CHQ			Unadjusted p-value, *P*			Odds Ratio (95% CI)		
				As	Eur	Hisp		As	Eur	Hisp	As	Eur	Hisp	As	Eur	Hisp
rs1234313	171.433	*TNFSF4 Intron1*	G	0.349/0.290	0.31/0.28	0.427/0.394	A	24.9	16.7	3.5	$6.17x10^{-7}$	$4.33x10^{-5}$	0.061	1.32(1.18-1.46)	1.17(1.08-1.24)	1.14(1-1.31)
rs16845607	171.440	*TNFSF4 Intron1*	A			0.147/0.08	G			33			$9.17x10^{-9}$			2.06(1.6-2.64)
rs1234314	171.444	*0.9 Kb 5' of TNFSF4*	G	0.466/0.384	0.472/0.426	0.421/0.533	C	43.4	33.2	41.2	$4.37x10^{-11}$	$8.13x10^{-9}$	$1.38x10^{-10}$	1.40(1.27-1.55)	1.20(1.13-1.28)	1.47(1.37-1.79)
rs1234315	171.445	*2Kb 5' of TNFSF4*	T	0.466/0.386	0.521/0.473	0.363/0.454	C	41.1	36.4	29	$1.42x10^{-10}$	$1.59x10^{-9}$	$7.24x10^{-8}$	1.39(1.26-1.54)	1.21(1.14-1.29)	1.47(1.28-1.69)
rs1234317	171.454	*11.3Kb 5' of TNFSF4*	T	0.321/0.252	0.306/0.256	0.391/0.285	C	36.7	48.4	39.3	$1.38x10^{-9}$	$3.56x10^{-12}$	$3.65x10^{-10}$	1.40(1.26-1.57)	1.28(1.19-1.38)	1.62(1.39-1.88)
rs2205960	171.458	*15Kb 5' of TNFSF4*	T	0.312/0.246	0.273/0.219	0.374/0.267	G	41.5	61.4	42.5	$1.18x10^{-10}$	$4.6x10^{-15}$	$7.24x10^{-11}$	1.44(1.29-1.61)	1.34(1.25-1.44)	1.65(1.42-1.91)
rs844644	171.476	*33Kb 5' of TNFSF4*	A	0.403/0.488	0.434/0.475	0.333/0.402	C	46.6	26.6	17	$8.61x10^{-12}$	$2.55x10^{-7}$	$3.75x10^{-5}$	0.7(0.64-0.78)	0.85(0.8-0.9)	0.74(0.64-0.86)
rs844645	171.477	*33.6Kb 5' of TNFSF4*	G	0.476/0.403			A	34			$5.44x10^{-9}$			1.35(1.22-1.49)		
rs12039904	171.479	*35.8Kb 5' of TNFSF4*	T	0.319/0.253	0.285/0.234	0.393/0.287	C	32.4	51.5	40.3	$1.23x10^{-8}$	$7.11x10^{-13}$	$2.23x10^{-10}$	1.38(1.23-1.54)	1.30(1.21-1.4)	1.61(1.39-1.87)
rs2795288	171.481	*37.5Kb 5' of TNFSF4*	A	0.489/0.410	0.468/0.417	0.523/0.42	T	38.6	40.3	33.9	$5.2x10^{-10}$	$2.21x10^{-10}$	$5.8x10^{-9}$	1.37(1.24-1.52)	1.23(1.15-1.31)	1.52(1.32-1.72)
rs1012507	171.486	*43Kb 5' of TNFSF4*	T	0.335/0.266	0.384/0.342	0.468/0.369	G	36.2	29.5	32.8	$1.78x10^{-9}$	$5.48x10^{-8}$	$1x10^{-8}$	1.39(1.25-1.55)	1.2(1.12-1.28)	1.51(1.31-1.76)
rs844648	171.490	*47.4Kb 5' of TNFSF4*	A	0.467/0.401			G	27.5			$1.61x10^{-7}$			1.31(1.18-1.44)		
rs844649	171.491	*47.9Kb 5' of TNFSF4*	C	0.422/0.356	0.333/0.279	0.441/0.338	T	29.1	53.8	35.8	$6.75x10^{-8}$	$2.19x10^{-13}$	$2.21x10^{-9}$	1.32(1.20-1-47)	1.29(1.21-1.38)	1.55(1.34-1.78)
rs844651	171.492	*48.7Kb 5' of TNFSF4*	G	0.438/0.361	0.426/0.386	0.504/0.405	T	38.1	23.3	31	$6.8x10^{-10}$	$1.9 x 10^{-6}$	$2.59x10^{-9}$	1.38(1.24-1.53)	1.18(1.1-1.26)	1.49(1.29-1.71)
rs704840	171.493	*49.7Kb 5' of TNFSF4*	G	0.412/0.351			T	29.8			$4.73x10^{-8}$			1.33(1.2-1.47)		
rs2840317	171.493	*50Kb 5' of TNFSF4*	A	0.318/0.252			T	33.1			$8.81x10^{-9}$			1.38(1.24-1.54)		
rs2901716	171.494	*51Kb 5' of TNFSF4*	A	0.319/0.252			G	34			$5.48x10^{-9}$			1.39(1.24-1.55)		
rs844654	171.499	*56.3Kb 5' of TNFSF4*	T	0.488/0.41	0.472/0.420	0.419/0.332	A	38.5	41.4	33.3	$5.4x10^{-10}$	$1.22x10^{-10}$	$7.91x10^{-9}$	1.37(1.24-1.52)	1.23(1.16-1.31)	1.49(1.30-1.72)
rs10489265	171.503	*59.6Kb 5' of TNFSF4*	G	0.314/0.251	0.284/0.234	0.386/0.287	T	30.8	50.4	34.3	$2.88x10^{-8}$	$1.25x10^{-12}$	$4.81x10^{-9}$	1.37(1.22-1.53)	1.3(1.21-1.40	1.56(1.35-1.82)
rs2022449	171.505	*62.3Kb 5' of TNFSF4*	T	0.321/0.252			G	36.2			$1.76x10^{-9}$			1.40(1.26-1.56)		
rs844663	171.510	*67.1Kb 5' of TNFSF4*	C	0.428/0.354	0.334/0.278	0.436/0.332	T	35.7	53.8	36.2	$2.34x10^{-9}$	$2.26x10^{-13}$	$1.78x10^{-9}$	1.36(1.23-1.51)	1.29(1.21-1.38)	1.56(1.35-1.80)
rs12049190	171.514	*57.8Kb ' of TNFSF4*	A	0.332/0.262	0.386/0.346	0.45/0.358	T	34.5	26.8	21.6	$4.37x10^{-9}$	$2.2x10^{-7}$	$3.28x10^{-6}$	1.41(1.27-1.55)	1.19(1.11-1.27)	1.46(1.25-1.71)
rs12750070	171.516	*73.3Kb 5' of TNFSF4*	T	0.326/0.251	0.385-0.344	0.489/0.362	C	33.5	28.6	35.9	$7.25x10^{-9}$	$8.75x10^{-8}$	$2.05x10^{-9}$	1.44(1.28-1.64)	1.2(1.12-1.28)	1.55(1.34-1.79)
rs12405577	171.517	*73.5Kb 5' of TNFSF4*	T	0.319/0.256	0.287/0.238	0.398/0.292	C	26.1	47.3	35.6	$3.31x10^{-7}$	$6.14x10^{-12}$	$2.45x10^{-9}$	1.34(1.20-1.50)	1.2991.2-1.39)	1.60(1.37-1.87)
rs10912580	171.523	*80.1Kb 5' of TNFSF4*	G	0.315/0.251	0.286/0.236	0.397-0.297	A	31.7	49.2	35.5	$1.81x10^{-8}$	$2.3x10^{-12}$	$2.55x10^{-9}$	1.37(1.23-1.53)	1.29(1.20-1.39)	1.56(1.35-1.81)
rs4916319	171.533	*90.1Kb 5' of TNFSF4*	G	0.405/0.336	0.447/0.431		A	31.7	3.7		$1.79x10^{-8}$	$5x10^{-2}$		1.34(1.21-1.49)	1.06(0.99-1.13)	
rs4916213	171.535	*92.4Kb 5' of TNFSF4*	T	0.403/0.341	0.429/0.405		C	26.6	7.38		$2.56x10^{-7}$	$6.7x10^{-3}$		1.31(1.18-1.45)	1.1(1.02-1.17)	
rs1342032	171.537	*94.2Kb 5' of TNFSF4*	T	0.398/0.339	0.408/0.390	0.473/0.399	G	23.3	4.82	17.4	$1.38x10^{-6}$	$2.8x10^{-2}$	$2.99x10^{-5}$	1.29(1.16-1.43)	1.08(1.0-1.15)	1.35(1.17-1.56)

A1**= minor allele code, A2= major allele code, F_A/F_U allele frequency in affected/unaffected, As, East Asian, Eur European, Hisp Hispanic, *P* unadjusted p- value, [a]P_p adjusted p- value, by 5000 permutations to give $Pp<10^{-4}$, CHQ chi-squared, OR odds ratio(95% CIn) confidence interval), aLocation anchored to most common transcript we found by RACE-PCR, which validated Ensembl bioinformatic data**

Table 4.3 Associated *TNFSF4* markers in African- Americans, Gullah and combined AA-Gullah

Marker	[a.] Location	A1/A2	African-American (1529 cases, 2048 controls)			Gullah (151 cases, 122 controls)			Combined (AA-Gullah) (1680 cases,2170 controls)		
			F_A/F_U	*p-value*	ORX (95% CI)	F_A/F_U	*p-value*	ORX(95% CI)	F_A/F_U	*p-value*	ORX(95% CI)
rs7553711	-20.96Kb 3' *TNFSF4*	T/T	0.29/0.24	4.01×10^{-5}	1.26(1.10-1.44)	0.23/0.16	0.06	1.51(0.98-2.34)	0.29/0.24	5.34×10^{-5}	1.26(1.10-1.44)
rs6676785	7.85Kb 3' *TNFSF4*	A/G	0.29/0.24	3.49×10^{-5}	1.26(1.10-1.44)	0.23/0.16	0.06	1.51(0.98-2.34)	0.20/0.24	6.23×10^{-5}	1.26(1.1-1.44)
rs10127728	1.72Kb 3' *TNFSF4*	G/T	0.51/0.46	1.30×10^{-4}	1.20(1.09-1.32)	0.46/0.39	0.09	1.34(0.95-1.89)	0.50/0.46	5.68×10^{-5}	1.21(1.1-1.32)
rs6691738	0.84Kb 3' *TNFSF4*	T/G	0.29/0.25	6.12×10^{-5}	1.25(1.09-1.43)	0.24/0.18	0.11	1.40(0.21-0.92)	0.29/0.25	$6.64 \text{x-} 10^{-5}$	1.25(1.09-1.44)
rs3861950	*TNFSF4*	T/C	0.23/0.20	1.20×10^{-4}	1.20(1.07-1.35)	0.15/0.11	0.12	1.50(0.89-2.50)	0.22/0.19	2.23×10^{-5}	1.19(1.07-1.33)
rs10798265	*TNFSF4*	T/C	0.37/0.32	5.17×10^{-5}	0.81(0.74-0.90)	0.42/0.4	0.62	1.09(0.77-1.54)	0.370/0.34	3.05×10^{-4}	0.84(0.76-0.92)
rs1234314	0.92Kb 5' *TNFSF4*	G/C	0.32/0.28	7.27×10^{-5}	1.22 (1.10-1.35)	0.34/0.24	0.01	1.62(1.10-2.37)	0.33/0.28	8.84×10^{-5}	1.25(1.13-1.38)
rs1234317	**11.3Kb 5' *TNFSF4***	**T/G**	**0.11/0.08**	**$8.15 \text{x} 10^{-5}$**	**1.34 (1.14-1.58)**	**0.09/0.05**	**$6.76 \text{x} 10^{-3}$**	**2.57(1.27-5.21)**	**0.11/0.08**	**1.45×10^{-5}**	**1.39(1.19-1.62)**
rs2205960	**15Kb 5' *TNFSF4***	**T/G**	**0.07/0.04**	**$8.29 \text{x} 10^{-4}$**	**1.38 (1.13-1.69)**	**0.09/0.02**	**$6.51 \text{x} 10^{-4}$**	**4.69(1.78-12.38)**	**0.07/0.05**	**3.79×10^{-5}**	**1.49(1.22-1.79)**
rs12039904	35.1 Kb 5' *TNFSF4*	**T/C**	**0.06/0.04**	**3.44×10^{-3}**	**1.36(1.11-1.69)**	**0.06/0.02**	**0.01**	**3.80(1.27-11.39)**	**0.06/0.04**	**7.60×10^{-4}**	**1.43(1.16-1.75)**
rs10912580	80.1Kb 5' *TNFSF4*	G/A	0.14/0.12	$9.24 \text{x} 10^{-4}$	1.25 (1.08-1.44)	0.13/0.07	0.01	2.18(1.19-3.99)	0.14/0.11	2.20×10^{-4}	1.28(1.11-1.46)

A1/A2- minor allele code/major allele code; F_A/F_U - allele frequency in affected/unaffected, CHISQ- chi square, ORX(95% CI)- Odds ratio (95% confidence interval). After Q.C filtering African-American(1510,2022), Gullah(152,122) and both (1680,2170)
[a.] Location anchored to our transcript data found by RACE-PCR (S. Guerra, KCL, UK) and in the Ensembl genome browser.

Table 4.4 Conditional regression results for 5´TNFSF4 variants in four SLE-control groups

				AA+Gullah p-value			Asian p-value			European p-value			Amerindian/Hispanic p-value		
Marker	A1	A2	Coordinate	rs1234314	rs1234317	rs2205960	rs1234314	rs1234317	rs2205960	rs1234314	rs1234317	rs2205960	rs1234314	rs1234317	rs2205960
rs1234314	*G*	*C*	*173177392*	*-1*	*0.088*	*0.103*	*-1*	*4×10^{-4}*	*0.013*	*-1*	*0.017*	*0.038*	*-1*	*0.005*	*0.009*
rs1234317	*T*	*C*	*173187775*	*0.004*	*-1*	*0.084*	*0.190*	*-1*	*0.024*	*$8.98 \text{x} 10^{-5}$*	*-1*	*0.57*	*0.007*	*-1*	*0.94*
rs2205960	*T*	*G*	*173191475*	*0.001*	*0.224*	*-1*	*0.008*	*9×10^{-4}*	*-1*	*$4.02 \text{x} 10^{-8}$*	*4.88×10^{-5}*	*-1*	*3.56×10^{-4}*	***0.010***	*-1*

Conditional analyses in SNPTESTv2 Case Control. Continuous covariate within a clustering framework. P-values selected using additional model and a frequentist paradigm

4.6 Conditional regression analysis of 5′ risk-haplotype associated single-markers

As expected, the 5′ upstream association data suggested pairwise LD between markers is weakest in African-Americans and strongest in Asians, and this correlated with haplotype length. In order to establish whether the signals identified by our trans ancestral fine-mapping study represent causal variants, independent risk factors, or if they are surrogate markers strongly correlated with causal variants, the association data from each population was conditioned with the marker which represented the best evidence of association **(Table 4.4**). In all populations, rs2205960-T*,* a risk-haplotype tag SNP with highest meta-analysis *p*-value and effect size, was associated with SLE after 5000 permutations, a similar trend was found for the adjacent marker rs1234317. In African-Americans, the single most associated risk marker was rs1234314, a marker adjacent to the *TNFSF4* promoter associated with SLE in all groups, and so rs2205960, rs1234317 and rs1234314 were included in a step-wise conditional regression analysis.

Conditioning on the presence of rs1234317 or rs2205960*,* found residual association at the intron1 marker rs16845607 in the Hispanic/Amerindian group, confirming association at this marker as an independent signal unique to this population. Association of all other intron 1 markers tested across all groups was lost, confirming these as secondary to *5′* risk associations. Conditional analysis used either rs1234317 or rs2205960 as a covariate: The signal at rs1234317 was lost after conditioning for rs2205960, and this was consistent for all populations tested (**Table 4.4**). On performing the reverse analysis, conditioning on the presence of rs1234317, there was residual association of rs2205960 in the East Asian, European and Hispanic cohorts ($P=0.024_{AS}$, $P<10^{-4}{}_{EUR}$, $P=0.015_{His}$).

4.7 Modelling the pattern of inheritance

For each population, in genotype-based analyses, the model that best fit the 5′ association of *TNFSF4* with SLE was the additive model.

4.8 Association of intragenic *TNFSF4* Single Markers

Examining the genetic association between SNPs within the *TNFSF4* gene and SLE, identified association of the intron 1 variant rs16845607-Awith Hispanic and Amerindian SLE ($P=9.17 \times 10^{-9}$, $P_p<10^{-4,}$ OR=2.06 (95%CI 1.6-2.64)) (**Table 4.5**). Association of rs1234313-G, within intron1, with SLE in Asians ($P=6.17 \times 10^{-7}$, $P_p<10^{-4}$, OR=1.32 (95%CI 1.18-1.46)), and Europeans ($P=4.32\times10^{-5}$, $P_p<10^{-3}$, OR=1.17 (95% CI 1.08-1.24)) (**Table 4.2, Figure 4.6**) was also identified. In both cohorts rs1234313-G is partitioned from other associated SNPs by recombination at the *TNFSF4*-5′ boundary. However, correlation coefficient r^2 values between this marker and risk-associated 5′ variants suggested strong correlation. Under representation of rs10798265-A in African-American SLE ($P=4.09\times10^{-5}$, $P_p<10^{-3}$, 0.81(95%CI 0.74-0.9)) was also identified. There were additional modest association signals ($P<10^{-4}$) from a series of SNPs located at the *TNFSF4*-3′UTR boundary in the same cohort (**Figure 4.6**).

4.9 Fixed-effects meta-analysis

A logistic regression model fitted with an interaction term (effect) in the R statistical package was used to investigate cross-study heterogeneity. P-values for individual associated SNPs were generated using the likelihood-ratio test. I found no evidence of cross-study heterogeneity for key haplotype-tagging common variants which span the locus: rs1234317, rs2205960, rs12039904, and rs10912580 were selected as representatives for this test. These data indicate the observed effects to differ by chance. P-values against the homogeneity of odds ratios are found in Appendix E.

The fixed-effects meta-analysis method combined the association results for East Asians, Europeans and Hispanics and African-Americans to more powerfully estimate the true effect size (**Table 4.6**). The sample size for each set of data after QC was used in the meta-analysis. Genetic complexity at *TNFSF4* in terms of pairwise SNP correlations and haplotype structure, and the strong association of identical alleles in the 5′*TNFSF4* region allowed use of a single set of assumptions and conditions in these diverse populations. The average effect size across all datasets was computed using inverse variance weighting of each study.

SNPs were organised into two categories (*TNFSF4* gene or 5′ region) and are highly correlated with one another ($r^2 > 0.7$) within each group. By combining three independent datasets we find the 5′ association of *TNFSF4* with SLE is greatly reinforced. Rs2205960-T, the most associated allele in Europeans and Hispanics, ($P = 7.1 \times 10^{-32}$, OR = 1.63, 95% CI = 1.58-1.79), and rs1234317-T ($P = 3.0 \times 10^{-30}$, OR = 1.62, 95% CI = 1.39-1.88) have the strongest combined associations with disease, these markers are adjacent to one another, separated by a 3kb section of chromosome 1.

Table 4.5 **Markers genotyped across intron1 of the *TNFSF4* gene in an Amerindian and Hispanic SLE-control group and combined association data**

Marker	[a]Location	A1/A2	Amerindian				Hispanic				Combined			
			F_A/F_U	CHISQ	p-value	ORX (95% CI)	F_A/F_U	CHISQ	p-value	ORX(95% CI)	F_A/F_U	CHISQ	p-value	ORX(95% CI)
rs13343108	171434853	C/T	0.201/0.229	1.35	0.2462	0.84(0.63-1.13)	0.422/0.384	3.03	0.082	1.17(0.98-1.4)	0.427/0.394	3.523	0.061	1.14(0.99-1.31)
rs7525284	171435020	A/G	0.272/0.29	0.46	0.4956	0.91(0.13-0.70)	0.203/0.259	9.32	0.002	0.73(0.59-0.89)	0.202/0.246	9.116	0.003	0.78(0.66-0.92)
rs10489267	171436775	A/C	0.119/0.101	0.95	0.3305	1.21(0.83-1.76)	0.265/0.293	2.01	0.156	0.87(0.72-1.06)	0.266/0.292	2.666	0.102	0.88(0.76-1.03)
rs11811856	171438296	G/C	0.272/0.292	0.56	0.4551	0.90(0.67-1.18)	0.115/0.116	0.01	0.96	0.99(0.75-1.31)	0.116/0.109	0.3956	0.529	1.07(0.86-1.33)
rs11811856	171438296	G/C	0.272/0.292	0.558	0.4551	0.90(0.67-1.18)	0.266/0.293	1.82	0.177	0.87(0.72-1.06)	0.268/0.293	2.615	0.106	0.88(0.76-1.03)
rs16845607	**171440240**	**A/G**	**0.152/0.073**	**14.9**	**1.1×10^{-4}**	**2.27(1.48-3.46)**	**0.145/0.080**	**17.0**	**3.6×10^{-5}**	**1.95(1.41-2.69)**	**0.147/0.077**	**33.01**	**9.17×10^{-9}**	**2.06(1.60-2.64)**
rs3850641	171442455	G/A	0.176/0.188	0.9717	0.3243	0.86(0.63-1.17)	0.174/0.137	5.07	0.024	1.33(1.04-1.71)	0.173/0.161	0.7936	0.373	1.09(0.90-1.31)

A1/A2- associated allele code/major allele code; F_A/F_U - allele frequency in affected/unaffected, CHISQ- chi square, ORX(95% CI)- Odds ratio (95% confidence interval).

After Q.C filtering, Amerindian (274 cases, 336 controls), Hispanic (959 cases, 336 controls) [a]Location anchored to our transcript data found by 5' RACE-PCR and in the Ensembl genome browser.

Table 4.6 Fixed effects meta-analysis of the association p-value for *TNFSF4* SNPs

MARKER NAME	COORDINATE (Mb)	A1	FREQ1	SE	ZSCORE	p-value	p-value (5000 PERM)
rs2205960	173.191	T	0.252	0.09	12.051	7.10×10^{-32}	7.12×10^{-12}
rs1234317	173.188	T	0.279	0.030	11.174	3.00×10^{-30}	3.06×10^{-11}
rs1234314	173.177	G	0.433	0.059	11.174	5.46×10^{-29}	6.06×10^{-11}
rs844663	173.244	C	0.372	0.049	11.011	3.37×10^{-28}	4.18×10^{-11}
rs12039904	173.212	T	0.310	0.039	10.888	1.32×10^{-27}	5.14×10^{-11}
rs844649	173.224	C	0.372	0.048	10.721	8.15×10^{-27}	7.06×10^{-11}
rs10912580	173.256	G	0.310	0.040	10.574	3.91×10^{-26}	7.14×10^{-11}
rs10489265	173.236	G	0.307	0.037	10.559	4.62×10^{-26}	8.08×10^{-11}
rs844654	173.233	T	0.467	0.022	10.332	5.05×10^{-25}	6.06×10^{-10}
rs2795288	173.214	A	0.465	0.022	10.286	8.16×10^{-25}	6.06×10^{-10}
rs12405577	173.250	T	0.313	0.040	10.214	1.71×10^{-24}	6.06×10^{-10}
rs1234315	173.178	T	0.482	0.057	9.966	2.14×10^{-23}	6.06×10^{-10}
rs1012507	173.219	T	0.386	0.041	9.446	3.51×10^{-21}	6.06×10^{-10}
rs12750070	173.250	T	0.385	0.044	9.374	6.99×10^{-21}	6.06×10^{-10}
rs844651	173.225	G	0.441	0.028	8.985	2.60×10^{-19}	6.06×10^{-10}
rs844644	173.209	C	0.431	0.049	8.978	2.76×10^{-19}	2.48×10^{-09}
rs12049190	173.247	A	0.384	0.036	8.744	2.25×10^{-18}	6.06×10^{-10}
rs1234313	173.166	G	0.320	0.055	6.354	2.10×10^{-10}	9.27×10^{-6}
rs1342032	173.271	T	0.416	0.025	5.73	1.00×10^{-08}	2.58×10^{-6}
rs4916319	173.267	G	0.447	0.033	5.515	3.49×10^{-08}	8.01×10^{-5}

The first three columns list SNP characteristics, the next six columns list meta-analysis results including allele frequencies (FREQ1)and two-tailed p-value for nominal and adjusted (5000 permutations)SNP associations

4.10 Bifurcation of *TNFSF4* haplotypes

Haplotypes significantly associated with risk of disease were identified for each population. To better visualise the breakdown of LD of associated haplotypes, I constructed bifurcation diagrams from phased genotypes for each cohort tested (**Figure 4.7**). The plots illustrate the breakdown of linkage disequilibrium (LD) at increasing distances in both directions from rs1234314, the most proximal genotyped SNP located at the *TNFSF4* gene-5′ boundary which was used as the core variant in the figure (labelled, circular core from which haplotype branches). The location of rs1234317 and rs2205960*,* which were best-associated in the fixed effects meta-analysis, are also marked onto the diagram. The thickness of the line in each plot corresponded to the number of samples with the haplotype, branches indicate breakdown of LD. For the risk haplotype*,* the lines were most robust in East Asians (**Figure 4.7*A*, risk),** followed by Hispanics and Europeans, and least robust in African-Americans. Branch junctions depicting breakdown of LD of the risk haplotype were coincident with the section of the *TNFSF4* locus encompassing rs1234317 and rs2205960.

The non-risk haplotype retained its thickness with distance from the core in the AA group, indicating long-range homozygosity (**Figure 4.7*B*, non-risk)**. Contrasting the recombination rate in risk and non-risk haplotype homozygotes, I found increased recombination in the risk individuals (**Figure 4.5),** data which support the visualized breakdown of haplotypes by these bifurcation plots.

Figure 4.7

A. $TNFSF4_{risk}$

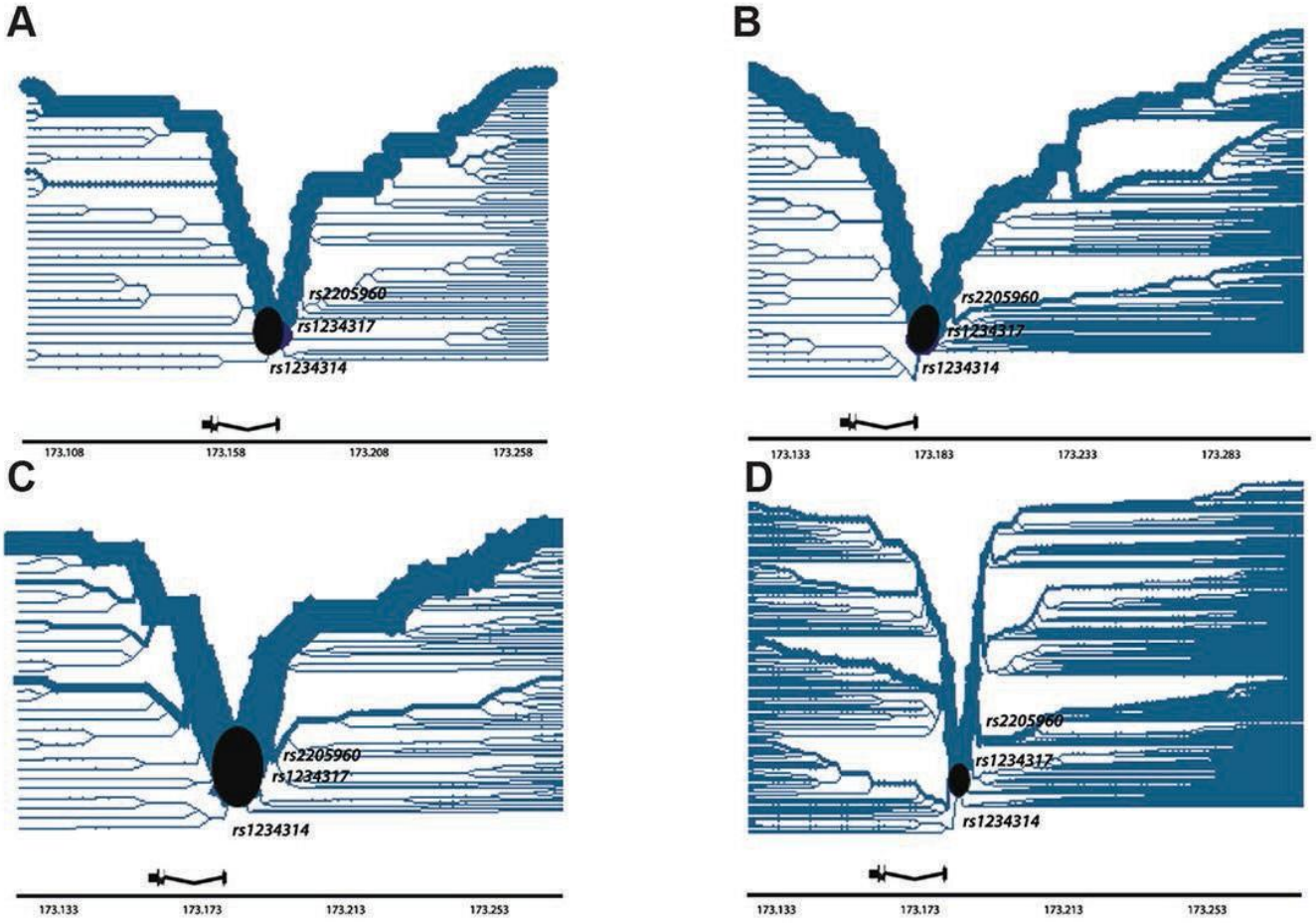


Figure 4.7

B. TNFSF4$_{non\text{-}risk}$

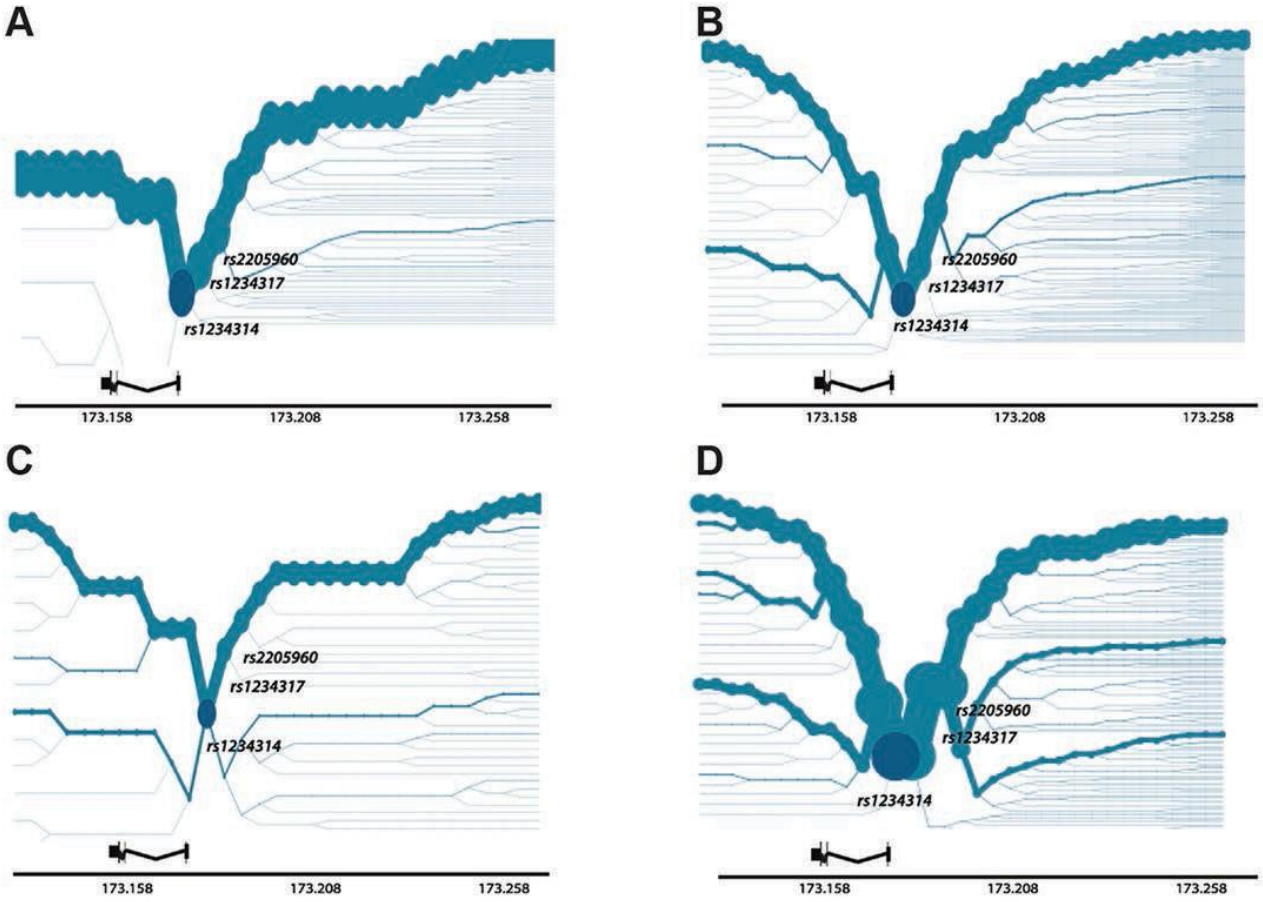


Figure 4.7 Haplotype bifurcation diagrams for A. *TNFSF4*$_{\text{risk}}$ and B. *TNFSF4*$_{\text{non-risk}}$ haplotype for four populations. Plots were constructed using phased haplotypes to illustrate breakdown of LD at increasing distances from a core proximal *TNFSF4* SNP (rs1234314*)* and are approximately to scale. rs1234314 is located at the *TNFSF4* gene-5′ boundary (black circle) this SNP is the most proximal 5′ marker associated with disease in all four populations. Gene location is depicted to scale by the black arrow below the plot, black ticks below each plot show the location of rs1234317 and rs2205960, the best-associated markers from the meta-analysis, additionally rs16845607 is marked under the Amerindian/Hispanic plot.

4.11 Conservation of *TNFSF4* haplotype structure across populations

There is a bipartite structure to the haplotype blocks at the *TNFSF4* locus in all but the African-American cohort (**Figure 4.8**). Significantly associated haplotypes were found in each population (**Figure 4.9**). Low recombination and similar location of hotspots at the *TNFSF4*-5′ boundary in East Asians, Europeans, and Hispanics allowed the construction of near-identical risk and non-risk haplotypes (designated *TNFSF4*$_{risk}$ and *TNFSF4*$_{non\text{-}risk}$, respectively) which extended at least 100kb into the *TNFSF4* 5′ region (**Figure 4.9**). Multiple associated risk alleles uniquely tag *TNFSF4*$_{risk,}$ overrepresented in SLE individuals in each population, whilst *TNFSF4*$_{non\text{-}risk}$ is the most frequent haplotype for all cohorts tested but underrepresented in SLE individuals.

Haplotype association data for *TNFSF4* $_{risk}$ and *TNFSF4*$_{non\text{-}risk}$ are presented in **Figure 4.9**. A shorter subdivision of the larger risk haplotype was associated with African-American SLE; rs1234317-T and rs2205960-T tags this 15.6kb AGTTCTT risk haplotype ($P=8.39 \times 10^{-5}$, OR=1.52*)*. This is anchored to the proximal 5′ region. Low frequency haplotypes (<4%) in all populations tested were not associated with disease after correction for multiple testing.

4.12 Intragenic haplotype confers risk uniquely in Amerindians and Hispanics

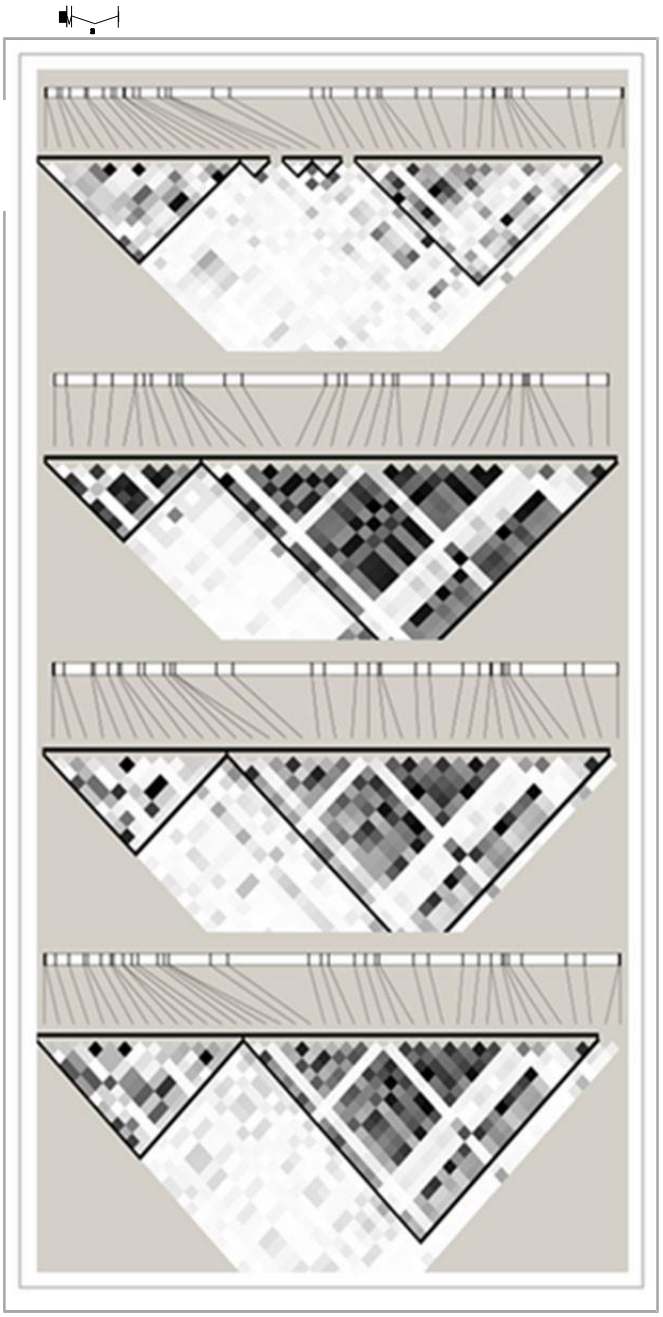
Examining haplotypes within the *TNFSF4* gene in relation to SLE identified association of a 22.5kb high-frequency haplotype which conferred risk with strong effect in the Hispanic and Amerindian cohorts ($P=9.17 \times 10^{-9}$, $Pp<10^{-4}$, OR=2.06) (**Figure 4.10**). This haplotype is uniquely tagged by rs16845607-A. a SNP located in intron 1 of the TNFSF4 gene which is monomorphic in African and European populations. Including a covariate for the rs2205960 variant, there was residual association at rs16845607-A in this population, suggesting it as an independent causal contributor to disease risk in this population.

Importantly, an intragenic common haplotype identical at all alleles but with rs16845607-G was found at reduced frequency in SLE cases ($P=0.04$, $P_p=0.22$, OR=0.84).

4.13 Conditional regression analysis of AA haplotypes

Testing the common haplotypes in the *TNFSF4* proximal promoter region for association with SLE revealed two haplotypes after permutation testing (**Table 4.7**): $_{AA}TNFSF4_{risk}$ ($P=8.39\text{x}10^{-5}$, $P_p<10^{-3}$, OR=1.52) and the most frequent haplotype, $_{AA}TNFSF4_{non\text{-}risk}$ ($P= 6.92\text{x}10^{-6}$, $P_p=1 \times 10^{-4}$, OR=0.82 (0.75-0.9)). Conditional analysis was undertaken in order to establish whether the two most associated haplotypes identified are likely to represent independent risk and protective factors or whether the association is confined to $_{AA}TNFSF4_{risk}$ (in this instance, $_{AA}TNFSF4_{non\text{-}risk}$ would only be associated because of the corresponding decrease in risk alleles). A covariate for the presence of $_{AA}TNFSF4_{risk}$ was included in the logistic regression model: There was residual association owing to $_{AA}TNFSF4_{non\text{-}risk}$. To explain further investigate the residual association signal at $_{AA}TNFSF4_{non\text{-}risk}$, I further conditioned on a low frequency haplotype (haplotype 1) which was weakly associated with SLE ($P= 0.03$, $P_p=0.151$, OR= 1.53 (1.04-2.25)) and found it not to explain the residual association attributed to $_{AA}TNFSF4_{non\text{-}risk}$.

Figure 4.8 Comparison of LD plots across 200kb of chromosome 1q25.1



This section of chromosome 1 encompasses the *TNFSF4* gene and upstream region. Pairwise LD relationships are defined using the custom algorithm in Haploview 4.2. Pairwise LD was used to compare the 44 successfully genotyped SNPs common to all cohorts, post QC. The pair-wise correlations between *TNFSF4* markers are illustrated in these plots by the correlation coefficient r^2 (where $r^2 = 0$ = no correlation, white; $0 < r^2 < 1$, gradations of grey; $r^2 = 1$ = complete correlation, black). The *TNFSF4* gene is positioned above the plots relative to haplotype blocks (black triangles) and grey ticks indicate SNP locations to scale.

Reversing the analysis by conditioning on presence of $_{AA}TNFSF4_{non\text{-}risk}$ also found residual association owing to $_{AA}TNFSF4_{risk}$. These analyses demonstrated the observed signals in the *TNFSF4* promoter region to independently confer risk and protection against SLE.

Table 4.7 Conditional regression of *TNFSF4* promoter haplotypes, AA SLE-control cohort

Haplotype ID	Seq	[a]Freq	Haplotypic association			Conditional regression analysis		
			ORX (95% CI)	*P*	P_p	$_{AA}TNFSF4_{risk}$ *pValue*	$_{AA}TNFSF4_{risk}$ +H1 P_u	$_{AA}TNFSF4_{OR<1}$ P_u
$_{AA}TNFSF4_{risk}$	AGTTCTT	0.052	1.52 (1.21-1.76)	$8.4x10^{-5}$	$1x10^{-3}$	NA	NA	0.011
Haplotype 1(H1)	GGTTCCG	0.014	1.53 (1.04-2.25)	0.031	0.151	0.022	NA	0.100
Haplotype 2 (H2)	GGCTCCG	0.043	1.01 (0.90-1.14)	0.817	1	0.404	0.307	0.027
Haplotype 3 (H3)	ACTCCCG	0.200	1.02 (0.87-1.27)	0.466	0.966	0.328	0.270	0.945
Haplotype 4 (H4)	AGTTCCG	0.142	1.10 (0.96-1.25)	0.171	0.644	0.071	0.062	0.643
Haplotype 5 (H5)	AGTTCTG	0.036	1.23 (1.02-1.50)	0.090	0.486	0.089	0.075	0.723
$_{AA}TNFSF4_{non\text{-}risk}$	ACTTCCG	0.496	0.85 (0.75-0.94)	$6.92x10^{-6}$	$1x10^{-4}$	0.001	0.004	NA

ORX, odds ratio, *P*, uncorrected p-value, P_p, adjusted p-value after 5000 permutations [a] Total freq of haplotypes 0.98, we have excluded rare (0.01 or less) haplotypes from this analysis

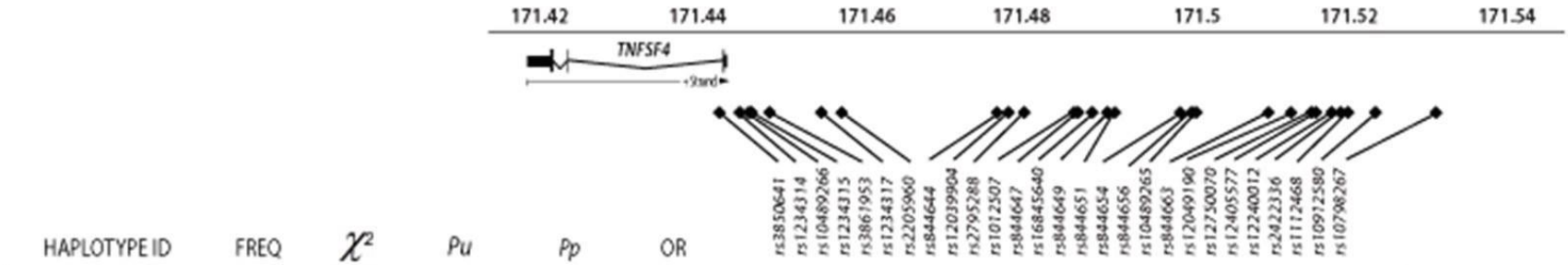
4.14 Neutral haplotypes

Informative neutral haplotypes supported the identification of causal SNPs by conditional analysis. Recombinant haplotypes in European and AA cohorts were identified: These are composites of the non-risk and risk haplotypes. The recombination point is between rs1234317 and rs2205960, the best-associated variants by meta-analysis. These recombinants are presented under the *Odds Ratio~1(NEUTRAL)* subheading in **Figure 4.9.** The European neutral haplotype $_{EUR}TNFSF4_{OR\sim1}$ extends 100kb into the 5′ region. $_{EUR}TNFSF4_{OR\sim1}$ has a proximal section of the risk haplotype, tagged by rs1234317-T and a distal section of the non-risk haplotype tagged by rs2205960-G: This haplotype was not associated with risk of disease. In AAs, the neutral $_{AA}TNFSF4_{OR\sim1*}$ haplotype is similarly structured to $_{EUR}TNFSF4_{OR\sim1}$; the proximal section of the risk haplotype is in combination with a single non-risk variant rs2205960-G. $_{AA}TNFSF4_{OR\sim1*}$ was not

associated with SLE. These data support the single marker conditional regression results which suggest rs2205960-T but not rs1234317-T, to drive the risk association signal.

In Hispanics, a high frequency 23.58kb intron 1 haplotype, black-boxed and designated #1 in **Figure 4.10** is identical to the risk–associated haplotype (red-boxed and designated #4, **Figure 4.10)** but instead with the G allele of rs16845607. Haplotype #4 was not associated with risk of disease (P=0.04, P_p=0.22, OR=0.84). This supports rs16845607-A as a causal variant in SLE individuals with Amerindian ancestry.

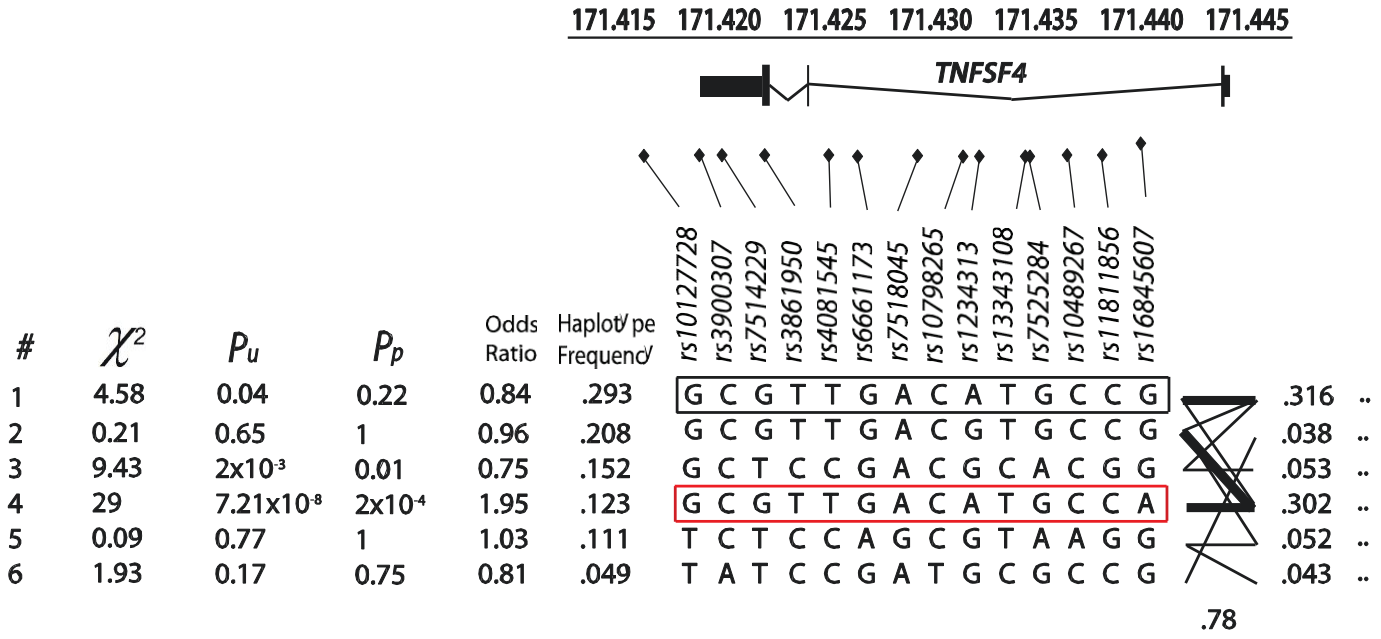
Figure 4.9 Fine-scale structural comparison of the *TNFSF4*$_{risk}$ and *TNFSF4*$_{non\text{-}risk}$ haplotypes and their association in four SLE case control populations



HAPLOTYPE ID	FREQ	χ^2	*Pu*	*Pp*	OR	
OR<1 (NON RISK)						
AS*TNFSF4*non-risk	0.520	28.1	2.59×10^{-6}	$<10^{-5}$	0.74	A C - T C C G A C T G T A T T A T T T T C C T A G A C
EUR*TNFSF4*non-risk	0.391	21.3	1.37×10^{-7}	$<10^{-5}$	0.80	A C T T - C G A C T G - A T T A - T T T C C T A G A
HIS*TNFSF4*non-risk	0.316	16.9	3.86×10^{-5}	8×10^{-4}	0.76	A C T T - C G A C T G - A T T A - T T T C T A G A
AA*TNFSF4*non-risk	0.491	15.5	3.34×10^{-5}	4×10^{-4}	0.82	A C T T C C
OR~1 (NEUTRAL)						
EUR*TNFSF4*OR~1	0.037	1.10	0.294	0.789	1.03	A G T T C T G A C T G - A T T A - T T T C C T A G A C
AA*TNFSF4*OR~1	0.142	2.70	0.274	0.824	1.07	A G T T C C G
AA*TNFSF4*OR~1*	0.0.36	3.60	0.090	0.486	1.23	A G T T C T G
OR>1 (RISK)						
AS*TNFSF4*risk	0.263	31.0	3.7×10^{-8}	$<10^{-5}$	1.34	A G - T C T T C T A T T A C G T T G C A T T T A G G C
EUR*TNFSF4*risk	0.199	48.2	2.30×10^{-13}	$<10^{-5}$	1.35	A G T T - T T C T A T - A C G T - G C A T T T A G G
HIS*TNFSF4*risk	0.302	34.6	4.02×10^{-9}	$<10^{-5}$	1.57	A G T T - T T C T A T - A C G T - G C A T T T A G G
AA*TNFSF4*risk	0.051	15.4	8.39×10^{-5}	1×10^{-3}	1.52	A G T T C T T

The black diamonds above SNPs give locations to scale. SNP IDs across the top of the haplotypes correspond to alleles listed for each population. Haplotype frequencies, uncorrected and permuted haplotypic association values, *P* and *Pp* respectively, and Odds Ratios (OR) are presented to the right of each haplotype. The extended haplotypes (>90kb) in East Asians (AS), Europeans (Eur) and Hispanics (His) cannot wholly be reconstructed in African-Americans (AA). Haplotype-tagging SNPs allow construction of the 11.9kb $_{AA}$*TNFSF4*$_{non\text{-}risk}$ and a 15.6kb subdivision of the risk haplotype ($_{AA}$*TNFSF4*$_{risk}$). Under the OR~1(neutral) subheading are European (Eur) and AA alternative recombinant haplotypes which differ from OR>1(risk) at rs1234317 and/or rs2205960. Haplotypes with rs2205960-G instead of rs2205960-T were not associated with risk of SLE in these populations.

Figure 4.10 Association results for intragenic *TNFSF4* haplotypes in Amerindian & Hispanic SLE-cohort



5′ RACE data generated by Sandra Guerra, King's College London, was used to anchor the 23.58kb *TNFSF4* gene to the scale above and black ticks below the gene indicate SNP location to scale. Haplotype # 4 (red-boxed) is uniquely tagged by rs16845607-A. This haplotype is associated with risk of SLE. Haplotype #1 (black-boxed) is identical to #4, but has rs16845607-G and is not associated with risk of disease (OR=0.84) Lines extending from the most proximal 5′ SNP (rs16845607) indicate degree of connection to corresponding promoter haplotypes, thin lines indicating 1% or more, and thick lines 10% connection or more. Approximately 20% of Hispanic individuals with the intragenic *TNFSF4* risk haplotype possess an extended 125kb haplotype encompassing the 5′ risk haplotype. Low frequency haplotypes are not depicted in this figure.

4.15 Sub-phenotype association analyses

Given *TNFSF4* surface expression on a range of cell types which control immune functionality, one might expect *TNFSF4* alleles to be associated with disease manifestations of SLE. Median (IQR) age at diagnosis, autoantibody production and renal disease were examined within SLE cases and against controls for each cohort. American College of Rheumatology (ACR) classification criteria were additionally examined in East Asians, Europeans and Hispanics. Phenotypic subsets of SLE cases are less heterogenous than SLE per se and so may enrich for risk variants with increased effect size or prove informative for causal mechanism. Clinical characteristics of SLE individuals sorted by population are presented with case-only and phenotype-control association results (**Table 4.8**).

The case-only and phenotype-control data presented here strongly reinforced the unique association of the intron 1 variant rs16845607-A in Hispanics. This marker was associated with leukopenia ($P=1.08 \times 10^{-8}$, $Pp<10^{-4,}$ OR=2.75 (95% CI 2.04-3.72)) and lymphopenia ($P=5.6 \times 10^{-12}$, $Pp<10^{-4,}$ OR=2.95 (95% CI 2.1-4.0)) with improved significance and strong effect size (**Table 4.8**). Conditioning on the presence of rs16845607 in the subset of Hispanic SLE individuals positive for leukopenia resulted in residual 5′ *TNFSF4* association ($P=7.7 \times 10^{-4}$, OR=1.47(95% CI 1.18-1.82), rs1234314). The same analyses in lymphopenia cases removed the association signal (most associated marker after conditioning on rs16845607, rs1234314, P=0.01, OR=1.23(95% CI 0.96-1.57)).

4.15.1 Association of TNFSF4 Markers with autoantibody production

Case-only analysis revealed association of *TNFSF4* risk variants with autoantibody production in AA, European and Hispanic SLE cohorts: Evidence of association of rs2205960-T with Anti-Sm autoantibodies in African-American cases ($P=5.1 \times 10^{-3}$, OR=1.57 (95% CI 1.14-2.16) was reinforced by testing this variant against controls ($P=6.67 \times 10^{-7}$, OR=1.91 (1.47-2.47)). This marker also segregated with anti-Sm autoantibodies in European case-only and phenotype-control analyses. In Europeans the adjacent variant rs1234317-T was associated with anti-Ro autoantibodies ($P = 9.5 \times 10^{-4}$, $P_p=0.01$, OR= 1.31(95% CI 1.12-1.54) and this was reinforced against controls ($P=9.5 \times 10^{-8}$, $P_p<10^{-3}$ OR=1.52 (1.3-1.76)). In African-Americans analyses of 5′ variants against controls improved the significance of risk-haplotype-tagging variants with anti-dsDNA autoantibodies (rs1234317-T, $P=5.36 \times 10^{-6}$, $P_p<10^{-3}$, OR=1.68 (95% CI 1.34-2.1.)). There was a trans-ancestral trend which suggested underrepresentation of *TNFSF4* intron 1 alleles with autoantibody production. (**Table4.8**).

Table 4.8 Association analysis of *TNFSF4* variants with SLE sub-phenotypes

	European			Hispanic			AA + Gullah		
	Presence/ Absence	Best Marker	*P*/ OR (95%CI)	Presence/ Absence	Best Marker	*P*/ OR (95%CI)	Presence/ Absence	Best Marker	*P*/ OR (95%CI)
Phenotype-case									
Age at diagnosis, median (IQR)	816/774	rs12405577	1.43x10^{-3}/0.78 (0.68-0.91)	139/138	rs1539259	1.44x10^{-3}/ 0.57(0.41-0.81)	337/ 269	rs844654	2.51 x 10^{-3}/0.70(0.5560.89)
Anti-dsDNA	1177/1904	rs12124768	3.5x10^{-5}/0.74(0.6-0.9)	515/522	rs12405577	3.8x10^{-4}/ 1.4(1.16-1.68)	752/830	rsS4250	3.53 x 10^{-3}/2.42(1.31-4.46)
Anti-Ro	**422/1742**	**rs1234317**	**9.5x10^{-4}/1.31 (1.12-1.54)**	**119/385**	**rs16845607**	**0.02/ 1.69(1.08-2.64)**	330/648	rs10127727	5.98x 10^{-3}/0.74(0.59-0.92)
Anti-Sm	**225/2461**	**rs2205960**	**0.05/1.23(1-1.52)**	204/772	rs12405577	3.1x10^{-3}/1.41(1.12-1.78)	**487/763**	**rs2205960**	**5.1x10^{-3}/1.57(1.14-2.16)**
Renal Disease	**1054/2020**	**rs2205960**	**2.87 x 10^{-4}/1.24 (1.1-1.4)**	439/449	rs16845607	0.03/1.3(1.02-1.77)	785, 784	rrrs7518045	1.77 x 10^{-3}/0.73(0.60-0.89)
Immunologic	2441/570	rs1234313	2.5 x 10^{-3}/0.81(0.7-0.93)	297/92	rs13343108	1.7x10^{-4}/0.52(0.36-0.73)	**1229/167**	**rs2205960**	**0.035/1.77(1.04-3.04)**
Leukopenia	684/1166	rs10798265	0.10/1.3(0.95-1.79)	103/597	**rs16845607**	**0.02/1.37(1.04-1.79)**	365/559	rs4916215	0.02/1.83(1.1-3.05)
Lymphopenia	682/1097	rs7553711	1.77x10^{-3}/0.76(0.63-0.90)	269/84	**rs16845607**	**9.5x10^{-3}/1.45(1.1-1.9)**	281/640	rs10489268	0.01/1.83(1.15-2.92)
Phenotype-control									
Age at diagnosis, median (IQR)	816/3580	rs1234317	6.7x10^{-4}/1.3(1.23-1.52)	139/615	rs2205960	1.89x10^{-3}/ 1.55(1.17-2.03)	337/2144	rs10489265	4.57 x 10^{-4}/1.39(1.11-1.74)
Anti-dsDNA	1177/3580	rs2205960	6.5x10^{-9}/1.37(1.23-1.52)	515/615	rs12039904	6.12x10^{-10}/ 1.74(1.46-2.07)	752/2144	rs2205960	1.59 x 10^{-6}/1.74(1.39-2.19)
Anti-Ro	**422/3580**	**rs1234317**	**9.5x10^{-8}/1.52 (1.3-1.76)**	**119/615**	**rs16845607**	**3.82x10^{-4}/ 2.15(1.4-3.3)**	330/2144	rs1234314	1.67 x 10^{-3}/1.32(1.11-1.58)
Anti-Sm	**225/3580**	**rs2205960**	**1.63x10^{-5} /1.58(128-1.94)**	204/615	rs2205960	2.4x10^{-7}/1.85(1.46-2.4)	**487/2144**	**rs2205960**	**6.67 x 10^{-7}/1.91(1.47-2.47)**
Renal Disease	**1054/3580**	**rs2205960**	**2.87x10^{-14} /1.53 (1.37-1.70)**	439/615	rs2205960	3.24x10^{-9}/1.75(1.45-2.1)	785, 2170	rs1234317	1.08 x 10^{-5}/1.52(1.26-1.83)
Immunologic	2441/3575	rs2205960	2.2x10^{-15}/1.4(1.29-1.53)	783/615	rs2205960	2.57x10^{-10}/1.69(1.43-1.98)	**1229, 2170**	**rs2205960**	**5.42 x 10^{-6}/1.59(1.30-1.95)**
Leukopenia	684/3575	rs2205960	5.5x10^{-9}/1.47(1.28-1.67)	**347/615**	**rs16845607**	**1.1x10^{-11}/2.75(2.04-2.37)**	365 ,2170	rs1234317	5.27x10^{-5}/1.64(1.29-2.09)
Lymphopenia	682/3575	rs2205960	9.4x10^{-12}/1.56(1.37-1.77)	**269/615**	**rs16845607**	**5.6x10^{-12}/2.94(2.14-4.05)**	281, 2170	rs1234314	9.27x10^{-5}/1.44(1.2-1.74)

For phenotypic definitions see web resources. Case and control numbers after filtering for QC.
Bold Indicates same marker is most-associated with phenotype in case-only and case-control analyses

(Hispanic P=1.7 x 10^{-4}, OR=0.52 (95% CI 0.36-0.73), European P=2.5 x 10^{-3}, OR= 0.81 (0.7-0.93) and East Asian P= 3.6 x 10^{-2}, OR=0.7 (95% CI 0.5-0.98)). Conditional regression analysis of the best-associated marker in each analysis removed all evidence of association.

4.15.2 Association of TNFSF4 markers with age at diagnosis

Examination within cases revealed association of distal 5′ *TNFSF4* alleles with age at diagnosis (IQR) across all cohorts apart from East Asians (**Table 4.8**). A trend for a reduced frequency of *TNFSF4* alleles in individuals with early age at SLE diagnosis was found in AA (P=9 x 10^{-4}, OR=0.69 (95% CI 0.56-0.86)), European (P=1.43 x 10^{-3}, OR=0.78(0.68-0.91)) and Hispanic (P=1.43 x 10^{-3}, OR=0.57(95% CI 0.41-0.81)) populations. These alleles are found in the distal 5′ region. A fixed effects meta-analysis found the best-associated marker associated with this phenotype to be rs844654 (P = 8.7 x 10^{-6}, Z score 4.45), located 60kb from the TNFSF4 gene. An additional four SNPs in this region illustrated the trend with age at diagnosis.

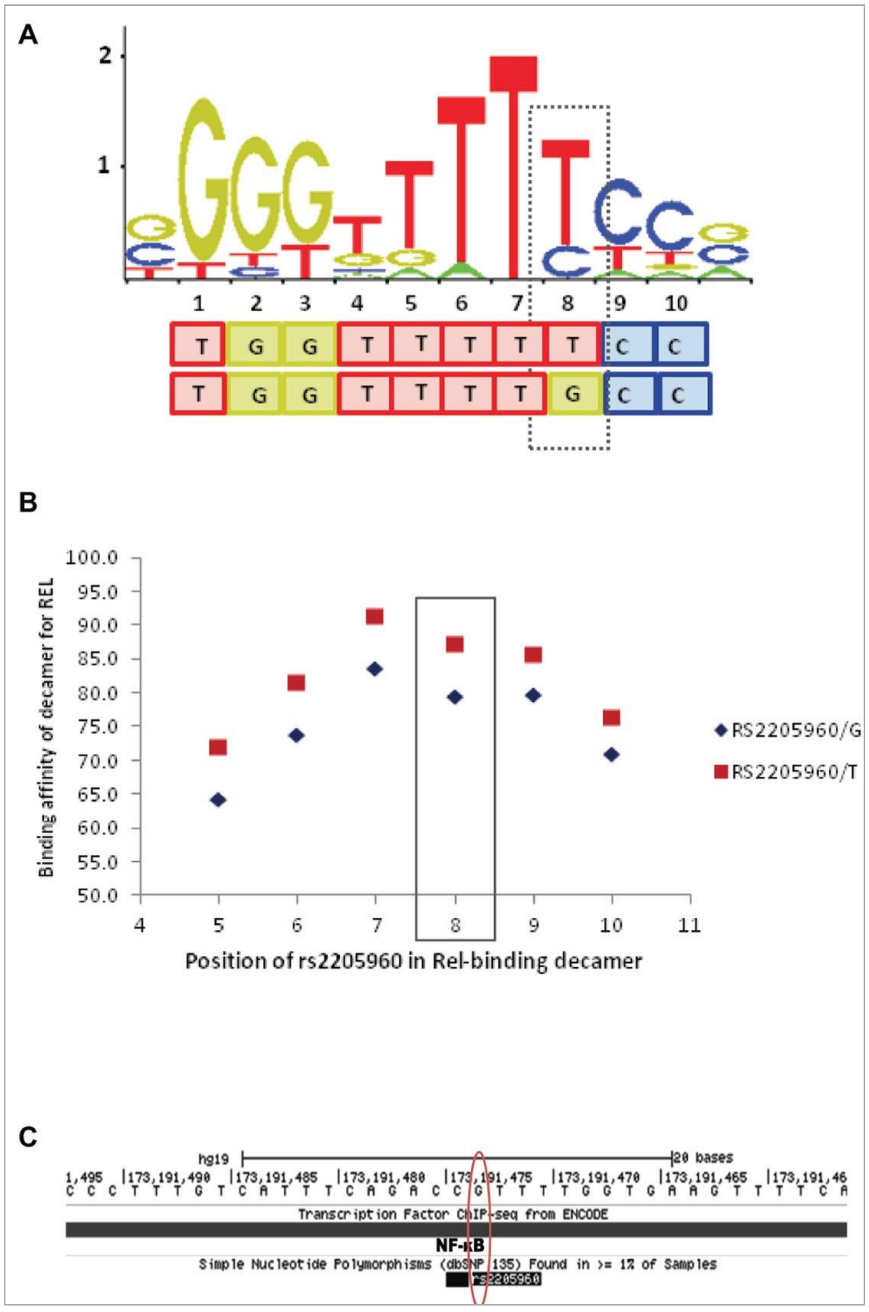
4.16 Bioinformatic analysis

The DNA sequence at rs2205960 was examined for interaction with regulatory proteins including transcription factors (TFs). A decameric DNA sequence including the rs2205960 variant was predicted to bind to the NF-κB p65 protein (RELA) with high confidence. Changing the associated (T) allele for the (G) rs2205960 allele was investigated for its impact on binding affinity of the motif for the p65 target protein. SELEX binding data and position weight matrix (PWM) profiles stored in the Jaspar core database (Portales-Casamar et al., 2010) were used to investigate the DNA sequence with rs2205960-T at the 8th nucleotide position: These data suggested a binding affinity of approximately 90% for NF-κB p65 **(Figure 4.11)**. Altering the allele to rs2205960-G decreased the binding affinity for NF-κB p65 by over 10%, but also highlighted degeneracy

of the motif (**Figure 4.11b**). Binding of NF-κB at rs2205960 has been confirmed by genome wide ChIP-seq experiments in EBV - B-cell lines as part of the ENCODE project (**Figure 4.11c**) (ENCODE Project Consortium, 2010). These ChIP-seq data indicate that signal intensity for NF-κB at rs2205960 in a heterozygous (G/T) cell-line (GM12878) was double that for a non-risk homozygote (G/G) cell line (GM06990). The sequence encompassing rs2205960 is conserved in primates but the degree of conservation lessensd for other Eutheria (**Figure 4.12**).

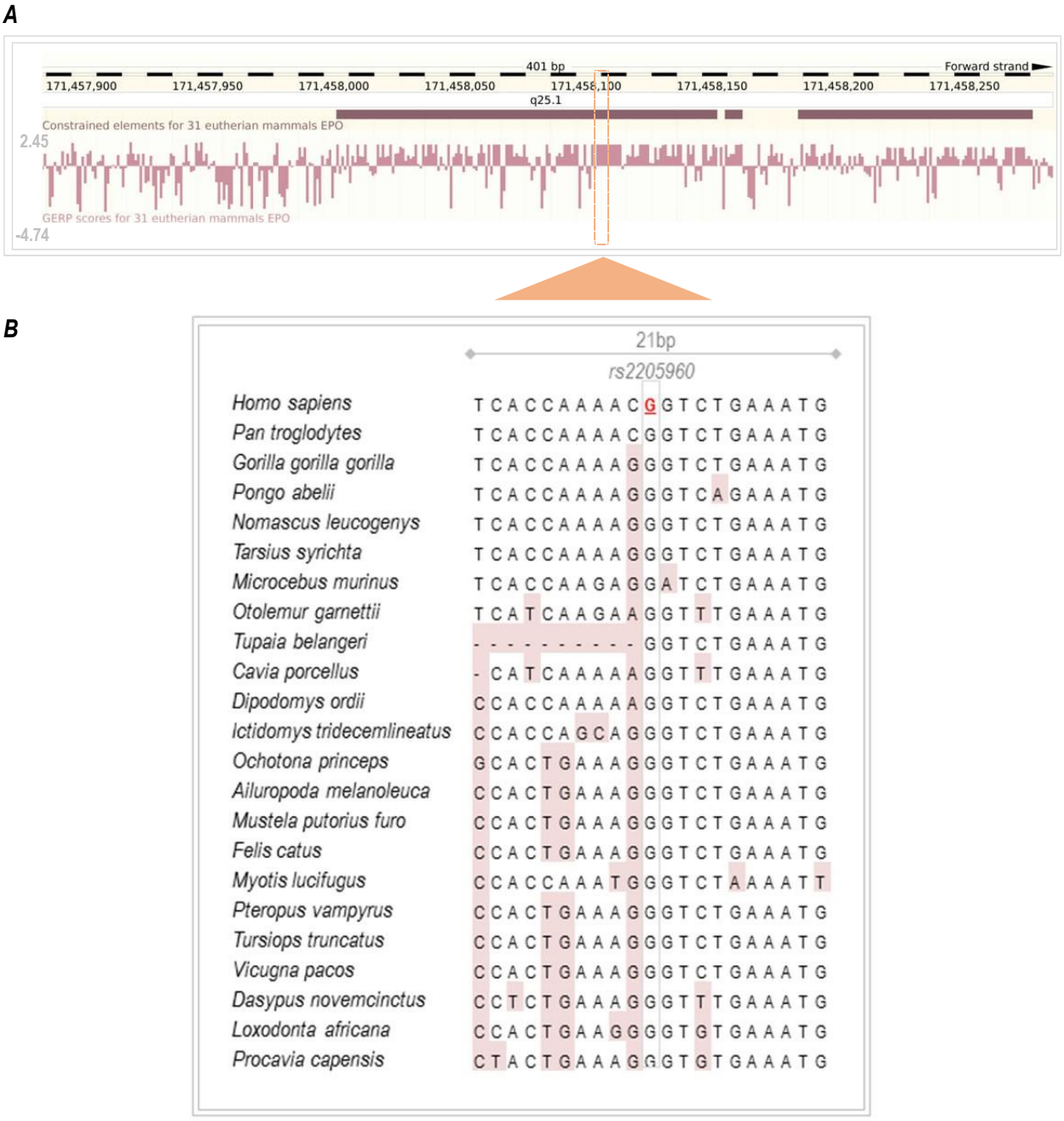
Annotation of the DNA sequence encompassing other haplotype-tagging *TNFSF4* variants was also investigated: The sequence encompassing rs1234314 was investigated for transcription factor binding. According to the conditional analysis, rs1234314 is the best-associated variant after including a covariate for the risk-association. Furthermore, the minor allele of this variant, under-represented in SLE individuals, tags the non-risk haplotype. Scanning the data held in the Ensembl genome browser revealed rs1234314 to be part of a 400bp segment which has repressed/ low activity in LCL cells but with no such activity in a T-cell line. The UCSC genome browser predicted rs1234314 to be located within a region associated with the H3K27Ac chromatin signature which is associated with active enhancers. Interrogating the sequence at rs1234314 with PWM binding data in the Jaspar core database gave no clear pattern of binding of either allele to the motif of a regulatory element.

Figure 4.11 SLE-associated rs2205960 predicted to be part of a decameric motif for NF-κB p65 (RELA)



A. Degeneracy within the core 10-base motif is illustrated at all positions apart from position 7 which is non-degenerate by the stacked letters at each position. The relative height of each letter is proportional to its over-enrichment in the motif. A dashed line is boxed around rs2205960-T, this SLE-associated allele is predicted to form the 8th nucleotide in the motif. Predictions were made using the non-degenerate set of matrix profiles in the Jaspar Core database. B. Altering the rs2205960 allele from -*T* to -*G* decreases the binding affinity for NF-κB p65 by over 10%. C. Binding of NF-κB *at* rs2205960, suggested by genome-wide ChIP-seq ENCODE data. Profiles were generated for lymphoblatoid cell lines and stored in the UCSC genome database

Figure 4.12 Phylogeny of for the sequence encompassing rs2205960



A.

B. Gerp (Cooper et al., 2005) was used to calculate conservation scores on the 31-way Eutherian species' multiple alignments for a 401bp section of chromosome 1q25.1 encompassing *rs2205960*. Positives scores indicate a greater degree of conservation. *B.* The phylogenetic context of the *rs2205960-T* polymorphism. The G allele of this variant (highlighted and underlined red) is depicted with 20bp flanking sequence. The sequence is aligned against that from eutherian mammal species. Pink highlighted nucleotides show differences in these species with respect to humans.

Examining the sequence with rs1234317-T against PWM binding data stored in the Jaspar Core database finds it completes a TATATT-binding motif and this motif was disrupted in the presence of rs1234317-C. The ENCODE project does not highlight binding of a TBP protein at this variant. Genome-wide ChIP-seq data from the ENCODE project has data for LCLs which carry the T allele of rs1234317. For LCLs carrying the risk (T) allele, there are currently no regulatory features annotated at this position in publically-available genome browsers.

4.17 Discussion

4.17.1 *Summary of findings: Recombination*

The European deCODE sex-averaged and female-only recombination maps (URL: http://www.decode.com/addendum/), are based on 15,257 and 8,850 directly observed recombinations, respectively. These maps, which have an effective resolution of 10kb, were compared to the HapMap phase III and 1000 Genomes population-averaged maps. Recombination differences at the *TNFSF4* locus were found between the aforementioned, established maps for the 5′ region where the associated variants are located. As such, I estimated background recombination rates in each population tested for association, using a Bayesian composite-likelihood method.

The deCODE map provides evidence of fine-scale differences in the recombination rate at *TNFSF4.* This map finds recombination at the bin closest to rs2205960 whilst the other maps do not. Furthermore, the observed recombination at this variant in the sex-averaged map is doubled in the female-only deCODE map and is highest at this variants relative to the 0.5Mb of encompassing sequence. The HapMap phase III and 1000 Genomes population-averaged maps depict very low recombination for this section of the *TNFSF4* locus. The trans-ancestral mapping study is predominantly in female cases and controls; I further investigated the likely pattern of recombination at *TNFSF4* in our populations in light of the deCODE data. The LD-based data I generated are

population-specific and inferred using multiple simulations and a large number of phased chromosomes from each group. The map data I generated better concur with the deCODE map for the section of the genome encompassing rs2205960.

The recombination data generated suggest that in Asians, Europeans and Hispanics the bulk of the recombination occurs in a fraction of the sequence. In African-Americans, there is increased recombination rate and higher density and proportion of hotspots across the locus. In all populations, peak recombination is consistently at the 5′ boundary of the *TNFSF4* gene and approximately 120kb into the 5′ region. A difference in African-Americans is that recombination extends 30kb from the *TNFSF4* gene boundary into the 5′ region, whilst there is negligible recombination in this section in the other populations, compatible with increased age of the genome in populations of African descent.

4.17.2 *Summary of findings: SNPs*

The data within this section collectively formed the first trans-ancestral fine-mapping association study of *TNFSF4* in SLE. Haplotype-tagging and proxy variants and major ancestry informative markers were genotyped in four distinct populations, including two admixed populations, across 200kb of 1q25 encompassing the *TNFSF4* gene, and 5′ and 3′ regions. Association data testing *TNFSF4* SNPs in African-American-Gullah SLE are also presented. Testing *TNFSF4* variants with disease status revealed strong association in all cohorts (Tables 4.2, 4.3, and 4.5) establishing *TNFSF4* as a global lupus susceptibility gene. Resolution of the association signal was accurate recombination data (Figure 4.1), and by increased power from the large numbers in our European cohort. Maximal power was achieved testing with a genetic model concordant with the major underlying mode of inheritance of the 5′ *TNFSF4* region in SLE, which is additive. The study confirmed the validity of large multi-ethnic cohorts where linkage disequilibrium is an obstacle. The novel association of rs16845607-A with Hispanic SLE was also reported. The data presented in this

chapter present two common *TNFSF4* signals responsible for the underlying association with SLE in the Hispanic population (Figure 4.5).

Rs16845607-A is located on intron 1 of the *TNFSF4* gene, and is associated with risk of Hispanic SLE with strong effect size (OR=2.06) (Table 4.5). The effect size is of the magnitude consistent with MHC risk alleles in SLE (Fernando et al., 2007). Rs16845607 is monomorphic in African-American and European populations but tags a high frequency haplotype in Hispanics (Figure 4.10). Amerindian ancestry is likely to drive increased frequency of rs16845607-A: This is supported by association data presented for independent Mestizo Amerindian and Hispanic Mestizo cohorts. Single marker and haplotype association data and results of conditional regression analysis suggest this allele represents a signal independent of the *TNFSF4* 5´ signal. A high frequency haplotype identical but instead carrying rs16845607-T was not associated with SLE risk (OR=0.86), suggesting a causal role for rs16845607-A. Sub-phenotype analyses demonstrated strong association of rs16845607-A with improved p-value and increased effect in leukopenia (OR=2.75) and lymphopenia (OR=2.94) (Table 4.8). Conditioning on the presence of *rs16845607-A* in Hispanic lymphopenia demonstrated rs16845607-A to drive the intron one association in this subgroup of lupus patients. Interrogating the DNA sequence at this variant located a DNase1 hypersensitivity site to within 1kb of rs16845607-A: Further experimental analyses, including association testing in independent Latino cohorts, and ChIP investigation of regulatory proteins that potentially bind the sequence, are required to support this variant and to understand underlying mechanism by which it regulates *TNFSF4* in lupus pathogenesis.

Mapping the alleles uniquely tagging *TNFSF4* haplotypes in each cohort helped establish the boundaries of $TNFSF4_{risk}$ $TNFSF4_{non\text{-}risk}$ in East Asians, Hispanics, and African-Americans and validated the haplotype boundaries previously defined in Europeans (Figure 4.9, and (Cunninghame Graham et al., 2008)). The spurious association of variants with disease through poor matching of cases with

controls was addressed by the removal of outlying individuals. The association of risk alleles was tested across all groups participating in this study.

Comparing recombination patterns in African-American individuals homozygous for the risk and non-risk haplotypes, there was increased recombination at the locus for risk-haplotype bearing individuals. These data provide evidence for global association of rs2205960-T with SLE. The contribution of rs2205960-T to disease risk was assessed by conditional regression, these data suggest that this allele drives the 5′ *TNFSF4* association in African-Americans, Europeans and Hispanics.

4.17.3 Summary of findings: Bioinformatics

I further investigated key associated variants at *TNFSF4* to further understand the biological processes underlying SLE pathogenesis: Curated and non-redundant profiles of SELEX binding experiments, stored in the JASPAR core database (Portales-Casamar et al., 2010) suggest rs2205960-T as the 8th nucleotide of a decameric motif with high binding affinity for NF-κB p65 (Figure 4.11). Altering the 8th nucleotide of the decamer to rs2205960-G reduced the binding affinity of this sequence for the NF-κB protein by approximately 10%, according to these data. ChIP-seq data generated for two HapMap phase III lymphoblastoid cell lines confirmed binding of NF-κB at this location. ENCODE ChIP-seq data also suggest binding of the transcription factors BCL11a, MEF2a and B-ATF at rs2205960, albeit with lower signal intensity compared to NF-κB. These data collectively suggest the genomic region encompassing rs2205960-T to have strong regulatory potential.

4.17.4 Cis-eQTL *data*

Expression profiling of common *TNFSF4* variants was carried out in a cis-eQTL study in LCL samples from 777 female TwinsUK participants (Grundberg E. et

al., 2012). Association of RNA expression with >2 x 10^6 SNPs was tested by two-step mixed model–based score test. To characterize likely independent regulatory effects, the identified cis-eQTLs were mapped to recombination hotspot intervals. For each gene, the most significant SNP per hotspot interval was selected, and LD filtering performed. The top cis-eQTL in the LD bin, for the probe located at *TNFSF4* (ILMN_2089875), was rs2205960 (P=3.75 x 10^{-4}).

4.17.5 Neutral haplotypes

Increased decay of 5′ LD at *TNFSF4* in AAs anchor the associated haplotype to the proximal 5′ region of *TNFSF4*. Examining the LD structure at *TNFSF4* in African-Americans and Europeans (Figure 4.8) supported the association data: Neutral haplotypes in these populations, recombinant between rs1234317 and rs2205960 for the risk and non-risk haplotype, confirmed our conditional regression results (Figure 4.9).

4.17.6 Sub-phenotypes

Investigating the association of *TNFSF4* risk alleles with biologically relevant lupus sub- phenotypes, strengthened the association p-value and effect size of rs2205960-T within the Anti-Smith autoantibody-positive AA lupus subgroup and this trend replicated in the European cohort (Table 4.7). The rs2205950-T allele was best associated when subgroups of Anti-Smith SLE cases were tested against AA, European and Hispanic controls.

Assessing the correlation of rs1234317-T with the presence of anti-Ro autoantibodies in European cases found increased significance of the association p-value. There was an underlying trend illustrating strong correlation of *TNFSF4* variants with autoantibodies, suggest a putative role for *TNFSF4* in their generation. The data suggest a mechanism by which *TNFSF4* variants might be

involved in lupus pathogenesis. The Genomatix SNP analysis web tool predicts rs1234317-T to destroy the DNA binding site for the transcriptional repressor E4BP4, a transcription factor with a role in the survival of early B-cell progenitors (Ikushima et al., 1997). These data were validated by the transcription-factor annotation tool in Genomatix, incorporating DNA sequences from *TNFSF4*$_{risk}$ homozygotes. However, at the time the data were generated, no other publically available data confirmed the Genomatix annotation.

Sub-phenotype analyses also demonstrate strong association of rs16845607-A with improved p-value and increased effect in leukopenia (OR=2.75) and lymphopenia (OR=2.94) (Table 4.8) subsets of Hispanic SLE cases. Conditional regression of rs16845607-A in Hispanic lymphopenia suggests that rs16845607-*A* drives the intron1 association in this subgroup of lupus patients, and is not dependent on the 5′ *TNFSF4* association. Sequence analysis locates a DNase1 hypersensitivity site to within 1kb of rs16845607-A in HapMap phase III Mexicans, however further experimental analyses are required to validate these data prior to functional experiments which investigate pathogenesis.

4.17.7 Comparison with existing studies

The association of rs2205950-T with African-American lupus concurs with data published previously by our group establishing a 5′ *TNFSF4* association with SLE in Northern Europeans (Cunninghame Graham et al., 2008). The risk-associated variants rs2205960-T and rs1234317-T are strongly associated in the Minnesota cohort consistent with our results in four racial groups. In this previous study LD was a major obstacle in delineation of causal variation. Testing the association using a very large number of Europeans and utilising an African-American cohort, the signal was refined. Conditional analyses and the presence of neutral recombinant haplotypes aided the process. With regards to previously published data for *TNFSF4* in SLE, the African-American data presented do not concur with data presented by Delgado-Vega and colleagues (Delgado-Vega et

al., 2009): These data suggest rs12039904-T and rs1234317-T to explain the entire haplotypic effect at *TNFSF4* with SLE. A possible explanation for the modest association of rs12039904-T in African-Americans presented in this thesis, is its low frequency in populations of West African descent. The data presented in this thesis also find rs12039904-T as a borderline rare allele in African-Americans, with a nominal allelic association with SLE. Conditioning on rs2205960 removed residual association at rs12039904 in all groups tested.

The key associated variant presented in this thesis, rs2205960-T, is correlated with risk of SLE in Amerindian-derived populations in a study by Sanchez and Colleagues (Sanchez et al., 2010). Sanchez and colleagues use *TNFSF4* rs2205960 and single markers at 15 other lupus susceptibility loci to illustrate correlation of Amerindian ancestry with increased frequency of lupus risk alleles. Delineation of rs2205960-T in the context of LD with adjacent markers isn't the aim of the Sanchez study, as a single SNP is typed at each locus. They find aggregation of deleterious alleles in Amerindian SLE individuals: These data are supported by the increased effect sizes we find for associated *TNFSF4* variants in Amerindians and Hispanics in this study.

The intron 1 marker rs16845607-A, was associated with risk of Hispanic SLE with strong effect size (OR=2.06) (Table 4.5). The effect size is of the magnitude consistent with MHC risk alleles in SLE (Fernando et al., 2007). Rs16845607 is monomorphic in African-American and European populations but the minor allele tags a high frequency haplotype in individuals of Amerindian ancestry (Figure 4.10).

4.18 Key points of study

In summary, the data presented in this chapter of the thesis confirmed *TNFSF4* as a global susceptibility gene in SLE. The 5′ association with disease was replicated in all racial groups; these data suggested the signal location in the

proximal *TNFSF4* promoter region. Efforts in African-Americans and a large cohort of European individuals were used to refine the association. Increased recombination in the proximal section of the 5′ upstream region of this locus in the AA population and the conditional regression strategies employed, better focused the association signal to the risk-haplotype-tagging variant, rs2205960-T. This marker was strongly associated with disease in all groups tested. This marker segregated with autoantibody subsets in African-Americans, European and Amerindian/Hispanic groups. ChIP-Seq and bioinformatic data suggest that the variant sits within a regulatory element flagged as a promoter-associated DNase1 site: Bioinformatic data mined from the Jaspar Core database suggest the risk allele forms part of a decameric motif for NF-κB RELA. ChIP-seq data from the ENCODE project additionally supported binding of NF-κB to this sequence. Collectively, these data suggest a causal mechanism for disease risk.

A novel association signal at *TNFSF4* was also identified from these data: The intron 1 *TNFSF4* variant rs16845607-A confers risk uniquely in SLE individuals with Amerindian ancestry. The results presented demonstrated segregation of this marker with lymphopenia and leukopenia which are sub-phenotypes associated with disease severity. These data suggest both global and population-specific *TNFSF4* associations with SLE exist and illustrate the use of trans-ancestral mapping in this complex trait.

4.19 Limitations

4.19.1 Limited ancestry informative data

Epidemiological studies suggest differences in the distribution of ancestry-associated susceptibility to SLE, the intron 1 association of rs16845607-A unique to mestizo Hispanic and Amerindian SLE cases described in this chapter support these data. The AIM panel of Halder and colleagues (Halder et al., 2008), limited to 347 markers, crudely distinguishes continental admixture but does not resolve hidden fine-scale genetic substructure. Populations of mestizo Hispanics with

complex disease are now being used frequently in genetic analyses; they often have large variance in the proportions of Amerindian (AMI) and European (EUR) genome (Tian et al., 2007). Assessing the relative contributions of EUR and AMI has been limited by the paucity of AIMs that distinguish them and which distinguish between different Amerindian groups such as the Uto-Aztecan speaking (eg. Pima Indians) and the Non-Uto-Aztecan language speaking (eg. Mayan) groups. The data presented in this chapter assumed homogeneity in the indigenous Amerindian population, but it is likely that they are heterogeneous as DNA samples were collected from multiple sites. A set of markers screened and validated for enrichment of EUR/AMI AIMs could provide a strong basis for future analyses which assess SLE risk loci in groups with AMI source ancestry (Tian et al., 2007)

4.19.2 Absent imputation of Amerindians and Hispanics

Imputation is a statistical approach that can be used to leverage genetic association data. Imputation of data allows estimation of untyped genotypes: This enabled comparison of haplotypes across African-Americans, Europeans and East Asians. Imputation thus aided resolution of population differences in haplotype structure at *TNFSF4*. The Hispanic cohort was not imputed due to cryptic structure owing to three or four-way admixture. At the time the data were imputed, publically-available genome-wide sequencing data representing Latinos was not available. Thus, reference haplotypes were not available for imputation; Re-sequencing 200 Latino individuals at high coverage would capture the majority of common genetic variants at this locus so that accurate, phased haplotype data could be generated for the purpose. With regards to the *TNFSF4* association, the frequency of common associated alleles is higher in Hispanics compared to the other groups tested: Pair-wise LD is also stronger. Therefore, if the Hispanic populations tested are representative, it is likely that the imputed genotypes captured by the other groups are representative for this population.

4.19.3 Recombination rate inference

Assessing the performance of *rhomap* as a recombination rate estimation tool, an upward bias of mean estimates is identified at very low rates. However,

simulations of the constant and variable recombination rate (Auton., 2007) found the variance of this estimate reduced compared to other programs so sample distribution is highly likely to contain the true rate for variable data. Comparing *rhomap* as a hotspot detection tool finds it less powerful than *HotspotFisher* or s*equenceLDhot* (Fearnhead., 2006; Krimmer et al., 2009). *Rhomap* is not suitable as an independent determinant for hotspot presence without recombination rate inference; however it is capable of identifying candidate hotspots with a lower false discovery rate than at least one other method assessed. Artificially thinning SNP density in a random but uniform manner affected the performance of *rhomap;* Detection power is reduced to below 10% for all hotspots. The markers used to infer recombination at *TNFSF4* are densely packed in all but a 10kb section of the upstream region; the section comprising SINE and LINE repeat elements. Strong LD across the encompassing upstream region in Europeans and East Asians suggest a hotspot has not been missed in the repeat region. Artificially thinning the SNPs to an allele frequency of 5% loses resolution to a lesser extent but increases the number of falsely assigned hotspots (Auton., 2007) and so *rhomap* is not suitable for simulating recombination from rare variants.

4.19.4 Imputation fall-out

A key limitation of this study is *TNFSF4* imputation may have missed common variations located in the distal 5′ *TNFSF4* region which could be causal. Accurate characterisation of variants remains challenging in low-complexity regions including the LINE element located in the distal 5′ section of this locus. As a result, variants in this region are systematically underrepresented in genetic association studies. Furthermore, an association signal may reside in the fraction of SNPs which have a lower imputation performance and were omitted using the info threshold of 0.7. This fraction is likely to include rare variants which are too infrequent to be imputed with confidence but which might have a large effect on risk. However, the association data suggest the true causal variants are likely to be common (>5% frequency) and located in the proximal section of the 5′ region. The standard error of the beta coefficients for most imputed variants included in later analyses reflects high imputation certainty.

4.19.5 East Asian Phenotype data

The data presented in this chapter finds association of *TNFSF4* alleles with phenotypic manifestations of SLE. The associated variants have improved effect size and increased significance and there are trends with autoantibody production and age at diagnosis across multiple populations. However there are no statistically significant associations between *TNFSF4* variants and the same phenotypes in the East Asian cohort, despite a similarly structured risk haplotype and strong association of *TNFSF4* with lupus per se. Numbers of cases included in analyses are on a weighted par with European cases, thus lack of association is perplexing. The phenotype datasets were collected at multiple institutions and a drawback in any large association study of this kind is the lack of standardisation in data collection, with especially large variance in reliability of self-reported phenotypes. One would not expect such errors given the East Asian cohort, and would expect it to equally apply to the other cohorts, if it was due to error alone. If absence of association is a result of error, and not a negative association, it is likely to be a systematic error introduced after collation of all phenotypes for this population as opposed to any random fluctuations in measurements.

Chapter 5

Targeted re-sequencing of *TNFSF4*

NGS re-sequencing has successfully refined the association of established susceptibility loci in lupus, including the *TNFAIP3* gene which regulates NF-κB (Graham et al., 2008). Sampling all available nucleotides at a locus catalogues previously untyped polymorphisms and allows full characterization of haplotype(s) associated with risk of disease. These data then enable comprehensive association fine-mapping, increasing the likelihood of identifying causal contributors, be they SNPs or structural variations, to SLE risk.

5.1 Targeted re-sequencing of the *TNFSF4* locus - study aims

5.1.1 Definition of variants unique to $TNFSF4_{risk}$ and $TNFSF4_{non\text{-}risk}$ haplotypes.

A sequencing study at the *TNFSF4* locus would better catalogue the genetic variation required for comprehensive association analysis of this gene in SLE. To this end, the TNFSF4 gene, 5′ and 3′ regions were deep-sequenced in UK-European SLE individuals who were homozygotes for the aforementioned $TNFSF4_{risk}$ or $TNFSF4_{non\text{-}risk}$ haplotypes. The conditional regression strategies applied to genotypic data in chapter 4 identified independent risk and protective effects owing to variants contained within $TNFSF4_{risk}$ and $TNFSF4_{non\text{-}risk.}$ Sequencing UK-European SLE individuals who are homozygotes for these haplotypes, I aimed to systematically identify the variants which implicate these signals in SLE.

5.1.2 Definition of full spectrum of variants underlying the upstream TNFSF4 association in lupus.

The trans-ancestral mapping experiment identified genetic association of *rs2205960-T* in all cohorts tested. The association of alleles at *TNFSF4* in multiple SLE populations confirmed *TNFSF4* as a global susceptibility gene for the disease. Utilising recombinant haplotypes and applying conditional regression strategies to these data, rs2205960-T best explains the association of the upstream region with disease. However, only a proportion of the available nucleotide positions available at the locus were sampled to generate these results: Imputation may have missed common variations located in the distal 5′ *TNFSF4* repetitive region; variants in this low-complexity region are systematically underrepresented in genetic association studies. Causal-variants, including indels, which might be in LD with rs2205960-T, but which have not yet been identified may be additional causal contributors to disease risk. Thus, by sequencing the

locus in many lupus individuals, I aimed to catalogue a more complete spectrum of variants, including structural variants, which potentially confer disease risk.

5.1.3 Definition of rare SLE-associated coding variants.

The rs2205960-T variant is a common allele with modest effect size; it is likely to explain a proportion of the variance in SLE disease risk owing to *TNFSF4*. Independently associated coding variations would more directly implicate the *TNFSF4* gene and also explain additional heritability. These variants are likely to reside in the fraction of SNPs which have a lower imputation performance as they are more likely to be deleterious, negatively selected, and so occur at lower frequency. An info threshold of 0.7 was used to omit variants in this study: This fraction of omitted variants is likely to include rare variants which are too infrequent to be imputed with confidence but which might have a large effect on risk. These rare and probably functional variants, once validated, would be used to dictate functional research of *TNFSF4* in SLE.

Results

5.2 Sequencing statistics

Sequencing of the 118kb section of chromosome 1q25.1 (the *TNFSF4* gene and genomic region up to the 5′ boundary of the risk-associated haplotype) was undertaken. Using this approach, 177.2Mb of sequence was generated in the form of 555,721 reads with a mean read length of 318 (**Figure 5.1*A***). The depth and breadth of sequence coverage (**Figure 5.2, Table 5.1**) was calculated using a custom script designed by Michael Simpson at Kings College London. The parallel-tagged sequencing (PTS) strategy of Meyer and colleagues (Meyer et al., 2008) enabled multiplexing of DNA sequence from 88 $TNFSF4_{\text{risk}}$ and $TNFSF4_{\text{non-risk}}$ homozygote SLE individuals using half a Roche-454 Titanium chip.

5.3 Variant-calling pipeline

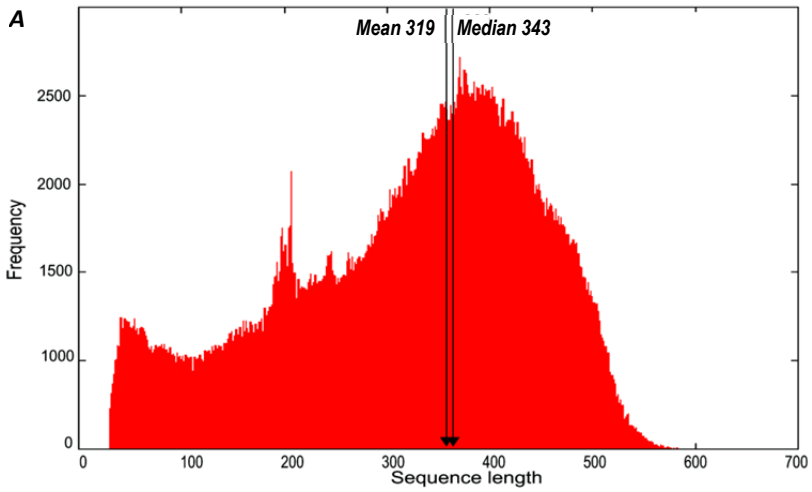
In order to search for new variants at *TNFSF4,* profiles were generated using an in-house variant calling pipeline. Sequence reads were aligned to the reference genome (UCSC hg18, NCBI Build36.3) and anomalous reads excluded from downstream analysis. Coverage sufficient for variant calling was achieved in 71 of the 88 Northern European SLE individuals selected for sequencing (**Table 5.1, Figure 5.2**). The 5′ region upstream of the *TNFSF4* gene, in particular the section corresponding to the 13.8kb amplicon with coordinates 171,462,182-171,476,028, predominantly comprises conserved repetitive elements: High quality sequencing data for this amplicon was generated with good coverage and depth in only three individuals; one risk (tagged by adapter 33) and two non-risk (tagged by 36 and 65) homozygotes. As a result, a reduced number of variants were called for this section of the locus.

5.4 Variant calling

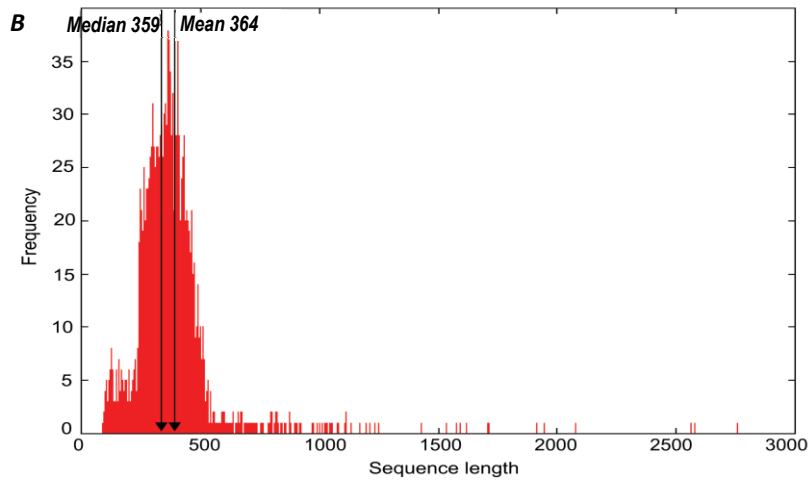
Single nucleotide substitutions and small insertion deletions (indels) were identified and further filtered using the overall mean quality scores (calculated from the sample-specific quality scores) generated by the variant calling tool. Variants were also visualised using the Integrated Genomes Viewer (IGV), as described in chapter 1 and depicted in **Figure 5.3**. Novel variants were aligned against the two main transcripts of the *TNFSF4* gene; data are presented in **Table 5.2.**

Figure 5.1 Sequencing statistics generated for *A.* all sequencing reads (above) and *B.* assembled contigs (below)

Metrics, all sequences	
Statistic	**Value**
n	555,721
sum	177,208,692
mean	318,88
sd	123.8
median	343
mode	371
q1	232
q3	414
min	28
max	1,438
range	1,410



Metrics, contigs	
Statistic	**Value**
n	5,655
sum	2,056,938
mean	363.74
sd	131.86
median	359
mode	363
q1	298
q3	417
min	92
max	2,756
range	2,664



Tag ID	Tag sequence	Number of Sequences	Coverage	Variants called	Tag ID	Tag sequence	Number of Sequences	Coverage	Variants called
TAG 01	TCTCTGTG	11680	35	24	**TAG 36**	ATATCACG	83284	249	105
TAG 02	TGTACGTG	22188	66	28	**TAG 39**	AGCACACG	22504	67	65
TAG 03	ATCGTCTG	31948	96	106	**TAG 40**	ATGTGTAG	27876	83	50
TAG 04	TAGCTCTG	33144	99	70	**TAG 41**	ACTCGTAG	48132	144	90
TAG 05	AGTATCTG	41892	125	71	**TAG 42**	TGCAGTAG	11608	35	35
TAG 06	TCGAGCTG	49536	148	139	**TAG 43**	TGATCTAG	22508	67	4
TAG 07	TCATACTG	38136	114	86	**TAG 44**	TACGCTAG	26044	78	53
TAG 08	TACGACTG	21084	63	95	**TAG 46**	AGACATAG	16968	51	59
TAG 09	ACTCACTG	63088	189	164	**TAG 47**	AGCGTGAG	35940	108	78
TAG 10	AGAGTATG	46420	139	75	**TAG 48**	ATGATGAG	32312	97	87
TAG 11	AGCTGATG	9860	29	32	**TAG 50**	TCTGCGAG	11968	36	66
TAG 12	TATCGATG	28184	84	61	**TAG 51**	ATAGAGAG	43640	131	121
TAG 13	ATGCGATG	44208	132	86	**TAG 52**	TATCAGAG	41056	123	104
TAG 14	ACGTCATG	23728	71	34	**TAG 53**	ACGCAGAG	10656	32	44
TAG 15	TCATGTCG	39312	118	145	**TAG 54**	ACAGTCAG	7580	23	23
TAG 17	TCTACTCG	56316	168	123	**TAG 55**	TCTATCAG	7964	24	16
TAG 19	ATCTATCG	44836	134	73	**TAG 58**	ATCAGCAG	26660	80	129
TAG 20	ACAGATCG	13060	39	5	**TAG 59**	TGCTACAG	15596	47	21
TAG 21	ATACTGCG	21076	63	40	**TAG 60**	AGTGACAG	39068	117	89
TAG 22	TATATGCG	57772	173	88	**TAG 62**	TACATGTC	36656	110	116
TAG 23	TGCTCGCG	12760	38	11	**TAG 63**	ATGACGTC	22428	67	100
TAG 24	ATCGCGCG	21948	66	64	**TAG 64**	AGCGAGTC	18576	56	62
TAG 26	AGATAGCG	57508	172	122	**TAG 65**	TCGCAGTC	32928	99	101
TAG 27	TGTGAGCG	11888	36	40	**TAG 66**	ATACAGTC	87652	262	148
TAG 28	TCACAGCG	32156	96	61	**TAG 68**	TCACTCTC	120408	360	199
TAG 29	ACTGTACG	15212	46	26	**TAG 69**	ATCTGCTC	10348	31	10
TAG 30	TGCGTACG	28824	86	58	**TAG 72**	TCTGACTC	90892	272	36
TAG 31	TCGCTACG	31612	95	71	**TAG 76**	TGACGATC	19204	57	1
TAG 32	TACTGACG	65004	194	132	**TAG 79**	ACAGCATC	10116	30	1
TAG 33	AGACGACG	39628	119	131	**TAG 82**	TGCGATGC	7468	22	16
TAG 34	TGTAGACG	19628	59	57	**TAG 85**	AGAGTCGC	19928	60	59
TAG 35	ACGAGACG	13148	39	34	**TAG 88**	TCATGCGC	32772	98	61

Table 5.1 **Number of sequence reads generated and coverage per tagged individual, *TNFSF4* sequencing study**

Figure 5.2 Bar chart illustrating *A*. Number of sequences per individual and *B.* The number of variants called per individual

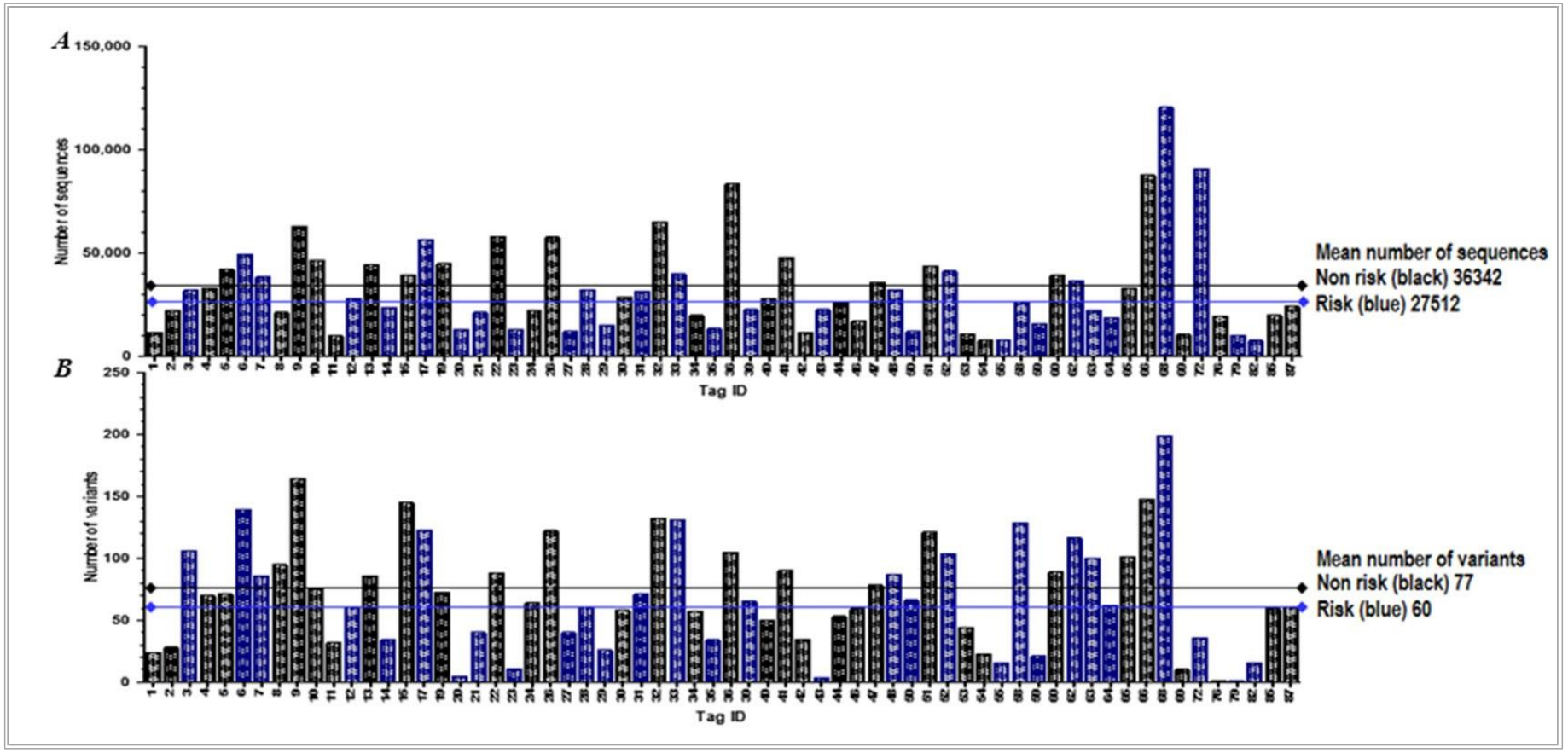
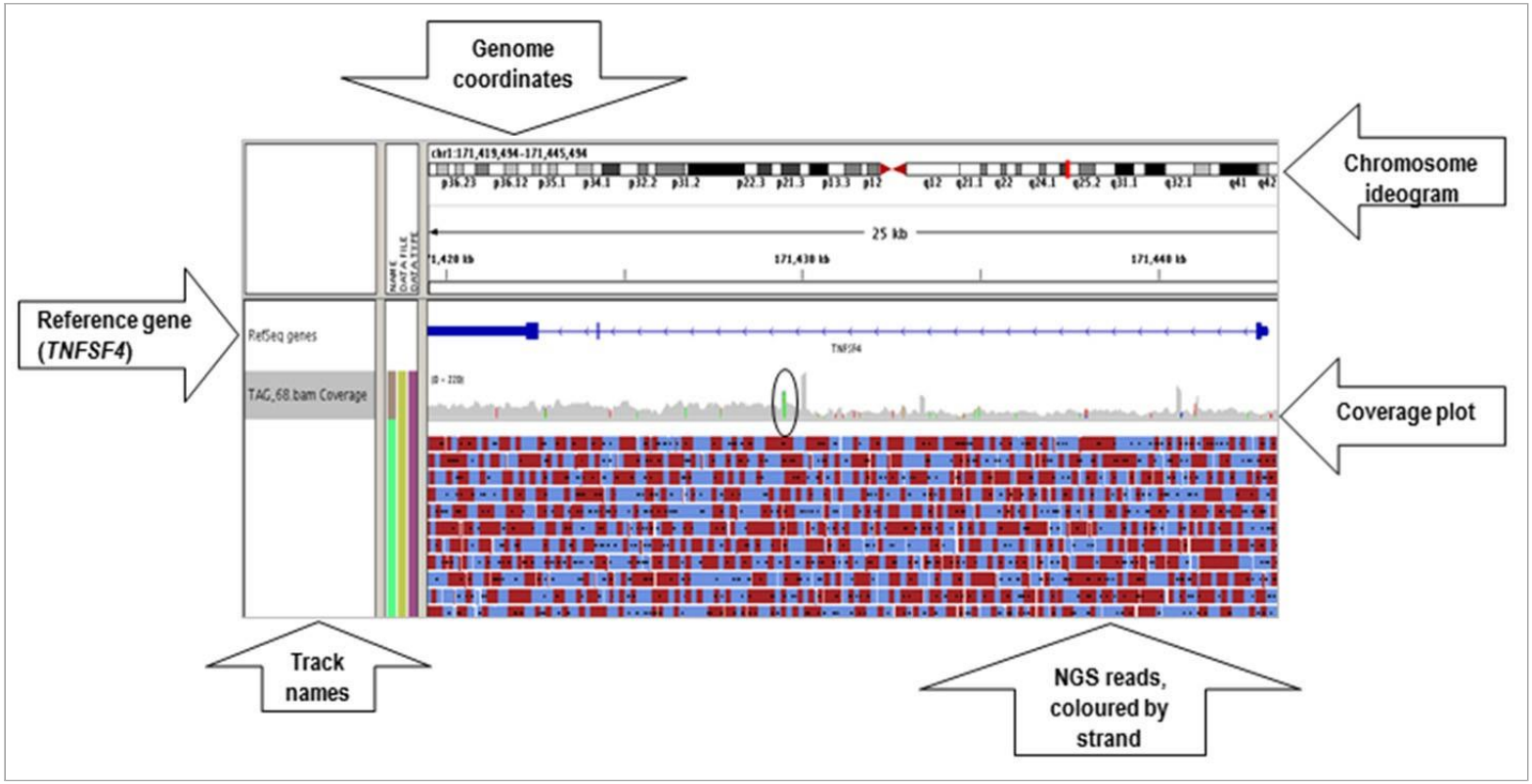


Figure 5.3 Integrated view of sequencing reads aligned against the *TNFSF4* gene (hg18) for a *TNFSF4*$_{risk}$ homozygote



The integrated genomes viewer (IGV) tool (Robinson et al., 2011) was used to explore aligned 454 reads aligned against the *TNFSF4* gene using the multi-resolution file formats bam and bam coverage. The example track depicts the full complement of sequencing reads at *TNFSF4,* post QC, for a single *TNFSF4*$_{risk}$ haplotype homozygote individual, tagged by the barcode adapter TAG68. Sporadic coloured vertical lines on the grey coverage plot immediately below the gene indicate a nucleotide base change at the position (example green line, black-circled).

Table 5.2 **All versus novel variants in two transcripts of the TNFSF4 gene**

	Transcript a.		Transcript b.	
TNFSF4 position	All variants	New variants	All variants	New variants
Exon 3	24	16	24	16
2+ individuals	24	16	24	16
3+	18	11	18	11
10+	4	1	4	1
Intron 2-3	17	10	17	10
2+	17	10	17	10
3+	14	8	14	8
10+	5	2	5	2
Exon 2	0	0	0	0
Intron 1-2	90	57	79	53
2+	90	57	79	53
3+	62	37	53	36
10+	20	5	11	5
Exon 1	-	-	3	3
2+	-	-	3	3
3+	-	-	2	2
10+	-	-	0	0
5'UTR	-	-	2	1
2+	-	-	2	1
3+	-	-	0	0
10+	-	-	0	0

5.5 Identification of novel variants

Novel variants were screened against UCSC hg19 (February 2009 high coverage assembly GRCh37) and also against SNPs and structural variations catalogued in the Ensembl Genome Browser 64, dbSNP132, HapMap data release 28, 1000 Genomes high coverage trios, 1000 Genomes high coverage exons and 1000 Genomes low coverage data. The URLs for the aforementioned browsers are described in Appendix B. of this thesis.

Exonic variants identified in the *TNFSF4* gene were probed against those catalogued from 350 control exomes sequenced and analysed by the method described (further described in **Table 5.3, Table 5.4**). In addition, the exonic

variants were probed against *TNFSF4* variants from the the first data freeze of 2500 European and African-American control exomes and contained within the Exome Variant Server from the NHLBI Exome Sequencing Project (ESP) (URL: http://evs.gs.washington.edu/EVS/). Variants with a phred quality score above 40 were identified as 'novel' if they did not concord with the variants in the two sets of exome data. The variants were interrogated using the IGV (**Figure 5.3**) to further ensure they were not a result of sequence ambiguity.

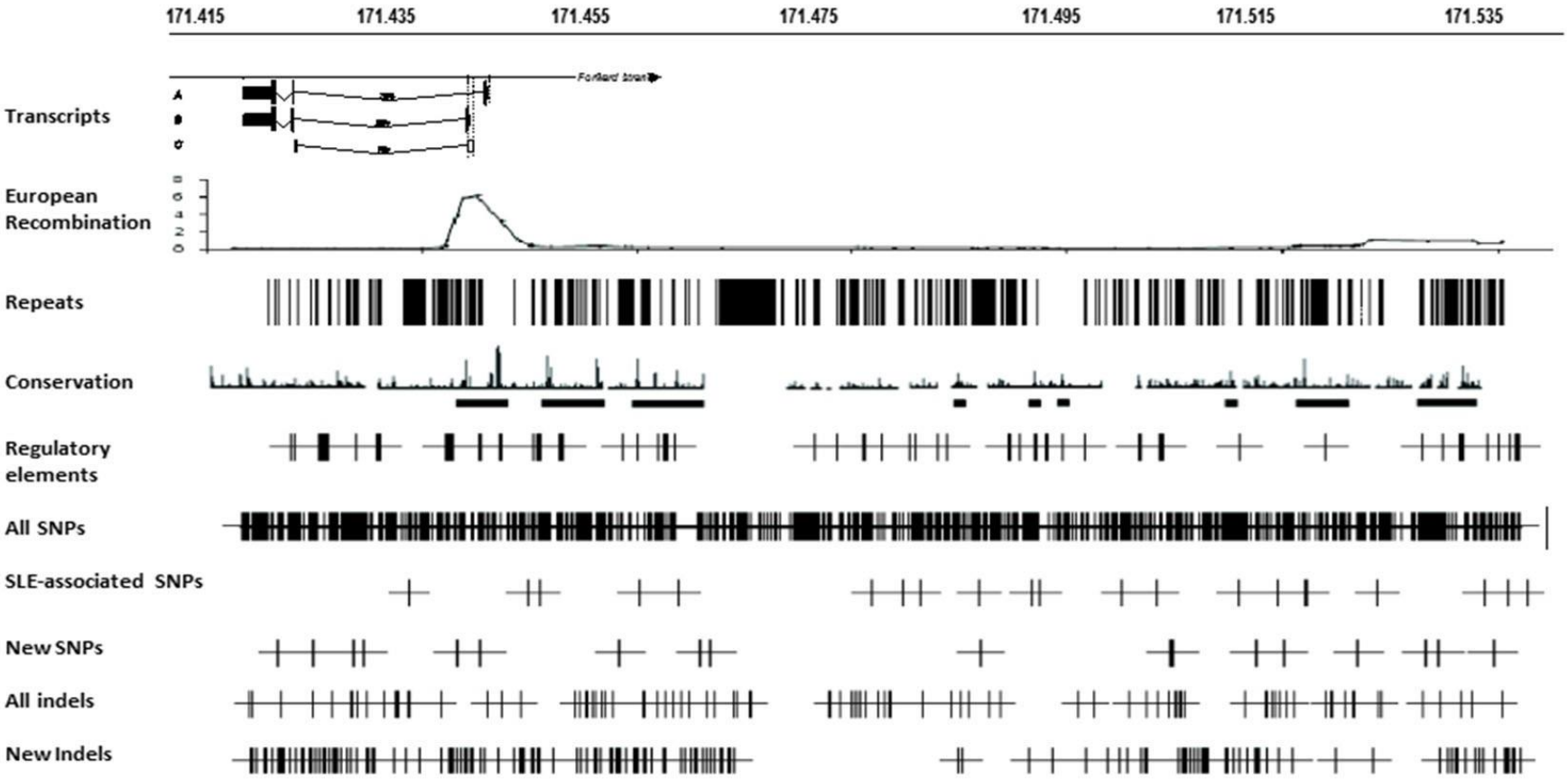
5.6 Polyphen-2 analysis

PolyPhen-2 (Polymorphism Phenotyping v2) is a tool which annotates coding non synonymous SNPs (Adzhubei et al., 2010). Predictions for potential impact of coding variants identified in this study on amino acid substitution and thus structure and function of human TNFSF4 were made. Orthologs and paralogs of the *TNFSF4* sequence were used to increase the accuracy of the predicted effect in the multiple sequence alignment (MSA). After identification and alignment of *TNFSF4* homologs, putative coding variants were interrogated for their predicted functional impact with respect to TNFSF4. The Polyphen-2 tool replaced amino acids where the variant caused a non-synonymous change and a naive Bayes classifier was used in two datasets (HumDiv and HumVar) to predict and classify the functional impact of each coding variant. These data are described in this chapter **(Table 5.4 and Figure 5.5)**.

ID	Build 36	Build37	Variant type	Total freq	Risk freq	Non risk freq	% Risk, sequencing data	% Non- risk sequencing data	Mean quality risk	Mean quality non-risk	Variant detail	Transcript 1	Transcript 2	Regulatory features
2	171419147	173152524	INDEL	13	4	9	16.00	30.00	62.5	69.39	T-TT	Exon 3	Exon 3	
3	**171420289**	**173153666**	**INDEL**	**7**	**0**	**7**	**0.00**	**23.33**		**87.37**	**TT-T**	**Exon 3**	**Exon 3**	
7	**171421041**	**173154418**	**INDEL**	**6**	**0**	**6**	**0.00**	**20.00**		**66.5**	**A-AA**	**Exon 3**	**Exon 3**	
16	**171423523**	**173156900**	**INDEL**	**7**	**0**	**7**	**0.00**	**23.33**	**69.29**		**A-AA**	**Intron 2**	**Intron 2**	SP1(1)
20	171423863	173157240	INDEL	22	7	15	28.00	50.00	92.92	83.77	T-TT	Intron 2	Intron 2	CTCF(2)
21	171424250	173157627	INDEL	17	10	7	40.00	23.33	84.85	87.07	A-AA	Intron 2	Intron 2	
30	171426673	173160050	INDEL	12	8	4	32.00	13.33	78.75	61.5	A-AA	Intron 1	Intron 1	Max(2), Cmyc(3), DNase1(1), E2F6(1)
53	171430108	173163485	INDEL	7	5	2	20.00	6.67	61.5	62.5	T-TT	Intron 1	Intron 1	
54	171430209	173163586	INDEL	6	4	2	23.53	8.70	65.6	81	A-AA	Intron 1	Intron 1	
57	**171432884**	**173166261**	**INDEL**	**3**	**0**	**3**	**0.00**	**13.04**	**84.67**		**T-TT**	**Intron 1**	**Intron 1**	
64	171437014	173170391	INDEL	5	1	4	5.88	17.39	78.15		T-TT	Intron 1	Intron 1	
67	**171437855**	**173171232**	**INDEL**	**3**	**0**	**3**	**0.00**	**13.04**	**63.83**		**A-AA**	**Intron 1**	**Intron 1**	
69	171438386	173171763	INDEL	5	1	4	5.88	17.39	81.625		A-AA	Intron 1	Intron 1	CTCF(1), Rad21(2), DNase1(4)
70	**171438729**	**173172106**	**INDEL**	**9**	**0**	**9**	**0.00**	**39.13**	**92.82**		**CCCTCCTC-C**	**Intron 1**	**Intron 1**	CTCF(1), Rad21(2), DNase1(4)
72	171438892	173172269	INDEL	7	2	5	11.76	21.74	74.75	78.88	A-AA	Intron 1	Intron 1	
75	171440072	173173449	INDEL	6	1	5	5.88	21.74	79.3		A-AA	Intron 1	Intron 1	
76	171441075	173174452	SNP	4	3	1	17.65	4.35	31		T-A	Intron 1	Exon 1	**Cmyc(1), DNase1(4)**
79	171443121	173176498	INDEL	4	1	3	7.14	18.75	61.83		T-TT	5' region	5' region	DNase1(1), BATF(1)
82	**171444200**	**173177577**	**INDEL**	**3**	**3**	**0**	**21.43**	**0.00**	**68.23**		**A-AA**	**5' region**	**5' region**	DNase1(1)
92	171449789	173183166	INDEL	10	3	7	21.43	43.75	87.33	73.21	A-AA	5' region	5' region	
102	171452784	173186161	INDEL	7	6	1	31.58	9.09	80.83		A-AA	5' region	5' region	
129	171461900	173195277	INDEL	8	6	2	31.58	18.18	76.08	99	T-TT	5' region	5' region	
141	**171497624**	**173231001**	**INDEL**	**4**	**4**	**0**	**44.44**	**0.00**	**85.5**		**CC-C**	**5' region**	**5' region**	
145	**171499486**	**173232863**	**INDEL**	**3**	**3**	**0**	**33.33**	**0.00**	**71.23**		**AA-A**	**5' region**	**5' region**	
151	171503293	173236670	INDEL	10	2	8	11.76	38.10	52.5	68.5	T-TT	5' region	5' region	
152	**171503339**	**173236716**	**SNP**	**20**	**19**	**1**	**111.76**	**4.76**	**86.72**		**T-C**	**5' region**	**5' region**	EBF(1), DNase1(3), BATF(1)
154	**171503679**	**173237056**	**INDEL**	**17**	**17**	**0**	**100.00**	**0.00**	**92.4**		**A-AACAGGA**	**5' region**	**5' region**	
164	**171507785**	**173241162**	**INDEL**	**4**	**0**	**4**	**0.00**	**19.05**	**71.5**		**A-AA**	**5' region**	**5' region**	
165	**171508341**	**173241718**	**INDEL**	**5**	**0**	**5**	**0.00**	**23.81**	**80.1**		**A-AA**	**5' region**	**5' region**	
166	**171508485**	**173241862**	**INDEL**	**6**	**0**	**6**	**0.00**	**28.57**	**84.92**		**A-AA**	**5' region**	**5' region**	
172	171511291	173244668	INDEL	9	2	7	11.76	33.33	83.75	78.07	T-TT	5' region	5' region	
177	171520353	173253630	SNP	27	15	12	150.00	70.59	87.94	82.19	T-C	5' region	5' region	
192	171531498	173264875	INDEL	8	2	6	8.70	20.00	96.25	84	A-AA	5' region	5' region	
193	171531632	173265009	INDEL	11	3	8	13.04	26.67	78.5	75.63	T-TT	5' region	5' region	
194	171531667	173265044	INDEL	6	5	1	21.74	3.33	61.9		T-TT	5' region	5' region	
197	171531931	173265308	INDEL	9	2	7	8.70	23.33	69.5	95.21	T-TT	5' region	5' region	
200	**171532531**	**173266001**	**SNP**	**29**	**25**	**4**	**108.70**	**13.33**	**85.4**	**99**	**C-T**	**5' region**	**5' region**	

Table 5.3 Putative novel variants at *TNFSF4* identified at higher frequency in $TNFSF_{risk}$ or $TNFSF4_{non\text{-}risk}$ individuals. These variants preferentially tag $TNFSF4_{risk}$ or $TNFSF4_{non\text{-}risk}$ haplotype. The variants, SNPs and indels, were identified in at last three sequenced individuals. Shading illustrates variants which segregate with $TNFSF4_{non\ risk}$ (shaded blue) or $TNFSF4_{risk}$ (shaded pink). Variants in bold are to be included in the primary round of validation.

Figure 5.4 *Novel SNP*s and indels in the context of known genomic signatures at the *TNFSF4* locus

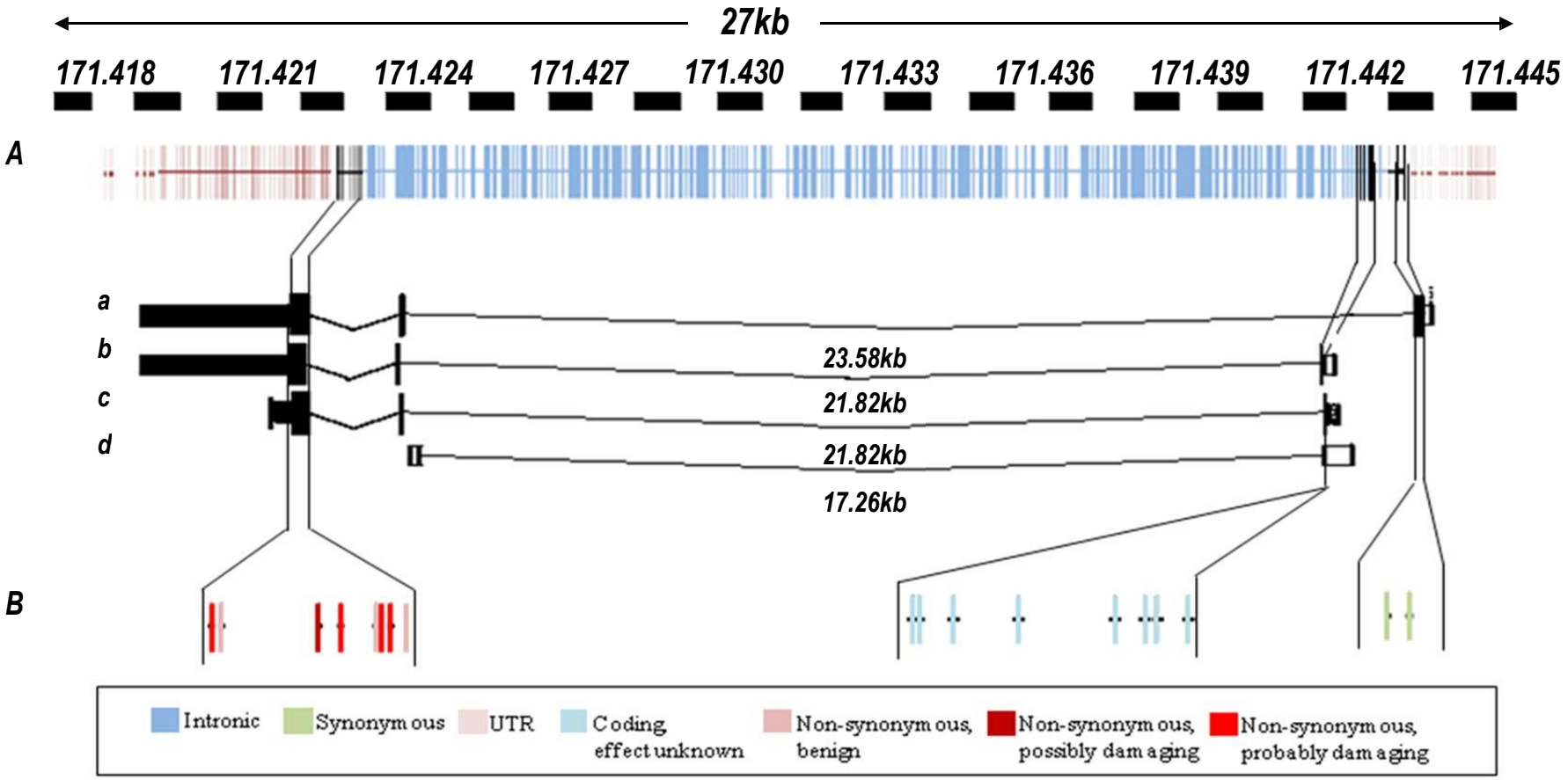


Variants were identified by targeted deep-sequencing of the 118kb section of chromosome 1q25.1 encompassing the *TNFSF4* gene and upstream 5′ region. Known SNP markers and indels from the December 2011 release of Ensembl, UCSC and 1000 Genomes browsers, along with SNPs from the NHLBI exome sequencing project and King's College London exome sequencing project are mapped in 'All SNPs' and 'All indels'; new variants in these classes, excluding singletons, are presented in the new SNP and indel sections. The data presented are aligned to the recombination data generated for 1568 control chromosomes from European individuals, which is presented in chapter 4 of this thesis.

Table 5.4 *TNFSF4 coding variants predicted by Polyphen-2 as benign or probably/possibly damaging*

Variant ID	hg18	hg19	Risk/ Non-risk	Quality Score	Position	AA1/AA2	Human Diversity Prediction	pph2_prob	Human Variation prediction	pph2_prob
173155664.AC.giv	**171422287**	**173155664**	**NR**	**21**	**131**	**C/W**	**probably damaging**	**0.961**	**possibly damaging**	0.54
173155664.AC.giw	**171422287**	**173155664**	NR	21	**181**	**C/W**	**probably damaging**	**1**	**probably damaging**	**0.941**
173155685.AC.giv	171422308	173155685	R	48	124	H/Q	possibly damaging	0.577	benign	0.168
173155685.AT.giv	171422308	173155685	R	48	124	H/Q	possibly damaging	0.577	benign	0.168
173155685.AC.giw	171422308	173155685	R	48	174	H/Q	possibly damaging	0.709	benign	0.256
173155685.AT.giw	171422308	173155685	R	48	174	H/Q	possibly damaging	0.709	benign	0.256
173155834.GA.giv	171422457	173155834	NR	78.5	75	P/S	possibly damaging	0.922	benign	0.243
173155834.GC.giv	171422457	173155834	NR	78.5	75	P/A	possibly damaging	0.557	benign	0.061
173155834.GT.giv	171422457	173155834	NR	78.5	75	P/T	possibly damaging	0.843	benign	0.243
173155834.GA.giw	**171422457**	**173155834**	**NR**	**78.5**	**125**	**P/S**	**probably damaging**	**0.999**	**possibly damaging**	**0.896**
173155834.GC.giw	**171422457**	**173155834**	**NR**	**78.5**	**125**	**P/A**	**probably damaging**	**0.992**	**possibly damaging**	**0.629**
173155834.GT.giw	**171422457**	**173155834**	**NR**	**78.5**	**125**	**P/T**	**probably damaging**	**0.998**	**possibly damaging**	**0.896**
173155921.TA.giv	**171422544**	**173155921**	**R**	**23**	**46**	**N/Y**	**probably damaging**	**0.993**	**possibly damaging**	**0.692**
173155921.TC.giv	171422544	173155921	R	23	46	N/D	possibly damaging	0.465	benign	0.159
173155921.TG.giv	**171422544**	**173155921**	**R**	**23**	**46**	**N/H**	**probably damaging**	**0.959**	**possibly damaging**	**0.576**
173155921.TA.giw	**171422544**	**173155921**	**R**	**23**	**96**	**N/Y**	**probably damaging**	**0.989**	**possibly damaging**	**0.669**
173155921.TC.giw	171422544	173155921	R	23	96	N/D	possibly damaging	0.579	benign	0.169
173155921.TG.giw	**171422544**	**173155921**	**R**	**23**	**96**	**N/H**	**probably damaging**	**0.988**	**possibly damaging**	**0.668**
173155944.AC.giw	**171422567**	**173155944**	**R**	**29**	**88**	**V/G**	**possibly damaging**	**0.65**	**possibly damaging**	**0.592**

Figure 5.5 Classification of novel variants identified in the TNFSF4 gene



A. Vertical lines depict novel variants identified in the *TNFSF4* gene after sequencing on the Roche-454 platform. This upper section of the diagram depicts the location of ticks against scale (coordinates as for UCSC hg18) and against the three known translated spliceoforms of *TNFSF4* (*a b and c)* and a single untranslated splice variant (*d*). The colour of each tick represents where it is located in the gene: black (exon), blue (intron) and pink (untranslated region, UTR). *B*. Functional effects of putative novel coding SNPs were predicted using the sequence and structure-based annotation tool *Polyphen-2 (*Adzhubei *et al.* 2010), coding variants are classified for their predicted effect on the TNFSF4 protein.

5.7 Discussion

Deep sequencing of *TNFSF4* using DNA from $TNFSF4_{risk}$ and $TNFSF4_{non\text{-}risk}$ SLE cases aimed to identify putative, high-quality variants which segregate with the aforementioned haplotypes for further validation, genotyping, association testing and functional assessment. To this end, 200 novel SNPs and indels, with quality scores above 40, were identified. These variants screened positively if they were absent in the Ensembl genome browser 64, UCSC genome browser, dbSNP132, HapMap data release 28 or the 1000 Genomes data release v2. The variants were identified in two or more of the individuals sequenced in this study. Exonic variants were probed against variants identified in 350 in-house control exomes and also against the first data freeze of 2500 European and African-American control exomes from the NHLBI Exome Sequencing Project (ESP) (URL: http://evs.gs.washington.edu/EVS/). The identified variants were categorised as 'novel, to be validated' if absent in the aforementioned datasets.

High frequency, probably low penetrant variants and low frequency variants with the potential for high disease penetrance were identified from the data presented in this chapter. Validation followed by genotyping is likely to find association of a proportion of these novel variants with SLE, given the high LD across the *TNFSF4* locus in the selected individuals. The contribution of *TNFSF4* variants to the genetic burden underlying SLE risk is likely to be from both common and rare variants: Targeted re-sequencing of susceptibility loci in IBD identified multiple rare and common associated variants at single loci (Rivas et al., 2011). Multiple independent signals were also identified at *TNFSF4* in SLE individuals of Amerindian descent: These loci had markedly different effect sizes and allele frequencies. A similar trend might be identified in Europeans should one of the novel, low-frequency coding variants identified exert an independent effect in Northern European SLE individuals tested for association.

5.7.1 Novel SNPs

The total number of SNPs identified in this study was 1399; of these 1213 were singletons and 187 identified in two or more individuals. As expected, the mean quality score for singletons was lower than that for high-frequency identified

alleles: 44.4% against a mean of 70.5% for SNPs found in two or more individuals. Removing the quartile of singleton SNPs with lowest quality score improved the mean score for this group to 56.9% for the remaining 'rare' variants: A considerably higher score compared to the same region sequenced as part of the pilot release of the 1000 Genomes study. The aforementioned 1000 Genomes release relied on low coverage (x2-4 coverage) data in intergenic regions and so early versions had higher base-calling error rates. The sequences from 71 individuals were included for variant calling: Uncertainty was reduced owing to high average read depths (x20 coverage, or greater) for this group. Sequencing data for a 13.8kb amplicon which represented a low complexity repeat region at the locus (see repetitive DNA and sequencing, below) was excluded from this analysis, and independently analysed where appropriate.

Excluding known common SNPs and singletons from the cleaned data gave 17 novel SNPs which were identified in two or more individuals; the 17 SNPs had a mean quality score of 57%. Investigating the location of these variants, there was a single exonic, six intronic and 11 intergenic novel SNPs. These variants will be validated by Sanger sequencing. Two novel upstream SNPs at coordinates 171,503,339 (frequency, 19 risk vs. one non- risk) and 171,532,531 (25 risk vs. four non-risk) are annotated for regulatory features at their genomic location in the Ensembl genome browser: The variant at 171,532,531 is found within a regulatory region, *in silico* ChIP-seq data predict the B-ATF transcription factor to bind the sequence encompassing the variant. B-ATF enhances the transcription of genes involved in B-cell proliferation (Ensembl, release 132). Increased B-ATF binding would be a plausible mechanism to cause *TNFSF4* overexpression thus increasing risk in SLE.

5.7.2 Novel Indels

The deep sequencing data presented in this chapter revealed 531 indels with a mean quality score of 69.9. Of the total, 226 ('high frequency') indels were identified in two or more individuals (mean quality score 73.4) and 186 ('novel') out of 226 were not identified in the aforementioned genome browsers or exome data. Of the 186 high frequency novel indels, 37 were annotated as having

regulatory potential (Ensembl version 132 (December 2011), UCSC genome browser).

Several novel high-frequency indels identified in this study segregate with the *TNFSF4*$_{\text{risk}}$ or *TNFSF4*$_{\text{non-risk}}$ haplotypes: A 7bp deletion (CCCTCCTC to C) in intron 1 of the gene, with start coordinate 171,440,072 (hg18, NCBI v36.3) was identified in nine *TNFSF4*$_{\text{non-risk}}$ individuals. This deletion was not found in a single risk individual sequenced. The motif may form part of a larger degenerate 13bp motif (CCNCCNTNNCCNC) which causes increased recombination and forms of genomic instability in humans (Hinch et al., 2011). The *TNFSF4* gene is on the reverse strand so all sequenced risk individuals appear to fulfil the criteria for this degenerate motif. Recombination rate simulation data presented in chapter 4 of this thesis highlight increased recombination in *TNFSF4*$_{\text{risk}}$ compared to *TNFSF4*$_{\text{non-risk}}$ homozygotes. The data relating to this motif, presented by Hinch and colleagues, has demonstrated association with disease-causing genomic rearrangements in individuals carrying it. The motif may serve as a binding site for the chromatin-modifying PRDM9 protein (Berg et al., 2010) which binds the chromatin signature histone3 lysine4 (H3K4), by binding the DNA sequence. Use of the 8bp motif to bring about folding of *TNFSF4* might bring enhancers and transcription factors within close proximity of the gene. The TF would be predicted to bind associated variants identified in chapter 4, a plausible mechanism to cause overexpression of *TNFSF4* and increase susceptibility to SLE. Folding of the sequence at *TNFSF4* could be interrogated using the chromatin conformation capture (3C) technique.

DNA sequences which are regulatory factor motifs often have biological function on factor binding: Sequence-specific transcription factor binding sites (TFBS) are sequences with strong regulatory potential: A high-frequency 6bp insertion (A to AACAGGA) identified in this study (Table 5.3, variant ID 154) was identified in 17 risk individuals but absent in the non-risk group. The Jaspar and Genomatix data did not offer a consensus for the binding regulatory factor, however the motif is likely to have strong potential for regulation as multiple factors are predicted to bind the sequence.

Successful validation of all high-frequency novel indels presented would represent an increase in the total number identified at *TNFSF4* by a factor of 4.7. A recent NGS study in Koreans identified approximately twice as many indels in each individual compared to previously sequenced personal genomes (Ju et al., 2011). This study, presented by Ju and colleagues, identified substantial heterogeneity in the number of indels between the sequenced individuals. Comparing the indels identified in a single individual in the *TNFSF4* study with the reference genome found a fold increase of 1.9 in the *TNFSF4* study; this compares favourably and supports the data generated by the Korean study. The *TNFSF4* data included in the overall fold increase take into account high-quality indels discovered by the pilot phase of the 1000 Genomes Project. The fold increase of 4.7 does not include 305 novel indels identified as singletons in this study. The lowest individual phred quality score for an indel which occurs once is 50: These variants had a higher base calling accuracy compared to single occurrence SNPs identified in this study.

The excess in novel indels at *TNFSF4* from the study described in this thesis may be attributed to a combination of more accurate read alignments from increased lengths obtained by 454 sequencing, an increased proportion of available paired-end sequencing reads (Ju et al., 2011) but also errors including homopolymers, which are a known consequence of 454 sequencing (see below). A locus-specific map summarizing the location of indels is depicted in this chapter (see Figure 5.4)**.**

5.7.3 Population or disease specific?

Variants identified by this sequencing project may be uniquely associated with SLE or relevant to other complex polygenic traits. Polymorphisms located in intron 1 of the *TNFSF4* gene are risk alleles in lupus but have also recently been associated with thick carotid plaque formation in the inflammatory process leading to severe atherosclerosis (Gardener et al., 2011). *TNFSF4* alleles associated with SLE risk have also been associated with cerebrovascular disease (Olofsson et al., 2009).

5.7.4 Rare variants and non-synonymous SNPs in TNFSF4

Accurate estimates of the number of rare variants in an individual genome are difficult to attain as they are likely to include many singleton variants that are false positives. An approximation of the total number can be calculated from positive predictive values of SNP detection for the experiment: The whole-genome Korean study estimated that 1.5% variants with an allele frequency of <1% were false (Ju et al., 2011). For the small-scale project presented in this chapter, most variations identified across the *TNFSF4* gene were singletons, and so are rare polymorphisms or, more likely, false-positives. Within this fraction of singletons, the number of SNPs which cause a change in the amino-acid sequence, so called non synonymous SNPs (nsSNPs) are likely to be enriched as negative selection drives the frequency of deleterious alleles lower. A nsSNP could potentially affect the function of the protein, subsequently altering the carrier's phenotype (Kumar et al., 2009). To assess the effect of missense substitutions on the *TNFSF4* protein, the Polyphen-2 tool, which uses a sequence-based and structural approach, was used to evaluate *TNFSF4* coding variants (Adzhubei et al., 2010).

The naive Bayes classifier is a simple predictive tool which makes very independent (naive) computational predictions of functional impact. The coding variants identified were predicted as 'benign', 'probably damaging' or 'potentially damaging' (Table 5.4). A missense substitution at position 171422287 (hg18), Cys181Trp (Table 5.4, shaded blue), was classified as probably damaging by both Human Diversity and Human Variation datasets: This variant is to be validated by Sanger sequencing and then further assessed for function using a suitable expression vector. Variants which alter the amino acid sequence at positions 46, 96, 125 and 131, were also predicted as 'probably damaging' by the Human Diversity panel but predicted as 'possibly damaging' by the Human Variation panel: These variants will be assessed in a second round of validation.

5.7.5 Repetitive DNA and sequencing

Repetitive DNA sequences have high homology or are identical to sequences elsewhere, accounting for around 50% of the human genome. These elements contribute to human evolution and may represent important phenomena in terms of biological function. Aligning repeats is technically challenging using Sanger sequences but more challenging with NGS reads because of the shorter read lengths and large volumes of data generated (Treangen and Salzberg, 2012). Assembly is therefore challenged by the intrinsic structure of the genome: Ignoring repeat sequences potentially introduces experimental bias which may result in over-interpretation of data.

Repetitive DNA sequences are abundant in the upstream risk-associated region 5′ to the *TNFSF4* gene as illustrated in Figure 5.4 of this chapter: LINE and SINE repetitive elements at the locus exhibited 100% sequence identity with >100 identical copies scattered elsewhere in the genome, causing difficulties in the alignment procedure. The 454 platform generates longer reads and the *TNFSF4* locus is small, so errors in the assembly of these sequences compared to a genome-wide assembly were greatly reduced. Even so, the section upstream of the *TNFSF4* gene corresponding to the 13.8kb amplicon (Figure 5.4) presented a major technical challenge and sequenced to high depth and uniformity in only three individuals. Adjacent amplicons generated an increased volume of better quality data, but with sporadic gaps in the majority of sequences for this part of chromosome 1q25.1. Historically, this section of the locus has generated poor quality sequencing and genotype data: There are a reduced number of identified variants owing to gaps in contigs in the Ensembl, NCBI and UCSC genome browsers.

5.7.6 Errors in PTS: false-assignment rate

PTS, and other multiplex approaches, exploit the capabilities of NGS well; however they also come with the risk of falsely assigning sequences to barcoding adapters. Rare variant analysis is compromised if sequence assignment is even slightly inaccurate. To avoid misalignment, the selection of barcoding adapters in

this study was particularly robust to sequencing error: Oligomers were highly distinguishable with at least three base changes required to transform them into another included adapter sequence: The false-assignment rate was calculated to be 0.27% for this study. A simple binomial error distribution model, as suggested by Meyer and colleagues (Meyer et al., 2008) also identified the edit distance of three to correspond to a false-assignment probability of $<10^{-7}$ for 8-nucleotide sequences. There was scope for cross-contamination between adaptors in the laboratory and during manufacture of the adapters: Sequential purification which re-used a HPLC column at the manufacture stage could have given low-level carry-over contamination (Meyer et al., 2008), however the aforementioned risk is low.

5.7.7 Limitations of PTS

One limitation of PTS arises through the use of the restriction enzyme *Srf*I, a rare cutter in mammalian genomes with a restriction site approximately every 150kb in the human genome. Sequence coverage immediately around a *Srf*I site is lost; although sequence loss is limited to the immediate sequence either side the site, as the 454 universal adaptors are added after restriction digestion. Methyltransferase methylation of CpG sites prior to adapter ligation eliminates the sequencing loss by preventing restriction from occurring within template DNA (Meyer et al., 2008).

5.7.8 454 sequencing errors

Assessing the quality and accuracy of the sequences generated by the Roche 454 GS-Titanium platform, different types of error might have been incorporated into the sequence during pyrosequencing. The 454 method does not call bases but instead calls light signals (flows). The length of brightness of a flow indicates the length of a run of identical bases (homopolymer), and homopolymer length is easy to mis-calibrate (URL: http://www.broadinstitute.org/crd/wiki/index.php/Homopolymer). Single base indels occur with considerable frequency both within and around homopolymer regions, can persist at high coverage and are the most common error in 454

sequencing (Kircher et al., 2012). Error rates vary for different versions of the 454 assembly programs newbler and runMapper and may also differ among runs.

Using an extensive dataset of Roche control DNA fragments, A. Gilles and colleagues (Gilles et al., 2011) identified insertions to be the most common errors (mean = 0.273% (0.269, 0.276)), followed by deletions (0.232% (0.229-0.235)); mismatches (0.022% (0.021-0.023)) and ambiguous base calls (0.007% (0.006-0.007)). Deep sequencing of multiple independent *TNFSF4* sequences should have corrected for random errors generated by the sequencer. The Gilles study found heterogeneous errors along the length of sequences tested. The distribution of error within each category did not fit a stochastic model: Although the majority of positions sequenced correctly, a few had error rates which exceeded 50%. For the results presented in this thesis, there was additional, unintentional, control for the aforementioned errors, by sequencing lupus individuals with two distinct haplotypes: Many novel variants clearly segregated with $TNFSF4_{\text{risk}}$ or $TNFSF4_{\text{non-risk.}}$ Thus, the novel polymorphisms identified were a condition of genotype and not a function of the sequencing platform.

5.7.9 Functional assessment of novel TNFSF4 variants

Validation by Sanger sequencing and genotyping of high-quality upstream and rare coding variants in our UK-European SLE cohorts will be undertaken before functional assessment of novel alleles, which might be associated with risk. The novel alleles may impact *TNFSF4* expression and provide further insight into lupus pathogenesis.

5.8 Further work

5.8.1 Targeted re-sequencing of multiple SLE risk-associated loci, prediction of threshold liability for SLE

A multi-locus genetic model which can fit complex disease is the liability threshold model. This assumes an additive effect at each risk locus and between

loci. The extensive genetic research into complex disease which are currentlynderway are revealing many novel genetic loci. The multi-locus model shifts the mean of the normal distribution of disease liability for each genotype class until it is so great that disease is the result. A sequencing study at the 30 or so SLE loci will be undertaken in a multi-ethnic cohort comprising 600 individuals, 150 from each population. The variants identified by a targeted sequencing study in African-American, East Asian, European, Hispanic and South Indian SLE individuals will increase the attributed proportion of heritability by increasing the number of true causal variants identified. These genotype data will give a better approximation of the threshold liability for the aforementioned ancestral groups in lupus.

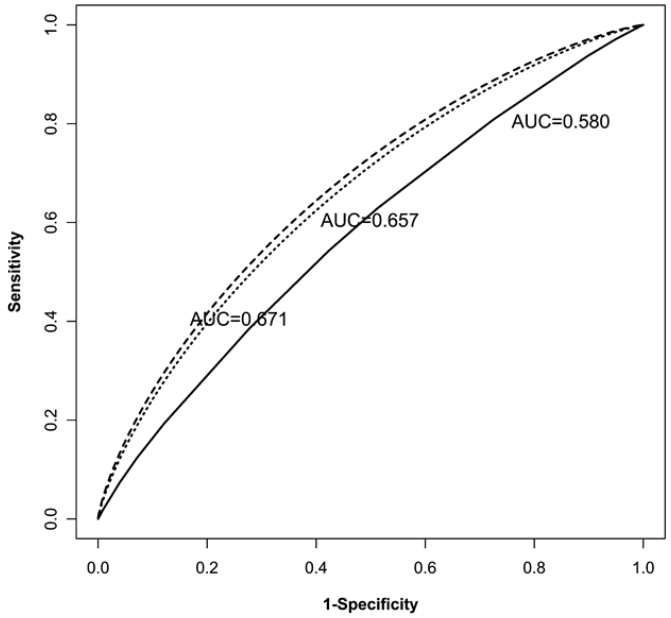
Novel and established genetic loci offer an important opportunity to improve clinical practice through the design of predictive genetic tests (Lu and Elston., 2008). Plotting population-specific receiver operating characteristic (ROC) curves using the test statistic from a likelihood ratio test would give a plot of sensitivity vs. specificity for multiple predictor, to help build a test for lupus (Figure 5.6). The tests using these curves would be asymptotically more powerful and would extend to situations where causal loci are linked or interact. ROC curves would also better establish whether the differences in SLE prevalence are attributed to heritable factors, or environmental and socioeconomic factors and access to care.

5.8.2 Targeted re-sequencing study of TNFSF4 RNA splice variants using PTS strategy

To investigate the transcriptional impact of the *TNFSF4* risk and non-risk haplotypes, RNA extracted from the individuals sequenced in this study will be sequenced, reverse transcribed and aligned to cDNA sequences stored in the Ensembl and UCSC browsers. This is to prevent misalignment of short reads which span exons to pseudogenes (pseudogenes are without junctions) so increase accuracy through *in situ* comparison. Transcriptional base modifications (TBMs) identified by comparison with the NGS DNA sequences will be mapped. TBMs

in coding regions are capable of influencing the translation of the protein, however modifications concentrated in AU-rich 3′ UTRs may function as mRNA editing targets (Rosenberg et al., 2011) which can regulate translation. The *TNFSF4* gene has an extended 3kb 3′UTR which should be investigated for TBMs. Quantifying the coverage of DNA and RNA reads and comparing them for each validated variant will address preferential expression of one allele relative to the other and the impact of allelic imbalance on expression of the *TNFSF4* gene.

Figure 5.6 Receiver Operating Characteristic Curves, exemplified with type 2 diabetes



ROC Curves for Type 2 Diabetes. The three lines in the plot from bottom to top correspond to the ROC curves of three type 2 diabetes predictive tests: the rebuilt existing predictive genetic test based on three SNP loci, a new predictive test combing the previously associated SNPs, four environmental factors, and four novel risk SNP loci and a improved new predictive test with five additional novel risk SNP loci. (Lu and Elston., 2008).

Chapter 6

Conclusion

The trans-ancestral fine-mapping of SNP variants presented in chapter 4 of this thesis has refined a risk signal at *TNFSF4* and mapped it in multiple independent populations*:* The SNP which demonstrates best evidence of association, rs2205960-T, is a common high-frequency variant which tags a risk-associated haplotype in all tested populations. This variant is the most associated variant in three independent populations (Europeans, Hispanics and African-American-Gullah). Rs2205960-T is also best-associated after fixed-effects meta-analysis and after conditional analysis of other key associated haplotype-tag SNPs. Assessing the regulatory potential of this variant; PWM binding data suggest rs2205960-T as the 8^{th} nucleotide of a motif which binds NF-κB p65 (RELA). Switching the rs2205960-T allele (risk) to G (non-risk), binding affinity for the motif is predicted to decrease by 10%. These data are supported by ChIP-seq data generated as part of the ENCODE project which find binding of NF-κB to the 346bp section of chromosome 1q25.1 encompassing rs2205960. These ChIP-seq data were generated in EBV-transformed LCLs.

At the time these doctoral studies were planned, the LD observed at the *TNFSF4* locus prevented delineation of specific causal contributors to SLE risk. The data presented in this thesis refine the signal to the proximal 5′ section of the locus. These data were generated from the best-available tag SNPs and proxies from the known repertoire of common *TNFSF4* SNPs. The rate at which novel SNPs are identified is rapidly decreasing as they near the total number. In contrast, as NGS technologies advance, structural variants, including small insertions and deletions, are being identified at increased frequency across the genome. The sequencing project described in this thesis increased the number of putative indels at *TNFSF4* by 4.7 fold. A proportion are unique to the risk haplotype and thus in strong LD with the best-associated variants identified in chapter 4. Many of the identified indels exhibit strong regulatory potential: Genotyping of the novel variations might find a better-associated or additional contributor to the aforementioned *TNFSF4* signal in lupus.

TNFSF4 variants associated with lupus are significantly more likely to be expression quantitative trait loci (eQTLs) than allele-frequency–matched SNPs spanning the locus: Presence of this polymorphism in a regulatory element makes it an attractive candidate for investigation as a cis-acting variant associated with transcript expression. Evidence from a genome-wide study supports the eQTL status of this variant. The cis-eQTL study in LCL samples from 777 female TwinsUK participants profiled common *TNFSF4* variants at the same time as *TNFSF4* transcript expression (Grundberg et al., 2010). In this study, association of RNA expression with >2 x 10^6 SNPs was tested by two-step mixed model–based score test. The identified cis-eQTLs were mapped to recombination hotspot intervals for likely independent regulatory effects. For each gene, the most significant SNP per hotspot interval was selected and LD filtering performed. The top cis-eQTL in the LD bin, for the probe located in the 3´UTR of *TNFSF4 (*ILMN_2089875), was rs2205960-T (P=3.75 x 10^{-4})

A limitation of the aforementioned eQTL data is that a large proportion of the available nucleotides were not tested for association with the *TNFSF4* transcripts. These nucleotides include newly identified rare coding variants and indels identified in chapter 5 of this thesis. In a refined eQTL study, the novel *TNFSF4* variants, in addition to the causal contributors identified in chapter 4, would be assayed at the same time as the three main transcripts of the gene (Cookson et al., 2009; Michaelson et al., 2009). Furthermore, the experiment would be repeated in different immune cell types, including activated B-cells, dendritic cells and CD4 T-cells. The gene is inducible and expression of the transcript is likely to be different to the expression data reported in the aforementioned eQTL study. Annotating SNPs with a score reflecting their eQTL potential could help clarify the nature of the mechanism driving the associations. Identifying trans-eQTLs which are associated with expression of *TNFSF4* transcripts might additionally prove informative for the gene-expression pathways which underlie disease pathogenesis (Kadota et al., 2007).

Data presented in this thesis identify an independent, population-specific risk variant unique to individuals of Amerindian-descent. This *TNFSF4* intron 1 allele, rs16845607-A, is annotated by ENCODE to sit within a regulatory region. A limitation of the Amerindian/Hispanic SLE data is that they were not imputed as part of this project, thus a large proportion of the available nucleotides were not tested for association. At the time of data generation, access to AMR phased reference haplotypes sequenced as part of the 1000 Genomes project was limited. For 80% of the *TNFSF4* locus, these haplotypes were inferred from low-coverage (x2-6) sequence data. The pilot release of the pilot phase of low coverage 1000 Genomes data were enriched for sequencing artefacts, thus unsuited as an imputation reference panel to use with these genotypes at the time.

Prior to functional assessment of the Hispanic rs16845607-A risk variant, a more complete set of the variant nucleotides will be tested for association: A first stage will be imputation of the Amerindian/Hispanic SLE-control genotypes, presented in chapter 4. These data will be imputed using in excess of 2000 phased chromosomes which form the 1000 Genomes Phase I integrated variant set v3 as a reference panel. The reference chromosomes would be included in an alternative imputation framework which captures unexpected allele sharing amongst populations: These data would be augmented for selecting a higher quality, more conservative variant subset of bi-allelic indels and SNPs. A subset of sequencing artefacts, including problem indels, has been removed from the 1000 Genomes integrated variant set v3 panel mentioned above: Power to detect SNPs present at a frequency of 1% is 99.3% using these data; thus an increased number of genetic variations at *TNFSF4* would be sampled. Preliminary runs using these data suggest reduced standard error for the beta coefficient of the probabilistic genotypes. This reflects greater imputation certainty, requisite for reliable imputation of populations with cryptic substructure.

If association testing of the imputed Hispanic data confirms the rs16045607-A allele as a causal contributor to risk, further exploration of the mechanism by

which it contributes to SLE pathogenesis is justified. Evaluation would include parallel experiments with the aforementioned associated variant rs2205960. Interactions of these variants with proteins in the same genomic regions could be assessed *in vivo:* Variation of chromatin state has been examined in pedigrees, and these studies demonstrate allele-specific clustering (Listgarten et al., 2010). Histone marks, transcription factors or other chromatin-associated proteins which bind these variants would be assessed by the chromatin immune precipitation (ChIP) assay in individuals selected by their genotype for increased heritability; a relevant group would include risk allele homozygote cases. In the first instance, ChIP would explore the regulation of *TNFSF4:* ENCODE data could be used to inform these preliminary investigations.

The sequencing data presented in chapter 5 are currently in validation: Many markers were identified that clearly segregated with genotype subgroup. A substantial proportion of these identified variants were predicted *in silico* to have regulatory activity. Sequencing also identified 'rare' non synonymous variants in the gene which map to regulator regions: Although the majority of these are likely to be sequencing artefacts, several putative variants, including the missense substitution at position 171422287 (Cys181Trp, Table 5.4 blue-shaded) warrant further assessment.

Preliminary functional data presented in chapter 3 suggest a difference in gene expression between $TNFSF4_{\text{risk}}$ and $TNFSF4_{\text{non-risk}}$ homozygote individuals, though these data have clear limitations due to sample heterogeneity and disease activity. The chromatin landscape of the *TNFSF4* gene may be influenced by the risk or non-risk signal. Histone modifications, which are likely integral to this landscape, are linked to gene expression through orchestration of DNA-based biological processes. Epigenetic features including the di and tri-methylations of Histone 3 lysine 4 (H3K4) are associated with gene activation, whereas tri-methylation of H3K27 and H3K9 are implicated in transcriptional repression through heterochromatin formation and gene silencing (Barski et al., 2007). High-

specificity, well-validated ChIP antibodies to these modifications are available, as are antibodies to NF-κBp65 (RELA), and B-ATF and PU1 transcription factors.

The functional data presented in this thesis identified statistically insignificant differences between *TNFSF4*$_{risk}$ and *TNFSF4*$_{non\text{-}risk}$ homozygote cell lines, but significance differences between gated PBMCs of the same genotype. The discordant findings may have resulted from background noise due to EBV-transformation or some other systematic effect in the cell lines, or because the influence the genetic variants have on expression is accelerated on a disease background. To avoid confounding effects of disease activity in lupus patients, which cause spurious or missed associations of variants with gene expression, ChIP and eQTL analysis would be undertaken in populations of *TNFSF4*-expressing cells from control individual participants, such as those from the UK Twins Registry of 11,000 twins (courtesy of Professor Tim Spector, King's College London). The GWAS and phenotype data available for this cohort would direct sample selection. If these studies prove informative, the experiments would be repeated to evaluate purified lymphocyte subsets from lupus cases.

Bibliography

1000 Genomes Project Consortium., Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E. and McVean, G.A. (2010). A map of human genome variation from population-scale sequencing. Nature *28*, 1061-73

1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. Nature. (2012). *491*, 56-65

Adrianto, I., Wen, F., Templeton, A., Wiley, G., King, J.B., Lessard, C.J., Bates, J.S., Hu, Y., Kelly, J.A., Kaufman, K.M., Guthridge, J.M., Alarcon-Riquelme, M.E., Anaya, J.M., Bae, S.C., Bang, S.Y., Boackle, S.A., Brown, E.E., Petri, M.A., Gallant, C., Ramsey-Goldman, R., Reveille, J.D., Vila, L.M., Criswell, L.A., Edberg, J.C., Freedman, B.I., Gregersen, P.K., Gilkeson, G.S., Jacob, C.O., James, J.A., Kamen, D.L., Kimberly, R.P., Martin, J., Merrill, J.T., Niewold, T.B., Park, S.Y., Pons-Estel, B.A., Scofield, R.H., Stevens, A.M., Tsao, B.P., Vyse, T.J., Langefeld, C.D., Harley, J.B., Moser, K.L., Webb, C.F., Humphrey, M.B., Montgomery, C.G. and Gaffney, P.M. (2011). Association of a functional variant downstream of *TNFAIP3* with systemic lupus erythematosus. Nat. Genet. *43*, 253-258.

Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S. and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat. Methods *7,* 248-249.

Alarcon, G.S., McGwin, G.Jr., Petri, M., Reveille, J.D., Ramsey-Goldman, R. and Kimberly, R.P. (2002). Baseline characteristics of a multiethnic lupus cohort: PROFILE. Lupus *11*, 95-101.

Alarcon-Segovia, D., Alarcon-Riquelme, M.E., Cardiel, M.H., Caeiro, F., Massardo, L., Villa, A.R. and Pons-Estel, B.A. (2005). Familial aggregation of systemic lupus erythematosus, rheumatoid arthritis, and other autoimmune diseases in 1, 177 lupus patients from the GLADEL cohort. Arthritis Rheum. *52*, 1138-1147.

Alba, P., Bento, L., Cuadrado, M.J., Karim, Y., Tungekar, M.F., Abbs, I., Khamashta, M.A., D'Cruz, D. and Hughes, G.R. (2003). Anti-dsDNA, Anti-Sm antibodies, and the lupus anticoagulant: Significant factors associated with lupus nephritis. Ann. Rheum. Dis. *62*, 556-560.

Altshuler, D., Pollara, V.J., Cowles, C.R., Van Etten, W.J., Baldwin, J., Linton, L. and Lander, E.S. (2000). A SNP map of the human genome generated by reduced representation shotgun sequencing. Nature *407*, 513-516.

Amos, B., Ward, F.E., Zmijewski, C.M., Hattler, B.G. and Seigler, H.F. (1968). Graft donor selection based upon single locus (haplotype) analysis within families. Transplantation *6,* 524-34

Ardlie, K.G., Lunetta, K.L. and Seielstad M. (2002). Testing for population subdivision and association in four case-control studies. Am J Hum Genet. *71,* 304-311

Auton, A. (2007). The Estimation of Recombination Rates from Population Genetic Data. Published D.Phil thesis. Hertford College, University of Oxford

Auton, A. and McVean, G. (2007). Recombination rate estimation in the presence of hotspots. Genome Res. *17*, 1219-1227.

Bae, S.C., Fraser, P. and Liang, M.H. (1998). The epidemiology of systemic lupus erythematosus in populations of African ancestry: A critical review of the "prevalence gradient hypothesis". Arthritis Rheum. *41*, 2091-2099.

Bain, S.C., Todd, J.A. and Barnett, A.H. (1990). The British Diabetic Association-Warren repository. Autoimmunity *7,* 83-85.

Balding, D. J. A tutorial on statistical methods for population association studies. Nat Rev Genet. *7*, 781-791.

Ballestar, E. (2011). Epigenetic alterations in autoimmune rheumatic diseases. Nat. Rev. Rheumatol. *7*, 263-271.

Barrett, J.C., Fry, B., Maller, J. and Daly, M.J. (2005). Haploview: Analysis and visualization of LD and haplotype maps. Bioinfomatics *21*, 263-265.

Barrett, J.C., Hansoul, S., Nicolae, D.L., Cho, J.H., Duerr, R.H., Rioux, J.D., Brant, S.R., Silverberg, M.S., Taylor, K.D., Barmada, M.M., Bitton, A., Dassopoulos, T., Datta, L.W., Green, T., Griffiths, A.M., Kistner, E.O., Murtha, M.T., Regueiro, M.D., Rotter, J.I., Schumm, L.P., Steinhart, A.H., Targan, S.R., Xavier, R.J., Libioulle, C., Sandor, C., Lathrop, M., Belaiche, J., Dewit, O., Gut, I., Heath, S., Laukens, D., Mni, M., Rutgeerts, P., Van, G.A., Zelenika, D., Franchimont, D., Hugot, J.P., de, V.M., Vermeire, S., Louis, E., Cardon, L.R., Anderson, C.A., Drummond, H., Nimmo, E., Ahmad, T., Prescott, N.J., Onnie, C.M., Fisher, S.A., Marchini, J., Ghori, J., Bumpstead, S., Gwilliam, R., Tremelling, M., Deloukas, P., Mansfield, J., Jewell, D., Satsangi, J., Mathew, C.G., Parkes, M., Georges, M. and Daly, M.J. (2008). Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat. Genet. *40*, 955-962.

Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I. and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. Cell *129*, 823-837.

Bentley, D.R. (2006). Whole-genome re-sequencing. Curr. Opin. Genet. Dev. *16*, 545-552.

Berg, I.L., Neumann, R., Lam, K.W., Sarbajna, S., Odenthal-Hesse, L., May, C.A. and Jeffreys, A.J. (2010). PRDM9 variation strongly influences recombination hot-spot activity and meiotic instability in humans. Nat. Genet. *42*, 859-863.

Biesecker, L.G. (2010). Exome sequencing makes medical genomics a reality. Nat. Genet. *42*, 13-14.

Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigo, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T., Thurman, R.E., Kuehn, M.S., Taylor, C.M., Neph, S., Koch, C.M., Asthana, S., Malhotra, A., Adzhubei, I., Greenbaum, J.A., Andrews, R.M., Flicek, P., Boyle, P.J., Cao, H., Carter, N.P., Clelland, G.K., Davis, S., Day, N., Dhami, P., Dillon, S.C., Dorschner, M.O., Fiegler, H., Giresi, P.G., Goldy, J., Hawrylycz, M., Haydock, A., Humbert, R., James, K.D., Johnson, B.E., Johnson, E.M., Frum, T.T., Rosenzweig, E.R., Karnani, N., Lee, K., Lefebvre, G.C., Navas, P.A., Neri,

F., Parker, S.C., Sabo, P.J., Sandstrom, R., Shafer, A., Vetrie, D., Weaver, M., Wilcox, S., Yu, M., Collins, F.S., Dekker, J., Lieb, J.D., Tullius, T.D., Crawford, G.E., Sunyaev, S., Noble, W.S., Dunham, I., Denoeud, F., Reymond, A., Kapranov, P., Rozowsky, J., Zheng, D., Castelo, R., Frankish, A., Harrow, J., Ghosh, S., Sandelin, A., Hofacker, I.L., Baertsch, R., Keefe, D., Dike, S., Cheng, J., Hirsch, H.A., Sekinger, E.A., Lagarde, J., Abril, J.F., Shahab, A., Flamm, C., Fried, C., Hackermuller, J., Hertel, J., Lindemeyer, M., Missal, K., Tanzer, A., Washietl, S., Korbel, J., Emanuelsson, O., Pedersen, J.S., Holroyd, N., Taylor, R., Swarbreck, D., Matthews, N., Dickson, M.C., Thomas, D.J., Weirauch, M.T., Gilbert, J., Drenkow, J., Bell, I., Zhao, X., Srinivasan, K.G., Sung, W.K., Ooi, H.S., Chiu, K.P., Foissac, S., Alioto, T., Brent, M., Pachter, L., Tress, M.L., Valencia, A., Choo, S.W., Choo, C.Y., Ucla, C., Manzano, C., Wyss, C., Cheung, E., Clark, T.G., Brown, J.B., Ganesh, M., Patel, S., Tammana, H., Chrast, J., Henrichsen, C.N., Kai, C., Kawai, J., Nagalakshmi, U., Wu, J., Lian, Z., Lian, J., Newburger, P., Zhang, X., Bickel, P., Mattick, J.S., Carninci, P., Hayashizaki, Y., Weissman, S., Hubbard, T., Myers, R.M., Rogers, J., Stadler, P.F., Lowe, T.M., Wei, C.L., Ruan, Y., Struhl, K., Gerstein, M., Antonarakis, S.E., Fu, Y., Green, E.D., Karaoz, U., Siepel, A., Taylor, J., Liefer, L.A., Wetterstrand, K.A., Good, P.J., Feingold, E.A., Guyer, M.S., Cooper, G.M., Asimenos, G., Dewey, C.N., Hou, M., Nikolaev, S., Montoya-Burgos, J.I., Loytynoja, A., Whelan, S., Pardi, F., Massingham, T., Huang, H., Zhang, N.R., Holmes, I., Mullikin, J.C., Ureta-Vidal, A., Paten, B., Seringhaus, M., Church, D., Rosenbloom, K., Kent, W.J., Stone, E.A., Batzoglou, S., Goldman, N., Hardison, R.C., Haussler, D., Miller, W., Sidow, A., Trinklein, N.D., Zhang, Z.D., Barrera, L., Stuart, R., King, D.C., Ameur, A., Enroth, S., Bieda, M.C., Kim, J., Bhinge, A.A., Jiang, N., Liu, J., Yao, F., Vega, V.B., Lee, C.W., Ng, P., Shahab, A., Yang, A., Moqtaderi, Z., Zhu, Z., Xu, X., Squazzo, S., Oberley, M.J., Inman, D., Singer, M.A., Richmond, T.A., Munn, K.J., Rada-Iglesias, A., Wallerman, O., Komorowski, J., Fowler, J.C., Couttet, P., Bruce, A.W., Dovey, O.M., Ellis, P.D., Langford, C.F., Nix, D.A., Euskirchen, G., Hartman, S., Urban, A.E., Kraus, P., Van, C.S., Heintzman, N., Kim, T.H., Wang, K., Qu, C., Hon, G., Luna, R., Glass, C.K., Rosenfeld, M.G., Aldred, S.F., Cooper, S.J., Halees, A., Lin, J.M., Shulha, H.P., Zhang, X., Xu, M., Haidar, J.N., Yu, Y., Ruan, Y., Iyer, V.R., Green, R.D., Wadelius, C., Farnham, P.J., Ren, B., Harte, R.A., Hinrichs, A.S., Trumbower, H. and Clawson, H. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature *447*, 799-816.

Bombardier, C., Gladman, D.D., Urowitz, M.B., Caron, D. and Chang, C.H. (1992). Derivation of the SLEDAI. A disease activity index for lupus patients. The Committee on Prognosis Studies in SLE. Arthritis Rheum. *35*, 630-640.

Botstein, D. and Risch, N. (2003). Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nat. Genet. *33*, 228-237.

Botto, M., Dell'Agnola, C., Bygrave, A.E., Thompson, E.M., Cook, H.T., Petry, F., Loos, M., Pandolfi, P.P. and Walport, M.J. (1998). Homozygous *C1q* deficiency causes glomerulonephritis associated with multiple apoptotic bodies. Nat. Genet. *19*, 56-59.

Bowness, P., Davies, K.A., Norsworthy, P.J., Athanassiou, P., Taylor-Wiedeman, J., Borysiewicz, L.K., Meyer, P.A. and Walport, M.J. (1994). Hereditary C1q deficiency and systemic lupus erythematosus. QJM. *87*, 455-464.

Broman, K.W., Murray, J.C., Sheffield, V.C., White, R.L. and Weber, J.L. (1998). Comprehensive human genetic maps: individual and sex-specific variation in recombination. Am. J. Hum. Genet. *63*, 861-869.

Buglio, D., Khaskhely, N.M., Voo, K.S., Martinez-Valdez, H., Liu, Y.J. and Younes, A. (2011). HDAC11 plays an essential role in regulating OX40 ligand expression in Hodgkin lymphoma. Blood *117*, 2910-2917.

Burgos, P.I., McGwin, G., Jr., Pons-Estel, G.J., Reveille, J.D., Alarcon, G.S. and Vila, L.M. (2011). US patients of Hispanic and African ancestry develop lupus nephritis early in the disease course: data from LUMINA, a multiethnic US cohort (LUMINA LXXIV). Ann. Rheum. Dis. *70*, 393-394.

Capriotti, E., Calabrese, R. and Casadio, R. (2006). Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. Bioinformatics. *22*, 2729-2734.

Chambers, S.A., Charman, S.C., Rahman, A. and Isenberg, D.A. (2007). Development of additional autoimmune diseases in a multiethnic cohort of patients with systemic lupus erythematosus with reference to damage and mortality. Ann. Rheum. Dis. *66*, 1173-1177.

Chan, O.T. and Shlomchik, M.J. (2000). Cutting edge: B-cells promote CD8+ T-cell activation in MRL-Fas(lpr) mice independently of MHC class I antigen presentation. J. Immunol. *164*, 1658-1662.

Chun, B.C. and Bae, S.C. (2005). Mortality and cancer incidence in Korean patients with systemic lupus erythematosus: results from the Hanyang lupus cohort in Seoul, Korea. Lupus *14*, 635-638.

Clark, A.G., Wang, X. and Matise, T. (2010). Contrasting methods of quantifying fine structure of human recombination. Annu. Rev. Genomics Hum. Genet. *11*, 45-64.

Compaan, D.M. and Hymowitz, S.G. (2006). The crystal structure of the costimulatory OX40-OX40L complex. Structure. *14*, 1321-1330.

Conrad, D.F., Keebler, J.E., DePristo, M.A., Lindsay, S.J., Zhang, Y., Casals, F., Idaghdour, Y., Hartl, C.L., Torroja, C., Garimella, K.V., Zilversmit, M., Cartwright, R., Rouleau, G.A., Daly, M., Stone, E.A., Hurles, M.E. and Awadalla, P. (2011). Variation in genome-wide mutation rates within and between human families. Nat. Genet. *43*, 712-714.

Cookson, W., Liang, L., Abecasis, G., Moffatt, M. and Lathrop, M. (2009). Mapping complex disease traits with global gene expression. Nat. Rev. Genet. *10*, 184-194.

Correa, P.A., Molina, J.F., Pinto, L.F., Arcos-Burgos, M., Herrera, M. and Anaya, J.M. (2003). *TAP1* and *TAP2* polymorphisms analysis in northwestern Colombian patients with systemic lupus erythematosus. Ann. Rheum. Dis. *62*, 363-365.

Cortes, S., Chambers, S., Jeronimo, A. and Isenberg, D. (2008). Diabetes mellitus complicating systemic lupus erythematosus - analysis of the UCL lupus cohort and review of the literature. Lupus *17*, 977-980.

Croft, M. (2010). Control of immunity by the TNFR-related molecule OX40 (CD134). Annu. Rev. Immunol. *28*, 57-78.

Cunninghame Graham, D.S., Akil, M. and Vyse, T.J. (2007). Association of polymorphisms across the tyrosine kinase gene, TYK2 in UK SLE families. Rheumatology. (Oxford) *46*, 927-930.

Cunninghame Graham, D.S., Wong, A.K., McHugh, N.J., Whittaker, J.C., and Vyse, T.J. (2006). Evidence for unique association signals in SLE at the CD28-CTLA4-ICOS locus in a family-based study. Hum. Mol. Genet. *15*, 3195-3205.

Cunninghame Graham, D.S., 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., and Vyse, T.J. (2008). Polymorphism at the TNF superfamily gene *TNFSF4* confers susceptibility to systemic lupus erythematosus. Nat. Genet. *40*, 83-89.

D'Angelo, G.M., Kamboh, M.I. and Feingold, E. (2010). A Likelihood-Based Approach for Missing Genotype Data. Hum. Hered. *69*, 171-183.

Daly, M.J., Rioux, J.D., Schaffner, S.F., Hudson, T.J. and Lander, E.S. (2001). High-resolution haplotype structure in the human genome. Nat. Genet. *29*, 229-232.

Danchenko, N., Satia, J.A. and Anthony, M.S. (2006). Epidemiology of systemic lupus erythematosus: a comparison of worldwide disease burden. Lupus *15,* 308-318.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., McVean, G. and Durbin, R. (2011). The variant call format and VCFtools. Bioinformatics. *27*, 2156-2158.

de Bakker, P.I., Yelensky, R., Pe'er, I., Gabriel, S.B., Daly, M.J. and Altshuler, D. (2005). Efficiency and power in genetic association studies. Nat. Genet. *37*, 1217-1223.

de Bakker, P.I., Burtt, N.P., Graham, R.R., Guiducci, C., Yelensky, R., Drake, J.A., Bersaglieri, T., Penney, K.L., Butler, J., Young, S., Onofrio, R.C., Lyon, H.N., Stram, D.O., Haiman, C.A., Freedman, M.L., Zhu, X., Cooper, R., Groop, L., Kolonel, L.N., Henderson, B.E., Daly, M.J., Hirschhorn, J.N. and Altshuler, D. (2006). Transferability of tag SNPs in genetic association studies in multiple populations. Nat. Genet. *38*, 1298-1303.

Deapen, D., Escalante, A., Weinrib, L., Horwitz, D., Bachman, B., Roy-Burman, P., Walker, A. and Mack, T.M. (1992). A revised estimate of twin concordance in systemic lupus erythematosus. Arthritis Rheum. *35*, 311-318.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del, A.G., Rivas, M.A., Hanna, M., McKenna, A., Fennell, T.J., Kernytsky, A.M., Sivachenko, A.Y., Cibulskis, K., Gabriel, S.B., Altshuler, D. and Daly, M.J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. *43*, 491-498.

Devlin, B. and Risch, N. (1995). A comparison of linkage disequilibrium measures for fine-scale mapping. Genomics *29*, 311-322.

Devlin B and Roeder K (1999). Genomic control for association studies. Biometrics. *55,* 997-1004

Devlin B, Bacanu S.A. and Roeder K (2004). Genomic Control to the extreme. Nat Genet. *36*, 1129-1130

Dhanya, R., Kishore, V.C., Sudha, K.C., Sreekumar, K. and Joseph, R. (2008). Ground state and excited state dipole moments of alkyl substituted para-nitroaniline derivatives. Spectrochim. Acta A Mol. Biomol. Spectrosc. *71*, 1355-1359.

Dib, C., Faure, S., Fizames, C., Samson, D., Drouot, N., Vignal, A., Millasseau, P., Marc, S., Hazan, J., Seboun, E., Lathrop, M., Gyapay, G., Morissette, J. and Weissenbach, J. (1996). A comprehensive genetic map of the human genome based on 5, 264 microsatellites. Nature *380,* 152-154.

Edberg, J.C., Langefeld, C.D., Wu, J., Moser, K.L., Kaufman, K.M., Kelly, J., Bansal, V., Brown, W.M., Salmon, J.E., Rich, S.S., Harley, J.B. and Kimberly, R.P. (2002). Genetic linkage and association of Fcgamma receptor IIIA (CD16A) on chromosome 1q23 with human systemic lupus erythematosus. Arthritis Rheum. *46*, 2132-2140.

Eichler, E.E., Flint, J,. Gibson, G., Kong, A., Leal, S.M., Moore, J.H., and Nadeau, J.H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. Nature Reviews Genetics *11*, 446-450.

ENCODE Project Consortium., Bernsteinm, BE., Birney. E., Dunham. I., Green, ED., Gunter, C. and Snyder M. (2012). Nature *489,* 57-74

Engler, J.B., Undeutsch, R., Kloke, L., Rosenberger, S., Backhaus, M., Schneider, U., Egerer, K., Dragun, D., Hofmann, J., Huscher, D., Burmester, G.R., Humrich, J.Y., Enghard, P. and Riemekasten, G. (2011). Unmasking of autoreactive CD4 T-cells by depletion of CD25 regulatory T-cells in systemic lupus erythematosus. Ann. Rheum. Dis. *70*, 2176-2183.

Ewing, B. and Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. *8*, 186-194.

Fanciulli, M., Norsworthy, P.J., Petretto, E., Dong, R., Harper, L., Kamesh, L., Heward, J.M., Gough, S.C., de, S.A., Blakemore, A.I., Froguel, P., Owen, C.J., Pearce, S.H., Teixeira, L., Guillevin, L., Graham, D.S., Pusey, C.D., Cook, H.T., Vyse, T.J. and Aitman, T.J. (2007). *FCGR3B* copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. Nat. Genet. *39*, 721-723.

Fearnhead, P. (2006). SequenceLDhot: detecting recombination hotspots. Bioinformatics. *22,* 3061-3066.

Fernandez, M., Alarcon, G.S., Calvo-Alen, J., Andrade, R., McGwin, G., Jr., Vila, L.M. and Reveille, J.D. (2007). A multiethnic, multicenter cohort of patients with systemic lupus erythematosus (SLE) as a model for the study of ethnic disparities in SLE. Arthritis Rheum. *57*, 576-584.

Fernando, M.M., Stevens, C.R., Walsh, E.C., De Jager, P.L., Goyette, P., Plenge, R.M., Vyse, T.J. and Rioux, J.D. (2008). Defining the role of the MHC in autoimmunity: a review and pooled analysis. PLoS. Genet. *4,* e1000024.

Fernando, M.M., Stevens, C.R., Sabeti, P.C., Walsh, E.C., McWhinnie, A.J., Shah, A., Green, T., Rioux, J.D. and Vyse, T.J. (2007). Identification of two independent risk factors for lupus within the MHC in United Kingdom families. PLoS. Genet. *3,* e192.

Feuerer, M., Hill, J.A., Mathis, D. and Benoist, C. (2009). Foxp3+ regulatory T-cells: differentiation, specification, subphenotypes. Nat. Immunol. *10,* 689-695.

Foster, M.W. and Sharp, R.R. (2004). Beyond race: towards a whole-genome perspective on human populations and genetic variation. Nat. Rev. Genet. *5*, 790-796.

Frazer, K.A., Murray, S.S., Schork, N.J., Topol, E.J. (2009). Human genetic variation and its contribution to complex traits. Nat Rev Genet. *10*, 241-251.

Frese, S. and Diamond, B. (2011). Structural modification of DNA-a therapeutic option in SLE? Nat. Rev. Rheumatol. *7*, 733-738.

Friedman, T.B., Liang, Y., Weber, J.L., Hinnant, J.T., Barber, T.D., Winata, S., Arhya, I.N. and Asher, J.H., Jr. (1995). A gene for congenital, recessive deafness *DFNB3* maps to the pericentromeric region of chromosome 17. Nat. Genet. *9*, 86-91.

Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S.N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E.S., Daly, M.J. and Altshuler, D. (2002). The structure of haplotype blocks in the human genome. Science *296*, 2225-2229.

Delgado-Vega, A.M., Abelson, A.K., Sanchez, E., Witte, T., D'Alfonso, S., Galeazzi, M., Jimenez-Alonso, J., Pons-Estel, B.A., Martin, J. and Alarcon-Riquelme, M.E. (2009). Replication of the *TNFSF4* (OX40L) promoter region association with systemic lupus erythematosus. Genes Immun. *10*, 248-253.

Gardener, H., Beecham, A., Cabral, D., Yanuck, D., Slifer, S., Wang, L., Blanton, S.H., Sacco, R.L., Juo, S.H. and Rundek, T. (2011). Carotid plaque and candidate genes related to inflammation and endothelial function in Hispanics from northern Manhattan. Stroke *42*, 889-896.

Gilkeson, G., James, J., Kamen, D., Knackstedt, T., Maggi, D., Meyer, A. and Ruth, N. (2011). The United States to Africa lupus prevalence gradient revisited. Lupus *20*, 1095-1103.

Gilles, A., Meglecz, E., Pech, N., Ferreira, S., Malausa, T. and Martin, J.F. (2011). Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. BMC. Genomics *12*, 245.

Graham, R.R., Kozyrev, S.V., Baechler, E.C., Reddy, M.V., Plenge, R.M., Bauer, J.W., Ortmann, W.A., Koeuth, T., Gonzalez Escribano, M.F., Pons-Estel, B., Petri, M., Daly, M., Gregersen, P.K., Martin, J., Altshuler, D., Behrens, T.W. and Alarcon-Riquelme, M.E. (2006). A common haplotype of interferon regulatory factor 5 (*IRF5*) regulates splicing and expression and is associated with increased risk of systemic lupus erythematosus. Nat. Genet. *38,* 550-555.

Graham, R.R., Ortmann, W.A., Langefeld, C.D., Jawaheer, D., Selby, S.A., Rodine, P.R., Baechler, E.C., Rohlf, K.E., Shark, K.B., Espe, K.J., Green, L.E., Nair, R.P., Stuart, P.E., Elder, J.T., King, R.A., Moser, K.L., Gaffney, P.M., Bugawan, T.L., Erlich, H.A., Rich, S.S., Gregersen, P.K. and Behrens, T.W. (2002). Visualizing human leukocyte antigen class II risk haplotypes in human systemic lupus erythematosus. Am. J. Hum. Genet. *71*, 543-553.

Graham, R.R., Cotsapas, C., Davies, L., Hackett, R., Lessard, C.J., Leon, J.M., Burtt, N.P., Guiducci, C., Parkin, M., Gates, C., Plenge, R.M., Behrens, T.W., Wither, J.E., Rioux, J.D., Fortin, P.R., Graham, D.C., Wong, A.K., Vyse, T.J., Daly, M.J., Altshuler, D., Moser, K.L. and Gaffney, P.M. (2008). Genetic variants near *TNFAIP3* on 6q23 are associated with systemic lupus erythematosus. Nat. Genet. *40*, 1059-1061.

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., Bauer, J.W., Gillett, C., Burtt, N., Cunninghame Graham, D.S., Onofrio, R., Petri, M., Gunnarsson, I., Svenungsson, E., Ronnblom, L., Nordmark, G., Gregersen, P.K., Moser, K., Gaffney, P.M., Criswell, L.A., Vyse, T.J., Syvanen, A.C., Bohjanen, P.R., Daly, M.J., Behrens, T.W. and Altshuler, D. (2007). Three functional variants of IFN regulatory factor 5 (*IRF5)* define risk and protective haplotypes for human lupus. Proc. Natl. Acad. Sci. U. S. A *104*, 6758-6763.

Gravel, S., Henn, B.M., Gutenkunst, R.N., Indap, A.R., Marth, G.T., Clark, A.G., Yu, F., Gibbs, R.A. and Bustamante, C.D. (2011). Demographic history and rare allele sharing among human populations. Proc. Natl. Acad. Sci. U. S. A *108*, 11983-11988.

Green, RE., Krause, J., Ptak, SE., Briggs, AW., Ronan, MT., Simons, JF., Du, L., Egholm, M., Rothberg, JM., Paunovic, M. and, Pääbo, S. (2006) Analysis of one million base pairs of Neanderthal DNA. Nature *16*, 330-336

Gri, G., Piconese, S., Frossi, B., Manfroi, V., Merluzzi, S., Tripodo, C., Viola, A., Odom, S., Rivera, J., Colombo, M.P. and Pucillo, C.E. (2008). CD4+CD25+ regulatory T-cells suppress masT-cell degranulation and allergic responses through OX40-OX40L interaction. Immunity *29,* 771-781.

Grundberg, E., Small, KS., Hedman, ÅK., Nica, AC., Buil, A., Keildson, S., Bell, JT., Yang, TP., Meduri, E., Barrett, A., Nisbett, J., Sekowska, M., Wilk, A., Shin, SY., Glass, D., Travers, M., Min, JL., Ring, S., Ho, K., Thorleifsson, G., Kong, A., Thorsteindottir, U., Ainali, C., Dimas, AS., Hassanali, N., Ingle, C., Knowles, D., Krestyaninova, M., Lowem, CE., Di Meglio, P., Montgomery, SB., Parts, L., Potter, S., Surdulescu, G., Tsaprouni, L., Tsoka, S., Bataille, V., Durbin, R., Nestle, FO., O'Rahilly, S., Soranzo, N., Lindgren, CM., Zondervan, KT., Ahmadi, KR., Schadt EE., Stefansson K., Smith, GD., McCarthy, MI., Deloukas, P., Dermitzakis, E.T. and Spector, T.D. Multiple Tissue Human .Expression Resource (MuTHER) Consortium. (2012). Mapping cis- and trans-regulatory effects across multiple tissues in twins. Nat Genet. *44,* 1084-1089.

Guo, S.W. (1998). Inflation of sibling recurrence-risk ratio, due to ascertainment bias and/or overreporting. Am J Hum Genet. *63*, 252-258.

Guzman, J., Cardiel, M.H., Arce-Salinas, A., Sanchez-Guerrero, J. and Alarcon-Segovia, D. (1992). Measurement of disease activity in systemic lupus erythematosus. Prospective validation of 3 clinical indices. J. Rheumatol. *19*, 1551-1558.

Halder, I., Shriver, M., Thomas, M., Fernandez, J.R. and Frudakis, T. (2008). A panel of ancestry informative markers for estimating individual biogeographical ancestry and admixture from four continents: utility and applications. Hum. Mutat. *29*, 648-658.

Han, J.W., Zheng, H.F., Cui, Y., Sun, L.D., Ye, D.Q., Hu, Z., Xu, J.H., Cai, Z.M., Huang, W., Zhao, G.P., Xie, H.F., Fang, H., Lu, Q.J., Xu, J.H., Li, X.P., Pan, Y.F., Deng, D.Q., Zeng, F.Q., Ye, Z.Z., Zhang, X.Y., Wang, Q.W., Hao, F., Ma, L., Zuo, X.B., Zhou, F.S., Du, W.H., Cheng, Y.L., Yang, J.Q., Shen, S.K., Li, J., Sheng, Y.J., Zuo, X.X., Zhu, W.F., Gao, F., Zhang, P.L., Guo, Q., Li, B., Gao, M., Xiao, F.L., Quan, C., Zhang, C., Zhang, Z., Zhu, K.J., Li, Y., Hu, D.Y., Lu, W.S., Huang, J.L., Liu, S.X., Li, H., Ren, Y.Q., Wang, Z.X., Yang, C.J., Wang, P.G., Zhou, W.M., Lv, Y.M., Zhang, A.P., Zhang, S.Q., Lin, D., Li, Y., Low, H.Q., Shen, M., Zhai, Z.F., Wang, Y., Zhang, F.Y., Yang, S., Liu, J.J. and Zhang, X.J. (2009). Genome-wide association study in a Chinese Han population identifies nine new susceptibility loci for systemic lupus erythematosus. Nat. Genet. *41*, 1234-1237.

Harley, I.T., Kaufman, K.M., Langefeld, C.D., Harley, J.B. and Kelly, J.A. (2009). Genetic susceptibility to SLE: new insights from fine mapping and genome-wide association studies. Nat. Rev. Genet. *10*, 285-290.

Harley, J.B., Alarcon-Riquelme, M.E., Criswell, L.A., Jacob, C.O., Kimberly, R.P., Moser, K.L., Tsao, B.P., Vyse, T.J., Langefeld, C.D., Nath, S.K., Guthridge, J.M., Cobb, B.L., Mirel, D.B., Marion, M.C., Williams, A.H., Divers, J., Wang, W., Frank, S.G., Namjou, B., Gabriel, S.B., Lee, A.T., Gregersen, P.K., Behrens, T.W., Taylor, K.E., Fernando, M., Zidovetzki, R., Gaffney, P.M., Edberg, J.C., Rioux, J.D., Ojwang, J.O., James, J.A., Merrill, J.T., Gilkeson, G.S., Seldin, M.F., Yin, H., Baechler, E.C., Li, Q.Z., Wakeland, E.K., Bruner, G.R., Kaufman, K.M. and Kelly, J.A. (2008). Genome-wide association scan in women with systemic lupus erythematosus identifies susceptibility variants in *ITGAM, PXK, KIAA1542* and other loci. Nat. Genet. *40*, 204-210.

Hastbacka, J., de la Chapelle, A., Kaitila, I., Sistonen, P., Weaver, A. and Lander, E. (1992). Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. Nat. Genet. *2*, 204-211.

Hinch, A.G., Tandon, A., Patterson, N., Song, Y., Rohland, N., Palmer, C.D., Chen, G.K., Wang, K., Buxbaum, S.G., Akylbekova, E.L., Aldrich, M.C., Ambrosone, C.B., Amos, C., Bandera, E.V., Berndt, S.I., Bernstein, L., Blot, W.J., Bock, C.H., Boerwinkle, E., Cai, Q., Caporaso, N., Casey, G., Cupples, L.A., Deming, S.L., Diver, W.R., Divers, J., Fornage, M., Gillanders, E.M., Glessner, J., Harris, C.C., Hu, J.J., Ingles, S.A., Isaacs, W., John, E.M., Kao, W.H., Keating, B., Kittles, R.A., Kolonel, L.N., Larkin, E., Le, M.L., McNeill, L.H., Millikan, R.C., Murphy, A., Musani, S., Neslund-Dudas, C., Nyante, S., Papanicolaou, G.J., Press, M.F., Psaty, B.M., Reiner, A.P., Rich, S.S., Rodriguez-Gil, J.L., Rotter, J.I., Rybicki, B.A., Schwartz, A.G., Signorello, L.B., Spitz, M., Strom, S.S., Thun, M.J., Tucker, M.A., Wang, Z., Wiencke, J.K., Witte, J.S., Wrensch, M., Wu, X., Yamamura, Y., Zanetti, K.A., Zheng, W., Ziegler, R.G., Zhu, X., Redline, S., Hirschhorn, J.N., Henderson, B.E., Taylor, H.A., Jr., Price, A.L., Hakonarson, H., Chanock, S.J., Haiman, C.A., Wilson, J.G., Reich, D. and Myers, S.R. (2011). The landscape of recombination in African- Americans. Nature *476*, 170-175.

Holliday, R. (1964). The induction of mitotic recombination by mitomycin c in ustilago and saccharomyces. Genetics *50*, 323-335.

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.V., Ballinger, D.G., Kosoy, R., Demirci, F.Y., Kamboh, M.I., Kao, A.H., Tian, C., Gunnarsson, I., Bengtsson, A.A., Rantapaa-Dahlqvist, S., Petri, M., Manzi, S., Seldin, M.F., Ronnblom, L., Syvanen, A.C., Criswell, L.A., Gregersen, P.K. and Behrens, T.W. (2008). Association of systemic lupus erythematosus with *C8orf13-BLK* and *ITGAM-ITGAX*. N. Engl. J. Med. *358*, 900-909.

Houghton, K.M., Page, J., Cabral, D.A., Petty, R.E. and Tucker, L.B. (2006). Systemic lupus erythematosus in the pediatric north American native population of British Columbia. J. Rheumatol. *33*, 161-163.

Howie, B.N., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS. Genet. *5*, e1000529.

Hu, X., Kim, H., Stahl, E., Plenge, R., Daly, M. and Raychaudhuri, S. (2011). Integrating autoimmune risk loci with gene-expression data identifies specific pathogenic immune cell subsets. Am. J. Hum. Genet. *89*, 496-506.

Ikushima, S., Inukai, T., Inaba, T., Nimer, S.D., Cleveland, J.L. and Look, A.T. (1997). Pivotal role for the NFIL3/E4BP4 transcription factor in interleukin 3-mediated survival of pro-B lymphocytes. Proc. Natl. Acad. Sci. U. S. A. *94*, 2609-2614.

Ito, T., Wang, Y.H., Duramad, O., Hanabuchi, S., Perng, O.A., Gilliet, M., Qin, F.X. and Liu, Y.J. (2006). OX40 ligand shuts down IL-10-producing regulatory T-cells. Proc. Natl. Acad. Sci. U. S. A. *103*, 13138-13143.

Jarvinen, P. and Aho, K. (1994). Twin studies in rheumatic diseases. Semin. Arthritis Rheum. *24*, 19-28.

Javierre, B.M., Fernandez, A.F., Richter, J., Al-Shahrour, F., Martin-Subero, J.I., Rodriguez-Ubreva, J., Berdasco, M., Fraga, M.F., O'Hanlon, T.P., Rider, L.G., Jacinto, F.V., Lopez-Longo, F.J., Dopazo, J., Forn, M., Peinado, M.A., Carreno, L., Sawalha, A.H., Harley, J.B. Siebert, R., Esteller, M., Miller, F.W. and Ballestar, E. (2010). Changes in the pattern of DNA methylation associate with twin discordance in systemic lupus erythematosus. Genome Res. *20*, 170-179.

Jeffreys, A.J. and May, C.A. (2004). Intense and highly localized gene conversion activity in human meiotic crossover hot spots. Nat. Genet. *36*, 151-156.

Johanneson, B., Steinsson, K., Lindqvist, A.K., Kristjansdottir, H., Grondal, G., Sandino, S., Tjernstrom, F., Sturfelt, G., Granados-Arriola, J., cocer-Varela, J., Lundberg, I., Jonasson, I., Truedsson, L., Svenungsson, E., Klareskog, L., Alarcon-Segovia, D., Gyllensten, U.B. and Alarcon-Riquelme, M.E. (1999). A comparison of genome-scans performed in multicase families with systemic lupus erythematosus from different population groups. J. Autoimmun. *13*, 137-141.

Johnson, G.C., Esposito, L., Barratt, B.J., Smith, A.N., Heward, J., Di, G.G., Ueda, H., Cordell, H.J., Eaves, I.A., Dudbridge, F., Twells, R.C., Payne, F., Hughes, W., Nutland, S., Stevens, H., Carr, P., Tuomilehto-Wolf, E., Tuomilehto, J., Gough, S.C., Clayton, D.G. and Todd, J.A. (2001). Haplotype tagging for the identification of common disease genes. Nat. Genet. *29*, 233-237.

Jorgenson, E., Tang, H., Gadde, M., Province, M., Leppert, M., Kardia, S., Schork, N., Cooper, R., Rao, D.C., Boerwinkle, E. and Risch, N. (2005). Ethnicity and human genetic linkage maps. Am. J. Hum. Genet. *76*, 276-290.

Ju, Y.S., Kim, J.I., Kim, S., Hong, D., Park, H., Shin, J.Y., Lee, S., Lee, W.C., Kim, S., Yu, S.B., Park, S.S., Seo, S.H., Yun, J.Y., Kim, H.J., Lee, D.S., Yavartanoo, M., Kang, H.P., Gokcumen, O., Govindaraju, D.R., Jung, J.H., Chong, H., Yang, K.S., Kim, H., Lee, C. and Seo, J.S. (2011). Extensive genomic and transcriptional diversity identified through massively parallel DNA and RNA sequencing of eighteen Korean individuals. Nat. Genet. *43*, 745-752.

Kadota, M., Yang, H.H., Hu, N., Wang, C., Hu, Y., Taylor, P.R., Buetow, K.H. and Lee, M.P. (2007). Allele-specific chromatin immunoprecipitation studies show genetic influence on chromatin state in human genome. PLoS. Genet. *3*, e81.

Kamen, D.L., Barron, M., Parker, T.M., Shaftman, S.R., Bruner, G.R., Aberle, T., James, J.A., Scofield, R.H., Harley, J.B. and Gilkeson, G.S. (2008). Autoantibody prevalence and lupus characteristics in a unique African-American population. Arthritis Rheum. *58*, 1237-1247.

Kelly, J.A., Kelley, J.M., Kaufman, K.M., Kilpatrick, J., Bruner, G.R., Merrill, J.T., James, J.A., Frank, S.G., Reams, E., Brown, E.E., Gibson, A.W., Marion, M.C., Langefeld, C.D., Li, Q.Z., Karp, D.R., Wakeland, E.K., Petri, M., Ramsey-Goldman, R., Reveille, J.D., Vila, L.M., Alarcon, G.S., Kimberly, R.P., Harley, J.B. and Edberg, J.C. (2008). Interferon regulatory factor-5 is genetically associated with systemic lupus erythematosus in African-Americans. Genes Immun. *9*, 187-194.

Kim, M.Y., Gaspal, F.M., Wiggett, H.E., McConnell, F.M., Gulbranson-Judge, A., Raykundalia, C., Walker, L.S., Goodall, M.D. and Lane, P.J. (2003). CD4(+)CD3(-) accessory cells costimulate primed CD4 T-cells through OX40 and CD30 at sites where T-cells collaborate with B-cells. Immunity. *18*, 643-654.

King, D.C., Taylor, J., Elnitski, L., Chiaromonte, F., Miller, W. and Hardison, R.C. (2005). Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences. Genome Res. *15*, 1051-1060.

Kircher, M., Sawyer, S. and Meyer, M. (2012). Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic Acids Res. *40*, e3.

Kong, A., Thorleifsson, G., Gudbjartsson, D.F., Masson, G., Sigurdsson, A., Jonasdottir, A., Walters, G.B., Jonasdottir, A., Gylfason, A., Kristinsson, K.T., Gudjonsson, S.A., Frigge, M.L., Helgason, A., Thorsteinsdottir, U. and Stefansson, K. (2010). Fine-scale recombination rate differences between sexes, populations and individuals. Nature *467*, 1099-1103.

Krimmer, D.I., Loseli, M., Hughes, J.M., Oliver, B.G., Moir, L.M., Hunt, N.H., Black, J.L. and Burgess, J.K. (2009). CD40 and OX40 ligand are differentially regulated on asthmatic airway smooth muscle. Allergy *64*, 1074-1082.

Kruglyak, L. and Nickerson, D.A. (2001). Variation is the spice of life. Nat Genet. *27,* 1253-1261.

Kumar, P., Henikoff, S. and Ng, P.C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat. Protoc. *4*, 1073-1081.

Kushner, A.L., Kamara, T.B., Groen, R.S., Fadlu-Deen, B.D., Doah, K.S. and Kingham, T.P. (2010). Improving access to surgery in a developing country: experience from a surgical collaboration in Sierra Leone. J. Surg. Educ. *67*, 270-273.

Lahita, R.G., Bradlow, H.L., Kunkel, H.G. and Fishman, J. (1979). Alterations of estrogen metabolism in systemic lupus erythematosus. Arthritis Rheum. *22*, 1195-1198.

Lander, E.S. and Schork, N.J. (1994). Genetic dissection of complex traits. Science *265*, 2037-2048.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A.,

Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R.A., Muzny, D.M., Scherer, S.E., Bouck, J.B., Sodergren, E.J., Worley, K.C., Rives, C.M., Gorrell, J.H., Metzker, M.L., Naylor, S.L., Kucherlapati, R.S., Nelson, D.L., Weinstock, G.M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D.R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H.M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R.W., Federspiel, N.A., Abola, A.P., Proctor, M.J., Myers, R.M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D.R., Olson, M.V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G.A., Athanasiou, M., Schultz, R., Roe, B.A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W.R., de la Bastide, M., Dedhia, N., Blocker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J.A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D.G., Burge, C.B., Cerutti, L., Chen, H.C., Church, D., Clamp, M., Copley, R.R., Doerks, T., Eddy, S.R., Eichler, E.E., Furey, T.S., Galagan, J., Gilbert, J.G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L.S., Jones, T.A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W.J., Kitts, P., Koonin, E.V., Korf, I., Kulp, D., Lancet, D., Lowe, T.M., McLysaght, A., Mikkelsen, T., Moran, J.V., Mulder, N., Pollara, V.J., Ponting, C.P., Schuler, G., Schultz, J., Slater, G., Smit, A.F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y.I., Wolfe, K.H., Yang, S.P., Yeh, R.F., Collins, F., Guyer, M.S., Peterson, J., Felsenfeld, A., Wetterstrand, K.A., Patrinos, A., Morgan, M.J., de, J.P., Catanese, J.J., Osoegawa, K., Shizuya, H., Choi, S. and Chen, Y.J. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860-921.

Leavy, O. (2010). Autoimmunity: New players in lupus nephritis. Nat. Rev. Immunol. *10*, 464.

Lee, D.M., Friend, D.S., Gurish, M.F., Benoist, C., Mathis, D. and Brenner, M.B. (2002). MasT-cells: a cellular link between autoantibodies and inflammatory arthritis. Science *297*, 1689-1692.

Lee, L.A., Gaither, K.K., Coulter, S.N., Norris, D.A. and Harley, J.B. (1989). Pattern of cutaneous immunoglobulin G deposition in subacute cutaneous lupus erythematosus is reproduced by infusing purified Anti-Ro (SSA) autoantibodies into human skin-grafted mice. J. Clin. Invest *83*, 1556-1562.

Lee-Kirsch, M.A., Gong, M., Chowdhury, D., Senenko, L., Engel, K., Lee, Y.A., de, S.U., Bailey, S.L., Witte, T., Vyse, T.J., Kere, J., Pfeiffer, C., Harvey, S., Wong, A., Koskenmies, S., Hummel, O., Rohde, K., Schmidt, R.E., Dominiczak, A.F., Gahr, M., Hollis, T., Perrino, F.W., Lieberman, J. and Hubner, N. (2007). Mutations in the gene encoding the 3'-5' DNA exonuclease TREX1 are associated with systemic lupus erythematosus. Nat. Genet. *39*, 1065-1067.

Li, H. and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. *26*, 589-595.

Liao, L.H., Zhang, H., Lai, M.P., Chen, S.L., Wu, M. and Shen, N. (2011). Single-nucleotide polymorphisms and haplotype of CYP2E1 gene associated with systemic lupus erythematosus in Chinese population. Arthritis Res. Ther. *13*, R11.

Lin, W.J., Su, Y.W., Lu, Y.C., Hao, Z., Chio, I.I., Chen, N.J., Brustle, A., Li, W.Y. and Mak, T.W. (2011). Crucial role for TNF receptor-associated factor 2 (TRAF2) in regulating NFkappaB2 signaling that contributes to autoimmunity. Proc. Natl. Acad. Sci. U. S. A *108*, 18354-18359.

Listgarten, J., Kadie, C., Schadt, E.E. and Heckerman, D. (2010). Correction for hidden confounders in the genetic analysis of gene expression. Proc. Natl. Acad. Sci. U. S. A *107*, 16465-16470.

Lohmueller, K.E., Mauney, M.M., Reich, D. and Braverman, J.M. (2006). Variants associated with common disease are not unusually differentiated in frequency across populations. Am. J. Hum. Genet. *78*, 130-136.

Lunetta, L.L., (2008). Genetic association studies. Cicrculation *118,* 96-101.

Lu, Q. and Elston, R.C. (2008). Using the optimal receiver operating characteristic curve to design a predictive genetic test, exemplified with type 2 diabetes. Am J Hum Genet *82*, 641–651.

Lyons, P.A., McKinney, E.F., Rayner, T.F., Hatton, A., Woffendin, H.B., Koukoulaki, M., Freeman, T.C., Jayne, D.R., Chaudhry, A.N. and Smith, K.G. (2010). Novel expression signatures identified by transcriptional analysis of separated leucocyte subsets in systemic lupus erythematosus and vasculitis. Ann. Rheum. Dis. *69*, 1208-1213.

MacDonald, M.E., Novelletto, A., Lin, C., Tagle, D., Barnes, G., Bates, G., Taylor, S., Allitto, B., Altherr, M., Myers, R., Lehrach, H., Collins, F.S., Wasmuth, J.J., Frontali, M. and Gusella, J.F. (1992). The Huntington's disease candidate region exhibits many different haplotypes. Nat. Genet. *1,* 99-103.

Madsen, A.M., Ottman, R. and Hodge, S.E. (2011). Causal models for investigating complex genetic disease: II. What causal models can tell us about penetrance for additive, heterogeneity, and multiplicative two-locus models. Hum. Hered. *72*, 63-72.

Marchini, J. and Howie, B. (2010). Genotype imputation for genome-wide association studies. Nat. Rev. Genet. *11*, 499-511.

Marchini, J., Howie, B., Myers, S., McVean, G. and Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. Nat. Genet. *39*, 906-913.

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., Dewell, S.B., Du, L., Fierro, J.M., Gomes, X.V., Godwin, B.C., He, W., Helgesen, S., Ho, C.H., Irzyk, G.P., Jando, S.C., Alenquer, M.L., Jarvie, T.P., Jirage, K.B., Kim, J.B., Knight, J.R., Lanza, J.R., Leamon, J.H., Lefkowitz, S.M., Lei, M., Li, J., Lohman, K.L., Lu, H., Makhijani, V.B., McDade, K.E., McKenna, M.P., Myers, E.W., Nickerson, E., Nobile, J.R., Plant, R., Puc, B.P., Ronan, M.T., Roth, G.T., Sarkis, G.J., Simons, J.F., Simpson, J.W., Srinivasan, M., Tartaro, K.R., Tomasz, A., Vogt, K.A., Volkmer, G.A., Wang, S.H., Wang, Y., Weiner, M.P., Yu, P., Begley, R.F. and Rothberg, J.M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature *437*, 376-380.

Marston, B. and Looney, R.J. (2010). Connective tissue diseases: Translating the effects of BAFF in SLE. Nat. Rev. Rheumatol. *6*, 503-504.

Martin, F. and Chan, A.C. (2006). B-cell immunobiology in disease: evolving concepts from the clinic. Annu. Rev. Immunol. *24*, 467-496.

Masi, A.T. and Kaslow, R.A. (1978). Sex effects in systemic lupus erythematosus: a clue to pathogenesis. Arthritis Rheum. *21*, 480-484.

McPeek, M.S. and Strahs, A. (1999). Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping. Am. J. Hum. Genet. *65*, 858-875.

McVean, G.A., Myers, S.R., Hunt, S., Deloukas, P., Bentley, D.R. and Donnelly, P. (2004). The fine-scale structure of recombination rate variation in the human genome. Science *304*, 581-584.

Metzker, M.L. (2010). Sequencing technologies - the next generation. Nat. Rev. Genet. *11*, 31-46.

Meyer, M., Stenzel, U. and Hofreiter, M. (2008). Parallel tagged sequencing on the 454 platform. Nat. Protoc. *3*, 267-278.

Michaelson, J.J., Loguercio, S. and Beyer, A. (2009). Detection and interpretation of expression quantitative trait loci (eQTL). Methods *48*, 265-276.

Mihara, M., Tan, I., Chuzhin, Y., Reddy, B., Budhai, L., Holzer, A., Gu, Y. and Davidson, A. (2000). CTLA4Ig inhibits T-cell-dependent B-cell maturation in murine systemic lupus erythematosus. J. Clin. Invest. *106*, 91-101.

Mihas, A.A., Foster, M.M., Barnes, S., Mihas, T.A. and Spenney, J.G. (1981). Effects of spironolactone on serum bile acids. Clin. Exp. Pharmacol. Physiol *8*, 87-88.

Mills, R.E., Walter, K., Stewart, C., Handsaker, R.E., Chen, K., Alkan, C., Abyzov, A., Yoon, S.C., Ye, K., Cheetham, R.K., Chinwalla, A., Conrad, D.F., Fu, Y., Grubert, F., Hajirasouliha, I., Hormozdiari, F., Iakoucheva, L.M., Iqbal, Z., Kang, S., Kidd, J.M., Konkel, M.K., Korn, J., Khurana, E., Kural, D., Lam, H.Y., Leng, J., Li, R., Li, Y., Lin, C.Y., Luo, R., Mu, X.J., Nemesh, J., Peckham, H.E., Rausch, T., Scally, A., Shi, X., Stromberg, M.P., Stütz, A.M., Urban, A.E., Walker, J.A., Wu, J., Zhang, Y., Zhang, Z.D., Batzer, M.A., Ding, L., Marth, G.T., McVean, G., Sebat, J., Snyder, M., Wang, J., Ye, K., Eichler, E.E., Gerstein, M.B., Hurles, M.E., Lee, C., McCarroll, S.A., Korbel, J.O, 1000 Genomes Project. (2011). Mapping copy number variation by population-scale genome sequencing. Nature *470*, 59-65.

Minton, K. (2011). B-cells: Short- and long-term memory. Nat. Rev. Immunol. *11*, 160.

Molina, J.F., Molina, J., Garcia, C., Gharavi, A.E., Wilson, W.A. and Espinoza, L.R. (1997). Ethnic differences in the clinical expression of systemic lupus erythematosus: a comparative study between African-Americans and Latin Americans. Lupus *6*, 63-67.

Mond, C.B., Peterson, M.G. and Rothfield, N.F. (1989). Correlation of anti-Ro antibody with photosensitivity rash in systemic lupus erythematosus patients. Arthritis Rheum. *32*, 202-204.

Monestier, M. and Kotzin, B.L. (1992). Antibodies to histones in systemic lupus erythematosus and drug-induced lupus syndromes. Rheum. Dis. Clin. North Am. *18*, 415-436.

Morris, D.L., Taylor, K.E., Fernando, M.M., Nititham, J., Alarcón-Riquelme, M.E., Barcellos, L.F., Behrens, T.W., Cotsapas, C., Gaffney, P.M., Graham, R.R., Pons-Estel, B.A., Gregersen, P.K., Harley, J.B., Hauser. S.L., Hom, G., International MHC and Autoimmunity Genetics Network., Langefeld. C.D., Noble, J.A., Rioux. J.D., Seldin, M.F., Systemic Lupus Erythematosus Genetics Consortium, Criswell, L.A. and Vyse, T.J. (2012) Unravelling multiple MHC gene associations with systemic lupus erythematosus: model choice indicates a role for HLA alleles and non-HLA genes in Europeans. Am J Hum Genet *91*, 778-793.

Mullikin, J.C., Hunt, S.E., Cole, C.G., Mortimore, B.J., Rice, C.M., Burton, J., Matthews, L.H., Pavitt, R., Plumb, R.W., Sims, S.K., Ainscough, R.M., Attwood, J., Bailey, J.M., Barlow, K., Bruskiewich, R.M., Butcher, P.N., Carter, N.P., Chen, Y., Clee, C.M., Coggill, P.C., Davies, J., Davies, R.M., Dawson, E., Francis, M.D., Joy, A.A., Lamble, R.G., Langford, C.F., Macarthy, J., Mall, V., Moreland, A., Overton-Larty, E.K., Ross, M.T., Smith, L.C., Steward, C.A., Sulston, J.E., Tinsley, E.J., Turney, K.J., Willey, D.L., Wilson, G.D., McMurray, A.A., Dunham, I., Rogers, J. and Bentley, D.R. (2000). A SNP map of human chromosome 22. Nature *407*, 516-520.

Murray, J.C., Buetow, K.H., Weber, J.L., Ludwigsen, S., Scherpbier-Heddema, T., Manion, F., Quillen, J., Sheffield, V.C., Sunden, S. and Duyk, G.M. (1994). A comprehensive human linkage map with centimorgan density. Cooperative Human Linkage Center (CHLC). Science *265*, 2049-2054.

Myers, S., Bottolo, L., Freeman, C., McVean, G. and Donnelly, P. (2005). A fine-scale map of recombination rates and hotspots across the human genome. Science *310*, 321-324.

Namjou, B., Sestak, A.L., Armstrong, D.L., Zidovetzki, R., Kelly, J.A., Jacob, N., Ciobanu, V., Kaufman, K.M., Ojwang, J.O., Ziegler, J., Quismorio, F.P., Jr., Reiff, A., Myones, B.L., Guthridge, J.M., Nath, S.K., Bruner, G.R., Mehrian-Shai, R., Silverman, E., Klein-Gitelman, M., McCurdy, D., Wagner-Weiner, L., Nocton, J.J., Putterman, C., Bae, S.C., Kim, Y.J., Petri, M., Reveille, J.D., Vyse, T.J., Gilkeson, G.S., Kamen, D.L., arcon-Riquelme, M.E., Gaffney, P.M., Moser, K.L., Merrill, J.T., Scofield, R.H., James, J.A., Langefeld, C.D., Harley, J.B. and Jacob, C.O. (2009). High-density genotyping of *STAT4* reveals multiple haplotypic associations with systemic lupus erythematosus in different racial groups. Arthritis Rheum. *60*, 1085-1095.

Napirei, M., Karsunky, H., Zevnik, B., Stephan, H., Mannherz, H.G. and Moroy, T. (2000). Features of systemic lupus erythematosus in Dnase1-deficient mice. Nat. Genet. *25*, 177-181.

Nath, S.K., Han, S., Kim-Howard, X., Kelly, J.A., Viswanathan, P., Gilkeson, G.S., Chen, W., Zhu, C., McEver, R.P., Kimberly, R.P., Alarcon-Riquelme, M.E., Vyse, T.J., Li, Q.Z., Wakeland, E.K., Merrill, J.T., James, J.A., Kaufman, K.M., Guthridge, J.M. and Harley, J.B. (2008). A nonsynonymous functional variant in integrin-alpha(M) (encoded by *ITGAM*) is associated with systemic lupus erythematosus. Nat. Genet. *40*, 152-154.

Ng, S.B., Turner, E.H., Robertson, P.D., Flygare, S.D., Bigham, A.W., Lee, C., Shaffer, T., Wong, M., Bhattacharjee, A., Eichler, E.E., Bamshad, M., Nickerson, D.A. and Shendure, J. (2009). Targeted capture and massively parallel sequencing of 12 human exomes. Nature *461*, 272-276.

Nohara, C., Akiba, H., Nakajima, A., Inoue, A., Koh, C.S., Ohshima, H., Yagita, H., Mizuno, Y. and Okumura, K. (2001). Amelioration of experimental autoimmune encephalomyelitis with anti-OX40 ligand monoclonal antibody: a critical role for OX40 ligand in migration, but not development, of pathogenic T-cells. J. Immunol. *166*, 2108-2115.

Olofsson, P.S., Soderstrom, L.A., Jern, C., Sirsjo, A., Ria, M., Sundler, E., de, F.U., Wiklund, P.G., Ohrvik, J., Hedin, U., Paulsson-Berne, G., Hamsten, A., Eriksson, P. and Hansson, G.K. (2009). Genetic variants of *TNFSF4* and risk for carotid artery disease and stroke. J. Mol. Med. (Berl) *87*, 337-346.

Patterson, N., Price, A.L. and Reich, D., (2006). Population structure and eigenanalysis. Plos Genet. 2, e190

Peschken, C.A., Katz, S.J., Silverman, E., Pope, J.E., Fortin, P.R., Pineau, C., Smith, C.D., Arbillaga, H.O., Gladman, D.D., Urowitz, M., Zummer, M., Clarke, A., Bernatsky, S., and Hudson, M. (2009). The 1000 Canadian faces of lupus: determinants of disease outcome in a large multiethnic cohort. J. Rheumatol. *36*, 1200-1208

Pfaff, C.L., Parra, E.J., Bonilla, C., Hiester, K., McKeigue, P.M., Kamboh, M.I., Hutchinson, R.G., Ferrell, R.E., Boerwinkle, E. and Shriver, M.D. (2001). Population structure in admixed populations: effect of admixture dynamics on the pattern of linkage disequilibrium. Am. J. Hum. Genet. *68*, 198-207.

Pollex, R. L. and Hegele, R. A. (2007). Copy number variation in the human genome and its implications for cardiovascular disease. Circulation *115*: 3130-3138.

Pons-Estel, B.A., Catoggio, L.J., Cardiel, M.H., Soriano, E.R., Gentiletti, S., Villa, A.R., Abadi, I., Caeiro, F., Alvarellos, A. and Alarcon-Segovia, D. (2004). The GLADEL multinational Latin American prospective inception cohort of 1, 214 patients with systemic lupus erythematosus: ethnic and disease heterogeneity among "Hispanics". Medicine (Baltimore) *83*, 1-17.

Portales-Casamar, E., Thongjuea, S., Kwon, AT., Arenillas, D., Zhao, X., Valen, E., Yusuf, D., Lenhard, B., Wasserman, W. and Sandelin, A. (2010). JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. Nucleic Acids Res. *38*, D105-10.

Preble, O.T., Rothko, K., Klippel, J.H., Friedman, R.M. and Johnston, M.I. (1983). Interferon-induced 2'-5' adenylate synthetase in vivo and interferon production in vitro by lymphocytes from systemic lupus erythematosus patients with and without circulating interferon. J. Exp. Med. *157*, 2140-2146.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A. and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. *38*, 904-909.

Price, A.L., Patterson, N., Yu, F., Cox, D.R., Waliszewska, A., McDonald, G.J., Tandon, A., Schirmer, C., Neubauer, J., Bedoya, G., Duque, C., Villegas, A., Bortolini, M.C., Salzano, F.M., Gallo, C., Mazzotti, G., Tello-Ruiz, M., Riba, L., guilar-Salinas, C.A., Canizales-Quinteros, S., Menjivar, M., Klitz, W., Henderson, B., Haiman, C.A., Winkler, C., Tusie-Luna, T., Ruiz-Linares, A. and Reich, D. (2007). A genomewide admixture map for Latino populations. Am. J. Hum. Genet. *80*, 1024-1036.

Pritchard, J.K., Stephens, M. and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. Genetics 155, 945–959.

Quinlan, A.R. and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. *26*, 841-842.

Ramos, P.S., Criswell, L.A., Moser, K.L., Comeau, M.E., Williams, A.H., Pajewski, N.M., Chung, S.A., Graham, R.R., Zidovetzki, R., Kelly, J.A., Kaufman, K.M., Jacob, C.O., Vyse, T.J., Tsao, B.P., Kimberly, R.P., Gaffney, P.M., Alarcon-Riquelme, M.E., Harley, J.B. and Langefeld, C.D. (2011). A Comprehensive Analysis of Shared Loci between Systemic Lupus Erythematosus (SLE) and Sixteen Autoimmune Diseases Reveals Limited Genetic Overlap. PLoS. Genet. *7*, e1002406.

Reddy, M.V., Velazquez-Cruz, R., Baca, V., Lima, G., Granados, J., Orozco, L. and Alarcon-Riquelme, M.E. (2007). Genetic association of *IRF5* with SLE in Mexicans: higher frequency of the risk haplotype and its homozygozity than Europeans. Hum. Genet. *121*, 721-727.

Risch, N. (1990). Linkage strategies for genetically complex traits. II. The power of affected relative pairs. Am J Hum Genet. *46*, 229-241.

Rhodes, B. and Vyse, T.J. (2008). The genetics of SLE: an update in the light of genome-wide association studies. Rheumatology. (Oxford) *47*, 1603-1611.

Rivas, M.A., Beaudoin, M., Gardet, A., Stevens, C., Sharma, Y., Zhang, C.K., Boucher, G., Ripke, S., Ellinghaus, D., Burtt, N., Fennell, T., Kirby, A., Latiano, A., Goyette, P., Green, T., Halfvarson, J., Haritunians, T., Korn, J.M., Kuruvilla, F., Lagacé, C., Neale, B., Lo, K.S., Schumm, P., Törkvist, L., National Institute of Diabetes and Digestive Kidney Diseases., Inflammatory Bowel Disease Genetics Consortium (NIDDK, IBDGC), United, Kingdom Inflammatory Bowel Disease Genetics Consortium, International Inflammatory Bowel Disease Genetics Consortium, Dubinsky, M.C., Brant, S.R., Silverberg, M.S., Duerr, R.H., Altshuler, D., Gabriel, S., Lettre, G., Franke, A., D'Amato, M., McGovern, D.P., Cho, J.H., Rioux, J.D., Xavier, R.J. and Daly, M.J. (2011) Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nat Genet. *43*, 1066-73.

Robinson, J.T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G. and Mesirov, J.P. (2011). Integrative genomics viewer. Nat. Biotechnol. *29*, 24-26.

Ronnblom, L. and Alm, G.V. (2002). The natural interferon-alpha producing cells in systemic lupus erythematosus. Hum. Immunol. *63*, 1181-1193.

Rosenberg, B.R., Hamilton, C.E., Mwangi, M.M., Dewell, S., and Papavasiliou, F.N. (2011). Transcriptome-wide sequencing reveals numerous APOBEC1 mRNA-editing targets in transcript 3' UTRs. Nat. Struct. Mol. Biol. *18*, 230-236.

Rothberg, J.M. and Leamon, J.H. (2008). The development and impact of 454 sequencing. Nat. Biotechnol. *26*, 1117-1124.

Rozen, S. and Skaletsky, H. (2000). Primer3 on the WWW for general users and for biologist programmers. Methods Mol. Biol. *132*, 365-386.

Rozzo, S.J., Vyse, T.J., Drake, C.G. and Kotzin, B.L. (1996). Effect of genetic background on the contribution of New Zealand black loci to autoimmune lupus nephritis. Proc. Natl. Acad. Sci. U. S. A. *93*, 15164-15168.

Sackton, T.B. and Clark, A.G. (2009). Comparative profiling of the transcriptional response to infection in two species of Drosophila by short-read cDNA sequencing. BMC. Genomics *10*, 259.

Salek-Ardakani, S. and Croft, M. (2010). Tumor necrosis factor receptor/tumor necrosis factor family members in antiviral CD8 T-cell immunity. J. Interferon Cytokine Res. *30*, 205-218.

Sanchez, E., Webb, R.D., Rasmussen, A., Kelly, J.A., Riba, L., Kaufman, K.M., Garcia-de la Torre, I., Moctezuma, J.F., Maradiaga-Cecena, M.A., Cardiel-Rios, M.H., Acevedo, E., Cucho-Venegas, M., Garcia, M.A., Gamron, S., Pons-Estel, B.A., Vasconcelos, C., Martin, J., Tusie-Luna, T., Harley, J.B., Richardson, B., Sawalha, A.H. and Alarcon-Riquelme, M.E. (2010). Genetically determined Amerindian ancestry correlates with increased frequency of risk alleles for systemic lupus erythematosus. Arthritis Rheum. *62*, 3722-3729.

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

Sasieni, P.D. (1997). From genotypes to genes: doubling the sample size. Biometrics *53,* 1253-1261.

Sankararaman, S., Sridhar, S., Kimmel, G. and Halperin, E. (2008). Estimating local ancestry in admixed populations. Am. J. Hum. Genet. *82*, 290-303.

Scheet, P. and Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am. J. Hum. Genet. *78*, 629-644.

Seldin, M.F., Qi, L., Scherbarth, H.R., Tian, C., Ransom, M., Silva, G., Belmont, J.W., Gamron, S., Allievi, A., Palatnik, S.A., Saurit, V., Paira, S., Graf, C., Guilleron, C., Catoggio, L.J., Prigione, C., Berbotto, G.A., Garcia, M.A., Perandones, C.E., Truedsson, L., Abderrahim, H., Battagliotti, C.G., Pons-Estel, B.A. and Alarcon-Riquelme, M.E. (2008). Amerindian ancestry in Argentina is associated with increased risk for systemic lupus erythematosus. Genes Immun. *9*, 389-393.

Service, S.K., Lang, D.W., Freimer, N.B. and Sandkuijl, L.A. (1999). Linkage-disequilibrium mapping of disease genes by reconstruction of ancestral haplotypes in founder populations. Am. J. Hum. Genet. *64*, 1728-1738.

Seshasayee, D., Lee, W.P., Zhou, M., Shu, J., Suto, E., Zhang, J., Diehl, L., Austin, C.D., Meng, Y.G., Tan, M., Bullens, S.L., Seeber, S., Fuentes, M.E., Labrijn, A.F., Graus, Y.M., Miller, L.A., Schelegle, E.S., Hyde, D.M., Wu, L.C., Hymowitz, S.G., and Martin, F. (2007). *In vivo* blockade of OX40 ligand inhibits thymic stromal lymphopoietin driven atopic inflammation. J. Clin. Invest *117*, 3868-3878.

Shen, N. and Tsao, B.P. (2004). Current advances in the human lupus genetics. Curr. Rheumatol. Rep. *6*, 391-398.

Shlomchik, M.J., Madaio, M.P., Ni, D., Trounstein, M. and Huszar, D. (1994). The role of B-cells in lpr/lpr-induced autoimmunity. J. Exp. Med. *180*, 1295-1306.

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., Weinstock, G.M., Wilson, R.K.,

Gibbs, R.A., Kent, W.J., Miller, W. and Haussler, D. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. *15*, 1034-1050.

Sigurdsson, S., Nordmark, G., Goring, H.H., Lindroos, K., Wiman, A.C., Sturfelt, G., Jonsen, A., Rantapaa-Dahlqvist, S., Moller, B., Kere, J., Koskenmies, S., Widen, E., Eloranta, M.L., Julkunen, H., Kristjansdottir, H., Steinsson, K., Alm, G., Ronnblom, L. and Syvanen, A.C. (2005). Polymorphisms in the tyrosine kinase 2 and interferon regulatory factor 5 genes are associated with systemic lupus erythematosus. Am. J. Hum. Genet. *76*, 528-537.

Simard, J.F. and Costenbader, K.H. (2007). What can epidemiology tell us about systemic lupus erythematosus? Int. J. Clin. Pract. *61*, 1170-1180.

Slatkin, M. (2008). Exchangeable models of complex inherited diseases. Genetics *179*, 2253-2261.

So, T., Lee, S.W. and Croft, M. (2006). Tumor necrosis factor/tumor necrosis factor receptor family members that positively regulate immunity. Int. J. Hematol. *83*, 1-11.

So, T., Choi, H., and Croft, M. (2011). OX40 complexes with phosphoinositide 3-kinase and protein kinase B (PKB) to augment TCR-dependent PKB signaling. J. Immunol. *186*, 3547-3555.

Stagakis, E., Bertsias, G., Verginis, P., Nakou, M., Hatziapostolou, M., Kritikos, H., Iliopoulos, D. and Boumpas, D.T. (2011). Identification of novel microRNA signatures linked to human lupus disease activity and pathogenesis: miR-21 regulates aberrant T-cell responses through regulation of PDCD4 expression. Ann. Rheum. Dis. *70*, 1496-1506.

Stuber, E., Neurath, M., Calderhead, D., Fell, H.P. and Strober, W. (1995). Cross-linking of OX40 ligand, a member of the TNF/NGF cytokine family, induces proliferation and differentiation in murine splenic B-cells. Immunity. *2*, 507-521.

Stumpf, M.P. and McVean, G.A. (2003). Estimating recombination rates from population-genetic data. Nat. Rev. Genet. *4*, 959-968.

Talal, N., Steinberg, A.D. and Daley, G.G. (1971). Inhibition of antigodies binding polyinosinic-polycytidylic acid in human and mouse lupus sera by viral and synthetic ribonucleic acids. J. Clin. Invest *50*, 1248-1252.

Tan, E.M., Schur, P.H., Carr, R.I. and Kunkel, H.G. (1966). Deoxybonucleic acid (DNA) and antibodies to DNA in the serum of patients with systemic lupus erythematosus. J. Clin. Invest *45*, 1732-1740.

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

The International HapMap Project (2003) Nature *426*, 789-796.

The International HapMap Project (2005). A haplotype map of the human genome. Nature *437,* 1299-1320.

The International HapMap Consortium. (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature *449*, 851–861.

The International HapMap 3 Consortium (2010) Integrating common and rare genetic variation in diverse human populations. Nature *467:* 52-58.

Tian, C., Hinds, D.A., Shigeta, R., Adler, S.G., Lee, A., Pahl, M.V., Silva, G., Belmont, J.W., Hanson, R.L., Knowler, W.C., Gregersen, P.K., Ballinger, D.G. and Seldin, M.F. (2007). A genomewide single-nucleotide-polymorphism panel for Mexican American admixture mapping. Am. J. Hum. Genet. *80*, 1014-1023.

Treangen, T.J. and Salzberg, S.L. (2012). Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat. Rev. Genet. *13*, 36-46.

Tsao, B.P. (2004). Update on human systemic lupus erythematosus genetics. Curr Opin Rheumatol *16*, 513-521.

Tucker, L.B., Menon, S., Schaller, J.G. and Isenberg, D.A. (1995). Adult- and childhood-onset systemic lupus erythematosus: a comparison of onset, clinical features, serology, and outcome. Br. J. Rheumatol. *34*, 866-872.

Urowitz, M.B., Ibanez, D., Jerome, D. and Gladman, D.D. (2006). The effect of menopause on disease activity in systemic lupus erythematosus. J. Rheumatol. *33*, 2192-2198.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., Gocayne, J.D., Amanatides, P., Ballew, R.M., Huson, D.H., Wortman, J.R., Zhang, Q., Kodira, C.D., Zheng, X.H., Chen, L., Skupski, M., Subramanian, G., Thomas, P.D., Zhang, J., Gabor Miklos, G.L., Nelson, C., Broder, S., Clark, A.G., Nadeau, J., McKusick, V.A., Zinder, N., Levine, A.J., Roberts, R.J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di, F., V, Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A.E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T.J., Higgins, M.E., Ji, R.R., Ke, Z., Ketchum, K.A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G.V., Milshina, N., Moore, H.M., Naik, A.K., Narayan, V.A., Neelam, B., Nusskern, D., Rusch, D.B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M.L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y.H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N.N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J.F., Guigo, R., Campbell, M.J., Sjolander, K.V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y.H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris,

M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N. and Nodell, M. (2001). The sequence of the human genome. Science *291*, 1304-1351.

Visscher, P.M., Hill, W.G., Wray, N.R. (2008). Heritability in the genomics era--concepts and misconceptions. Nat Rev Genet. *9*, 255-266.

Vyse, T.J. and Kotzin, B.L. (1998). Genetic susceptibility to systemic lupus erythematosus. Annu. Rev. Immunol. *16*, 261-292.

Vyse, T.J., Todd, J.A. and Kotzin, B.L. (1998). Non-MHC Genetic Contributions to Autoimmune Disease. In The autoimmune diseases, Rose, N.R. and Mackay, I. R. eds. Academic Press, San Diego & London, pp. 85-118.

Wandstrat, A. and Wakeland, E. (2001). The genetics of complex autoimmune diseases: non-MHC susceptibility genes. Nat. Immunol. *2*, 802-809.

Wang, X., Ria, M., Kelmenson, P.M., Eriksson, P., Higgins, D.C., Samnegard, A., Petros, C., Rollins, J., Bennet, A.M., Wiman, B., de, F.U., Wennberg, C., Olsson, P.G., Ishii, N., Sugamura, K., Hamsten, A., Forsman-Semb, K., Lagercrantz, J. and Paigen, B. (2005). Positional identification of *TNFSF4*, encoding OX40 ligand, as a gene that influences atherosclerosis susceptibility. Nat Genet *37*, 365-372.

Watts, T.H. (2005). TNF/TNFR family members in costimulation of T-cell responses. Annu. Rev. Immunol. *23*, 23-68.

Weinberg, A.D., Bourdette, D.N., Sullivan, T.J., Lemon, M., Wallin, J.J., Maziarz, R., Davey, M., Palida, F., Godfrey, W., Engleman, E., Fulton, R.J., Offner, H. and Vandenbark, A.A. (1996). Selective depletion of myelin-reactive T-cells with the anti-OX-40 antibody ameliorates autoimmune encephalomyelitis. Nat. Med. *2*, 183-189.

Weiss, K.M. and Clark, A.G. (2002). Linkage disequilibrium and the mapping of complex human traits. Trends Genet. *18*, 19-24.

Winkler, C.A., Nelson, G.W. and Smith, M.W. (2010). Admixture mapping comes of age. Annu. Rev. Genomics Hum. Genet. *11*, 65-89.

Wu, J., Xie, F., Qian, K., Gibson, A.W., Edberg, J.C. and Kimberly, R.P. (2011). FAS mRNA editing in Human Systemic Lupus Erythematosus. Hum. Mutat. *32*, 1268-1277.

Xu, L., Zhang, L., Yi, Y., Kang, H.K. and Datta, S.K. (2004). Human lupus T-cells resist inactivation and escape death by upregulating COX-2. Nat. Med. *10*, 411-415.

Yang, W., Shen, N., Ye, D.Q., Liu, Q., Zhang, Y., Qian, X.X., Hirankarn, N., Ying, D., Pan, H.F., Mok, C.C., Chan, T.M., Wong, R.W., Lee, K.W., Mok, M.Y., Wong, S.N., Leung, A.M., Li, X.P., Avihingsanon, Y., Wong, C.M., Lee, T.L., Ho, M.H., Lee, P.P., Chang, Y.K., Li, P.H., Li, R.J., Zhang, L., Wong, W.H., Ng, I.O., Lau, C.S., Sham, P.C. and Lau, Y.L. (2010). Genome-wide association study in Asian populations identifies variants in ETS1 and WDFY4 associated with systemic lupus erythematosus. PLoS. Genet. *6*, e1000841.

Yasutomo, K., Horiuchi, T., Kagami, S., Tsukamoto, H., Hashimura, C., Urushihara, M. and Kuroda, Y. (2001). Mutation of DNASE1 in people with systemic lupus erythematosus. Nat. Genet. *28*, 313-314.

Zaini, J., Andarini, S., Tahara, M., Saijo, Y., Ishii, N., Kawakami, K., Taniguchi, M., Sugamura, K., Nukiwa, T. and Kikuchi, T. (2007). OX40 ligand expressed by DCs costimulates NKT and CD4+ Th cell antitumor immunity in mice. J. Clin. Invest *117*, 3330-3338.

Appendices

Appendix A:

Table A1 **The 1982 revised criteria for classification of systemic lupus erythematosus**

Criterion	Definition
1. Malar rash	Fixed erythema, flat or raised, over the malar eminences, to spare the nasolabial folds
2. Discoid rash	Erythematous raised patches with adherent keratotic scaling and follicular plugging;
3. Photosensitivity	Skin rash as a result of unusual reaction to sunlight, by patient history or physician
4. Oral ulcers	Oral or nasopharyngeal ulceration, usually painless, observed by physician
5. Arthritis	Nonerosive arthritis involving 2 or more peripheral joints, characterized by tenderness,
6. Serositis	a) Pleuritis--convincing history of pleuritic pain or rubbing heard by a physician or evidence of pleural effusion *OR*
7. Renal disorder	a) Persistent proteinuria greater than 0.5 grams/day or greater than 3+ if quantitation not performed *OR*
8. Neurologic disorder	a) Seizures-in the absence of offending drugs or known metabolic derangements; e.g., uremia, ketoacidosis, or electrolyte imbalance *OR* b) Psychosis--in the absence of offending

9. Hematologic disorder	a) Hemolytic anemia--with reticulocytosis *OR* b) Leukopenia--less than 4,000/mm<>3<> total on 2 or more occasions *OR* c) Lymphopenia--less than 1,500/mm<>3<> on 2 or more occasions
10. Immunologic disorder	a) Positive LE cell preparation *OR* b) Anti-DNA: antibody to native DNA in abnormal titer *OR* c) Anti-Sm: presence of antibody to Sm nuclear antigen
11. Antinuclear antibody	An abnormal titer of antinuclear antibody by immunofluorescence or an equivalent assay at any point in time and in the absence of drugs known to be associated with "drug-induced

(URL: http://www.rheumatology.org/practice/clinical/classification/SLE/sle.asp)

Appendix B:

Web addresses/ uniform resource locators (URLs)

http://www.rheumatology.org/practice/clinical/classification/SLE/sle.asp
http://www.ncbi.nlm.nih.gov/
http://genome.ucsc.edu/
http://www.ensembl.org/index.html
http://www.appliedbiosystems.com/absite/us/en/home.html
https://www.wtccc.org.uk/index.shtml
http://www.appliedbiosystems.com/absite/us/en/home/applications-technologies/real-time-pcr/taqman-probe-based-gene-expression-analysis/taqman-gene-expression-assay-selection-guide.html
http://www.illumina.com/support/array/array_software/assay_design_tool.ilmn
http://www.illumina.com/technology/goldengate_genotyping_assay.ilmn
http://www.illumina.com/technology/infinium_hd_assay.ilmn
http://www.illumina.com/Documents/products/datasheets/datasheet_beadstudio.pdf
www.my454.com
www.novocraft.com/userfiles/file/NovoBarcode.pdf
http://www.uniprot.org/
http://www.decode.com/addendum/
http://www.broadinstitute.org/crd/wiki/index.php/Homopolymer
http://evs.gs.washington.edu/EVS/
http://genome.ucsc.edu/cgi-bin/hgLiftOver
http://mathgen.stats.ox.ac.uk/genetics_software/snptest/snptest.v2.pdf

Appendix C:

1000 Genomes allele frequencies, associated variants at *TNFSF4*

1000 Genomes allele frequencies, *rs2205960*

	G allele	T allele
ASW	0.9	0.1
AMR	0.71	0.29
ASN	0.77	0.23
EUR	0.79	0.21

1000 Genomes allele frequencies, *rs1234314*

	C allele	G allele
ASW	0.64	0.36
AMR	0.54	0.46
ASN	0.62	0.38
EUR	0.58	0.42

1000 Genomes allele frequencies, *rs1234317*

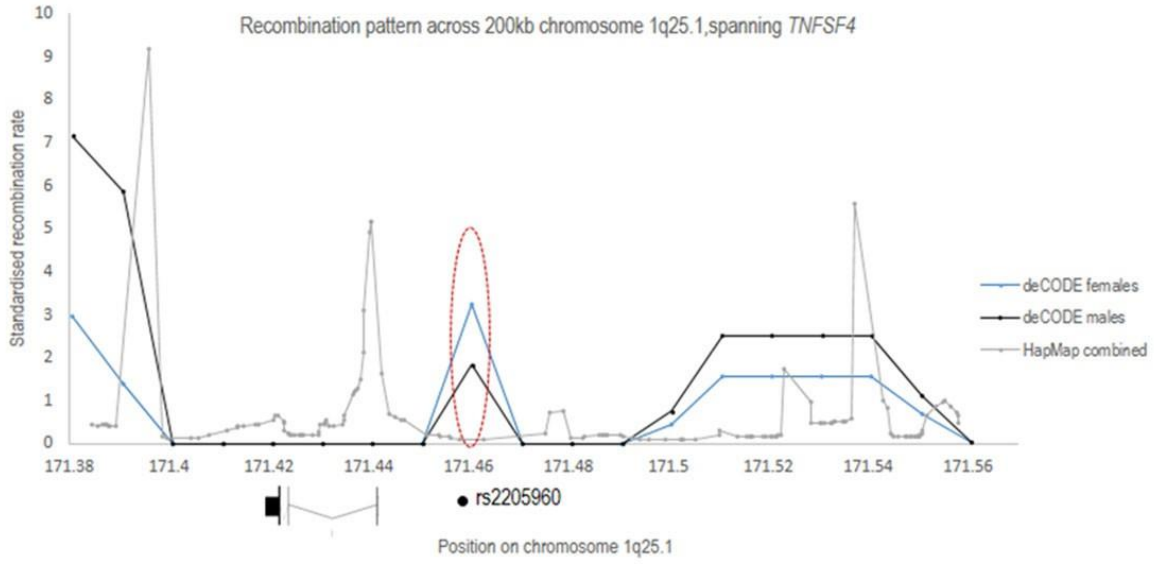
	C allele	T allele
ASW	0.87	0.13
AMR	0.69	0.31
ASN	0.77	0.23
EUR	0.75	0.25

1000 Genomes allele frequencies, *rs16845607*

	G allele	A allele
ASW	1	0
AMR	0.87	0.13
ASN	0.91	0.9
EUR	1	0

Appendix D:

Comparison of recombination at *TNFSF4*, deCODE females vs. deCODE males vs. HapMap combined data



Appendix E:

Test for cross-study heterogeneity

A logistic regression model fitted with an interaction term (effect) in the R statistical package was used to investigate cross-study heterogeneity for associated variants spanning *TNFSF4*. Variants were assessed across the African-American, East Asian, European and Hispanic SLE-control cohorts. P-values for individual associated SNPs were generated using the likelihood-ratio test. P-values are against homogeneity of odds ratio per marker for association data.

Marker	P-value (against homogeneity of odds ratios)
rs10489265	0.095
rs10912580	0.012
rs12039904	0.056
rs1234317	0.089
rs12405577	0.031
rs2205960	0.142

Table A2 Test for cross study heterogeneity

Appendix F:

Publications

Manku, H., Langefeld, C.D., Guerra, S.G., Malik, T.H., Alarcon-Riquelme, M., Anaya, J.M., Bae, S.C., Boackle, S.A., Brown, E.E., Criswell, L.A., Freedman, B.I., Gaffney, P.M., Gregersen, P.A., Guthridge, J.M., Han, S.H., Harley, J.B., Jacob, C.O., James, J.A., Kamen, D.L., Kaufman, K.M., Kelly, J.A., Martin, J., Merrill, J.T., Moser, K.L., Niewold, T.B., Park, S.Y., Pons-Estel, B.A., Sawalha, A.H., Scofield, R.H., Shen, N., Stevens, A.M., Sun, C., Gilkeson, G.S., Edberg, J.C., Kimberly, R.P., Nath, S.K., Tsao, B.P. and Vyse, T.J. (2013). Trans-Ancestral Studies Fine Map the SLE-Susceptibility Locus *TNFSF4*. PLoS Genet. *9*, e1003554.

Simpson, N., Gatenby, P.A., Wilson, A., Malik, S., Fulcher, D.A., Tangye, S.G., Manku, H., Vyse, T.J., Roncador, G., Huttley, G.A., Goodnow, C.C., Vinuesa, C.G. and Cook, M.C. (2010). Expansion of circulating T-cells resembling follicular helper T-cells is a fixed phenotype that identifies a subset of severe systemic lupus erythematosus. Arthritis Rheum. *62*, 234-44.

Manku, H., Graham, D.S. and Vyse, T.J. (2009) Association of the co-stimulator OX40L with systemic lupus erythematosus. J Mol Med (Berl) *87*, 229-34.

Cunninghame Graham, D.S., 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. and Vyse T.J. (2008). Polymorphism at the TNF superfamily gene TNFSF4 confers susceptibility to systemic lupus erythematosus. Nat Genet. *40*, 83-9.