De novo mutations implicate novel genes with burden of rare variants in Systemic Lupus Erythematosus

Venu Pullabhatla[1*], Amy L. Roberts[2*], Myles J. Lewis[3], Daniele Mauro[3], David L. Morris[2], Christopher A. Odhams[2], Philip Tombleson[2], Ulrika Liljedahl[4], Simon Vyse[2,7], Michael A. Simpson[2], Sascha Sauer[5,8], Emanuele de Rinaldis[1], Ann-Christine Syvänen[4], Timothy J. Vyse[2,6,9]

* Contributed equally to this work

[1] NIHR GSTFT/KCL Comprehensive Biomedical Research Centre, Guy's & St. Thomas' NHS Foundation Trust, London, UK

[2] Division of Genetics and Molecular Medicine, King's College London, London, UK

[3] Centre for Experimental Medicine and Rheumatology, William Harvey Research Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London, UK

[4] Department of Medical Sciences, Uppsala University, Uppsala, Sweden

[5] Max Planck Institute for Molecular Genetics, Berlin, Germany

[6] Division of Immunology, Infection and Inflammatory Disease, King's College London, London, UK

[7] Current address: Division of Cancer Biology, The Institute of Cancer Research, London, UK

[8] Current address: Max Delbrück Centre for Molecular Medicine (BIMSB/BIH), Berlin, Germany

[9] Corresponding author: timothy.vyse@kcl.ac.uk

Abstract

The heritability of most complex diseases, including autoimmune disease Systemic Lupus Erythematosus (SLE), remains largely unexplained by common variation, and few examples of rare variant associations have been identified. Here, using complementary whole-exome sequencing (WES) and high-density imputation, we identify candidate genes through *de novo* mutation discovery and demonstrate collective rare variant associations at novel SLE-susceptibility genes. Using extreme-phenotype sampling, we sequenced the exomes of 30 SLE parent-affected-offspring trios and identified 14 genes with missense *de novo* mutations, none of which are within the >70 SLE susceptibility loci implicated through genome-wide association studies (GWAS). In a follow-up cohort of 10,995 individuals of matched European ancestry, including 4,036 SLE cases, we imputed genotype data to the density of the combined UK10K-1000 genomes Phase III reference panel across the 14 candidate genes. We identify a burden of rare exonic variants across *PRKCD* associated with SLE risk (P=0.0028), and across *DNMT3A* associated with two severe disease prognosis sub-phenotypes (P=0.0005 and P=0.0033). Additionally, we show the p.His198Gln *de novo* mutation within the candidate gene *C1QTNF4* inhibits NF-κB activation following TNF exposure. Exome sequencing studies typically lack power to detect rare variant associations for complex traits. Our results support extreme-phenotype sampling and using *de novo* mutation gene discovery to aid the search for rare variation contributing to the heritability of complex diseases.

Introduction

Considerable progress has been made in elucidating the genetic basis of complex disease traits. The associated genetic polymorphisms identified are usually relatively common in the population and the risk alleles impart a modest individual increment to the likelihood of developing disease. The allelic identities of such genetic factors are established using large-scale genotyping chips with a predetermined composition of tagging variants. Advances in DNA sequencing technology, such as next-generation sequencing (NGS), enable the characterization of rare and unique genetic variants. Targeting the exome, which comprises approximately 1% of the genome, facilitates (by scale) analysis of NGS data and provides a highly enriched source of highly penetrant disease causing mutations[1]. The impact of NGS has been dramatic in monogenic disease; in a recent review it was stated that the genes underlying approximately half of known Mendelian disorders had been discovered[2]. In contrast, the role of rare genetic variants in complex diseases is unknown. It had been proposed that rare variants might explain, at least in part, the missing heritability that is frequently described in complex disease traits[3]. Yet very few examples of individual rare variants contributing to complex disease risk have been identified[4]. However, rare variants are collectively very common; indeed the Exome Aggregate Consortium (ExAC) have demonstrated that the vast majority of genetic variation is extremely rare[5]. Therefore, as opposed to case-control analyses on the *variant*-level – as routinely employed for common polymorphisms – the contribution of rare variation can be assessed on the *gene*-level by aggregating all observed rare variants in a defined region and performing burden tests.

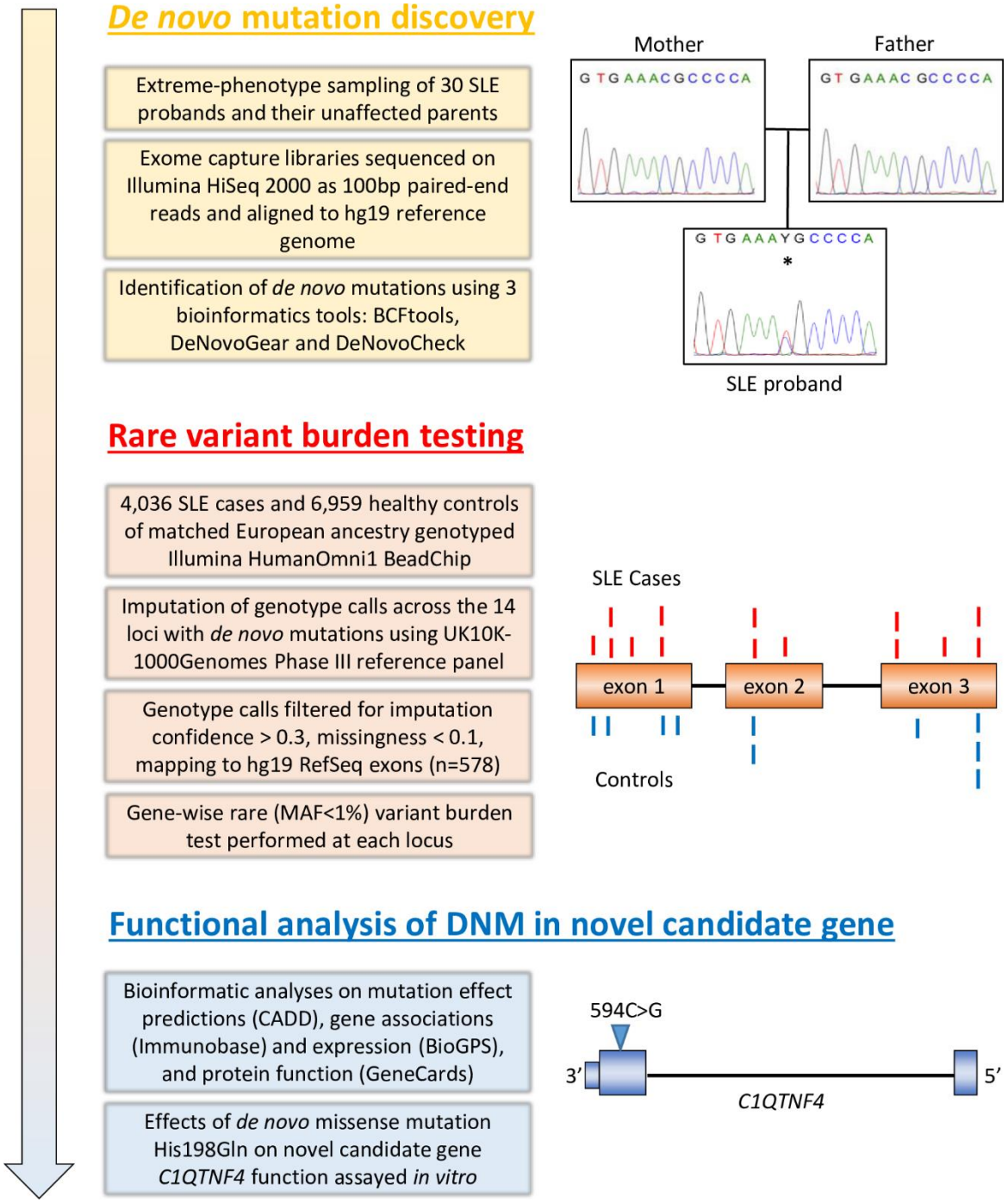
Although there are examples of disease predisposition genes, typically identified through genome-wide association studies (GWAS), harbouring associated variants across a spectrum of allele frequencies[4,6,7], studies focusing on canonical disease-associated loci have been far from fruitful, suggesting these loci by and large do not harbour additional risk through rare variation[8]. Exome-wide searches have similarly revealed limited numbers of rare variation associated with complex diseases; a recent large-scale whole-exome

sequencing case-control study in Type 2 Diabetes concluded that rare variants do not play a major role in disease predisposition[9]. Furthermore, widely used gene-based association tests have been shown to lack power at the exome-wide level, even with sample sizes up to 10,000, suggesting the need for reduced number of tested regions[10].

Our strategy to address this problem is outlined here and summarised in **Fig. 1**. We selected SLE cases with a severe phenotype (young age of onset and clinical features associated with poorer outcome) and hypothesized that these individuals would exhibit unique mutation events in their protein coding DNA that predisposed to disease risk. Therefore, we undertook whole exome sequencing (WES) in 30 family trios (both parents and affected offspring) and scrutinized the data for *de novo* mutations in the individual with SLE to identify a group of candidate genes for an independent follow-up rare variant analysis. This method allowed the identification of novel loci harbouring disease risk through collective rare variation.

Figure 1 Overview of study

De novo mutations (DNM) in a discovery cohort revealed candidate genes for imputation-based rare variant burden testing using a follow-up cohort. Independent functional analyses demonstrate the functional effects of one DNM in a candidate gene.



Results

Identification of de novo *mutations in SLE cases*

We screened for *de novo* mutations by WES of 30 family trios with an affected offspring with more severe SLE (**Supplementary Fig. 1**). A total of 584,798 variants (≥20X), including single nucleotide variants and indels, were identified in the 30 affected probands. Using three bioinformatic tools and employing conservative parameters, 26 putative exonic *de novo* mutations were identified across the 30 trios, at an average of 0.87 variants per proband, including 17 missense mutations across 17 genes (**Supplementary Table 1; Supplementary Fig. 2**). These data fit with the estimation that each exome contains one *de novo* mutation[11].

We also analysed the SLE proband WES data alone, without the unaffected parents. This revealed 1,194 non-silent, heterozygous, rare variants in 1,067 genes, which would make prioritisation for downstream analysis a difficult task, highlighting the benefit of parent-offspring trio sequencing (**Supplementary Fig. 3**).

Through Sanger sequencing all three members of the parent-offspring trio, plus any additional unaffected siblings (**Supplementary Table 2**), we confirmed 14 true non-silent *de novo* mutations (**Table 1**) in 11 of the 30 probands (36.7%) for further analysis. These true *de novo* mutations were absent in both parents and any unaffected siblings, but present in the SLE proband.

Of the three mutations that did not pass Sanger verification (**Supplementary Table 1**), one, within *LAMC2*, is likely a result of germline mosaicism because, although not observed in either parent, it is observed in an unaffected sibling in addition to the SLE proband[12]. It is therefore perhaps unlikely to be pathogenic and was not taken forward in downstream analyses. The two remaining putative mutations in *KLRC1* and *KRTAP10-2* are both member of highly homologous gene families. Such sequence identity may have caused false positive identification of *de novo* mutations in the WES analysis. Indeed the *KRLC1*

p.Ile225Met missense variant appears to be a polymorphic Paralogous Sequence Variant (PSV) – the paralogous variant being Met223Ile in *KLRC2* - once more making this variant unlikely to be pathogenic.

Supporting evidence of a role of candidate genes in SLE

We explored the function, expression (BioGPS), known associations with autoimmunity (ImmunoBase), and gene-level constraint against missense mutations (ExAC), of the genes with *de novo* mutations to build a profile of *a priori* evidence of a role in SLE pathogenesis (**Table 2**). The candidate genes include autoimmune susceptibility genes (*PRKCD, DNMT3A* and *ANXA3*), although none have been previously associated with SLE through GWAS in any population[13,14]. We also identify candidate genes through known/predicted function and expression profiles (*C1QTNF4, SRRM2, HMSD*), and four genes (*PRKCD, DNMT3A, C1QTNF4* and *LRP1)* with a significant (Z>3.09) constraint against missense variants (**Table 2**).

Functional characterisation of de novo *mutations*

To analyse the potential impact of the *de novo* mutations, we used the ExAC database[5] and Combined Annotation Dependent Depletion (CADD) scores[15] to characterise their frequency and predicted functional effects, respectively. Five of the 14 *de novo* mutations – found in *MICALL1*, *LRP1*, *PNPLA1*, *PLD1*, and *GFTP2* - have been observed, at very rare frequencies, in the ~60,000 exomes documented in ExAC (**Table 1**). All five mutations are CpG transitions and therefore likely to be identity-by-state, reflecting the higher mutability rate of these sites. Within the mutation set, five (35.7%) – found in *DNMT3A, PRKCD, MICALL1, LRP1,* and *PNPLA1* – have CADD Phred scores >30, placing them in the top 0.1% of possible damaging mutations in the human genome (**Table 1**)**.**

Independent gene-based analysis of rare variants in de novo *mutated genes*

We hypothesised that, while some observed *de novo* mutations were random background variation as present in the exome of every individual regardless of disease status[11], others could be contributing to disease through their deleterious effects on the encoded protein, and would thus be indicative of a hitherto unknown gene which contributes to SLE risk. Such putative genes could also harbour a burden of additional rare variation. Therefore, in a follow-up cohort of 10,995 individuals of matched European ancestry previously genotyped on the Illumina HumanOmni1 BeadChip[13], including 4,036 SLE cases, we imputed genotype data to the density of the combined UK10K and 1000 genomes Phase III reference panel (UK10K-1000GP3) across the 14 genes with *de novo* mutations to assay rare variation risk.

Using a collapsing burden test[16], we surveyed each of the 14 genes for an excess of aggregated rare (minor allele frequencies (MAF)<1%) exonic variants in SLE cases compared to healthy controls, and we identify an association of rare exonic variation in *PRKCD* with SLE (**Table 3**; P=0.0028). In sub-phenotype analyses, using healthy controls and only SLE cases with anti-dsDNA (n_{cases}=1261) or renal-involvement with hypocomplementemia (n_{cases}=186), both of which are markers of more severe disease, we identify collective rare exonic variants in *DNMT3A* associated with both anti-dsDNA (**Table 3;** P=0.0005) and renal involvement with hypocomplementemia (**Table 3;** P=0.0033). We also collapsed all exons from the 14 genes together to test for an overall burden of rare variants across these loci. These analyses revealed no excess of rare exonic variants across the grouped genes, reflecting the hypothesis that some/most genes will not be relevant to disease status because the observed *de novo* mutations are random background variation only.

Using gene-level constraint metric data from ExAC[5], *DNMT3A* and *PRKCD* are two of the four genes with *de novo* mutations with a significant constraint against missense variants (Z >3.09; **Table 2**). However, across the entire gene set, there was no difference in the median Z-score (0.50) compared with the median Z-score across all genes in ExAC (0.51). These

data reflect the results of our rare variant burden tests, in which the aggregated gene set do not contribute to disease risk.

Common variant analysis across de novo mutation genes

Loci harbouring autoimmune risk through both common and rare variants have been reported[6,17]. Therefore, using the high-density UK10K-1000GP3 imputed data, we reassessed the contribution of common variation to SLE risk across these loci. No significant association at any locus was observed with overall risk in a case-control comparison, as previously reported[13], nor with anti-dsDNA (n_{cases}=1261) or renal-involvement with hypocomplementemia (n_{cases}=186) sub-phenotypes (**Supplementary Table 3**). A candidate gene study previously reported a trend of association between the common *DNMT3A* intronic SNP rs1550117 (MAF~7%) and SLE in a European cohort[18]. Our analysis did not replicate this finding (*P*=0.23).

Underfunctioning of C1QTNF4 p.His198Gln

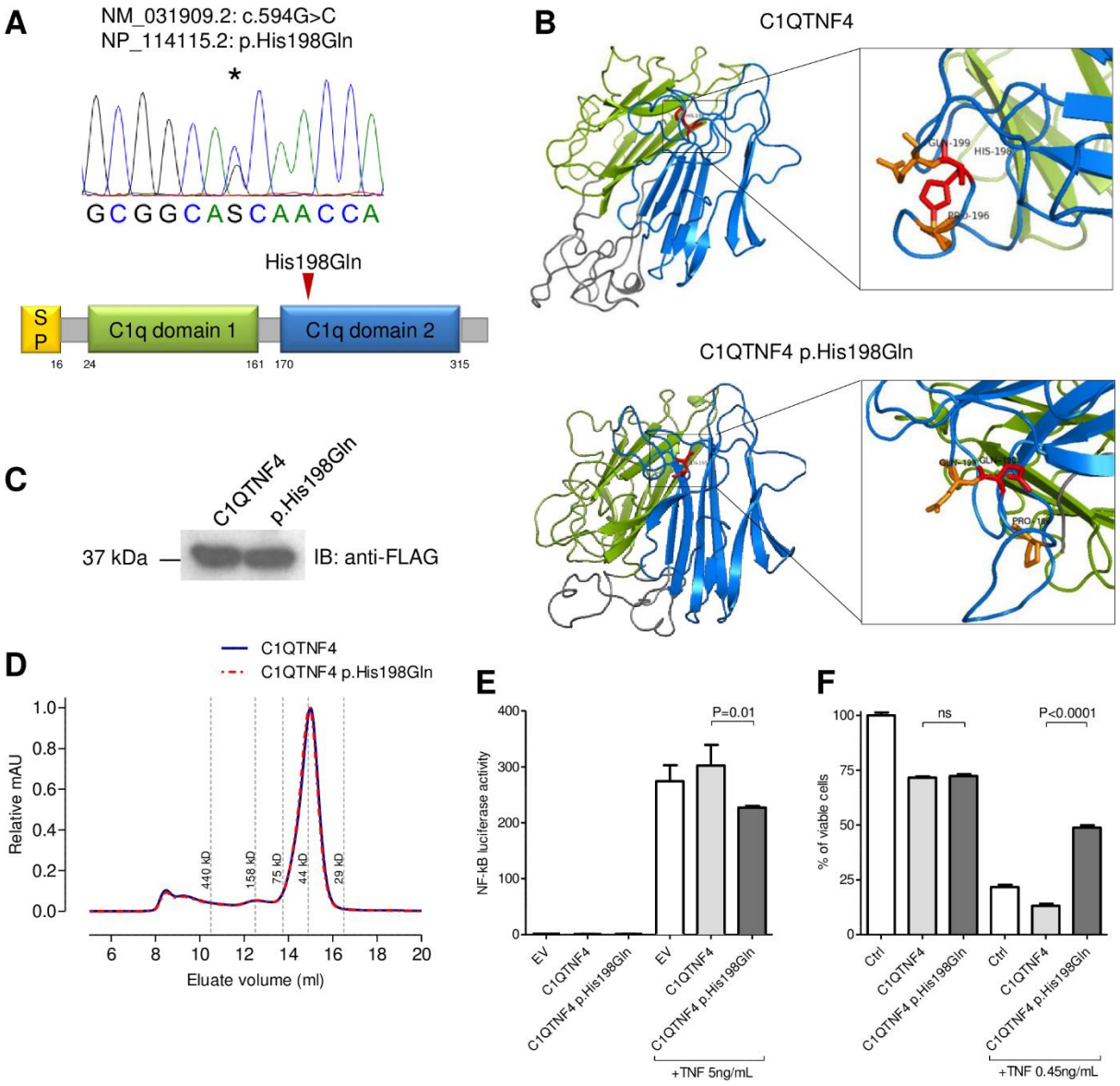
The candidate gene *C1QTNF4* is a very small gene with <1Kb coding sequence over two exons and is one of four genes constrained against missense variants (ExAC gene-level constraints Z=3.17, **Table 2**). Although gene coding length does not correlate with missense constraint scores[5], it may contribute to insufficient power to detect rare variant associations when using imputed data derived from reference haplotypes from healthy individuals alone.

The *de novo* mutation in *C1QTNF4* generates a His198Gln protein sequence change with a modest CADD score of 12.3 (**Table 1**). Although they are useful in the absence of suitable functional assays, the sensitivity of bioinformatic prediction tools is known to be suboptimal. Where functional assays are available, previous studies have also demonstrated functional effects of variants predicted to be tolerated/benign[19].

We therefore pursued a functional analysis of the His198Gln *de novo* mutation detected in the *C1QTNF4* gene. Although its function is rather poorly understood, the protein product, C1QTNF4 (CTRP4) is secreted and may act as a cytokine, as it has homology with TNF and the complement component C1q (**Fig. 2**). C1QTNF4 has been shown to influence NF-κB activation[20], a pathway known to be implicated in SLE pathogenesis. In order to study the effect of the *C1QTNF4* mutation, we looked for an effect on NF-κB production. Using a HEK293-NF-κB reporter cell line, we showed that C1QTNF4 p.His198Gln mutant protein was expressed and that it inhibited the NF-κB activation generated by exposure to TNFα (**Fig. 2**). Furthermore, we showed that the fibroblast L929 cell line, which is sensitive to TNF-induced cell death, was rescued by exposure to C1QTNF4 p.His198Gln, but not by wild type C1QTNF4. Thus, the mutant form of C1QTNF4 appears to inhibit some of the actions of TNF, which may promote antinuclear autoimmunity[21–23].

Figure 2. Structural and functional characterization of C1QTNF4 p.His198Gln substitution

(A) Domain organization of human C1QTNF4, showing signal peptide (yellow), first C1q domain (green), second C1q domain (blue) and linker peptides (grey). Arrow highlights substitution site. (B) 3D structure prediction of C1QTNF4 and C1QTNF4 p.His198Gln. Ribbons show the interaction between the positively charged Histidine 198 and Proline 196 lost in C1QTNF4 p.His198Gln due to the substitution of Histidine with Glutamine. (C) Immunoblot demonstrating that p.His198Gln does not affect secretion of C1QTNF4 in HEK293 supernatants. (D) Size exclusion chromatography profile showing no difference in oligomerisation between supernatant containing C1QTNF4 (blue) and C1QTNF4 p.His198Gln (red). (E) Luciferase assay in HEK293-NF-κB reporter cell line showing that C1QTNF4 p.His198Gln inhibits NF-κB activation in response to 4h stimulation with 5ng/mL TNFα. Error bars represent standard error of the mean. (F) Inhibition of L929 induced cell death by C1QTNF4 p.His198Gln after 24h of stimulation with 0.45 ng/mL TNFα in presence of Actinomycin 1µg/ml.



Discussion

Following the unexplained heritability left in the wake of massive GWAS in most complex diseases, searching the GWAS-identified canonical disease susceptibility genes for rare variants has added little to the heritability explained. Although there are examples – and perhaps more to discover – of canonical disease genes harbouring both common and rare risk alleles, the vast majority of such loci do not[6]. Indeed the common-variant associated loci which have also been shown to harbor rare coding variant risk are often those loci where the

common polymorphisms are non-silent coding variants (e.g. *NCF2*[8]). Our data suggest that rare genetic risk may be found in a discrete set of non-canonical susceptibility genes, as we report an association of collective rare variation at *PRKCD* and *DNMT3A*, and found no evidence of an association with common variants across these loci. This, to the best of our knowledge, is the first WES study in polygenic cases of autoimmune disease to use *de novo* mutation discovery to identify candidate genes for rare variant analyses.

We saw a relatively high drop-out rate (11.7%) from our NGS to Sanger sequencing confirmed mutations due to the two variants – *KRTAP10-2* and *KLRC1* - found in members of highly homologous gene families. This suggests our NGS error-prone genes (NEPG) filter, which removes loci known to be problematic for genome mapping during NGS analyses, should have been more conservative.

DNMT3A and *PRKCD*, although hitherto not associated with polygenic SLE, are known autoimmunity susceptibility loci; *DNMT3A* is associated with Crohn's disease (CD)[24] and *PRKCD* is associated with both CD and ulcerative colitis (UC)[25]. The notion that a locus could harbour common variants contributing to one autoimmune disease and rare variants contributing to another is intriguing, and could provide hypothesis-driven searches in the hunt for 'missing heritability'.

A study by Berlot et al. identified a functional missense variant p.G510S (c.G1528A) in *PRKCD* in a consanguineous family with monogenic SLE[26]. It was demonstrated that the *PRKCD*-encoded protein, PRCδ, was essential in the regulation of B cell tolerance and affected family members with the homozygous mutation had increased numbers of immature B cells. Our study implicates the role of rare variants in *PRKCD* in the broader context of SLE susceptibility, beyond a monogenic recessive disease model. Furthermore, *PRKCB*, another member of the protein kinase C gene family, has been implicated in SLE risk in a Chinese study[27].

297 *DNMT3A*, a DNA methyltransferase, is a very intriguing candidate gene for SLE as altered
298 patterns of DNA methylation are reported in autoimmune diseases[28], and hypomethylation of
299 apoptotic DNA has been reported to induce autoantibody production in SLE[29]. DNA
300 methylation changes are also associated with monozygotic twin discordance in SLE[30].
301 Although previously implicated through a candidate gene study[18], we found no evidence of a
302 common variant association at this locus. Instead we find an association of collective rare
303 variants and SLE sub-phenotypes.

304 Together with the results of Hunt et al., in which autoimmune cases were aggregated and
305 rare variants were found to play a negligible role[8], our study suggests the importance of
306 deep phenotyping and the potential role of rare variants in specific sub-phenotype, or indeed
307 autoimmune, manifestations. Despite progress with diagnosis and treatment, particular SLE
308 sub-phenotypes – including those used in this study - are still associated with reduced life
309 expectancy. Therefore, elucidating the specific underlying genetic risk is of paramount
310 importance.

311 Our rare variant study is likely to be underpowered given the burden testing was conducted
312 on imputation data from an, albeit large, reference panel of healthy individuals and therefore
313 will not include putative SLE-specific variants. Therefore, other loci with suggestive p-values
314 at candidate genes (e.g. *HMSD*) warrant follow-up analyses through sequencing and/or
315 SLE-weighted imputation reference panels.

316 Through two in vitro assays, we demonstrated the functional effect of a *de novo* variant,
317 His198Gln in *C1QTNF4*, despite this variant being predicted to be of little functional
318 importance across multiple prediction tools. We showed the mutated protein product of
319 *C1QTNF4*, C1QTNF4, inhibits some TNF-mediated cellular responses, including activation
320 of NF-κB and TNF-induced apoptosis. The role of TNF in SLE is complex and incompletely
321 understood, although, in this context, it is noteworthy that TNF inhibition may promote
322 antinuclear autoimmunity[22]. A recent transancestral Immunochip analysis found an
323 association at the chromosome 5-encoded *C1QTNF2* with SLE (Nature Communications, in

press), and chromosome 22-encoded *C1QTNF6* is a known susceptibility locus for Type 1 Diabetes and is implicated in Rheumatoid Arthritis[31,32]. Together these data suggest a potential role of the hitherto understudied *C1QTNF* superfamily of genes in autoimmunity.

Each human - regardless of the disease status - is estimated to have one *de novo* mutation in their exome[11]. The simple presence of a provisionally functional *de novo* mutation in a proband is therefore not sufficient evidence that it contributes to disease risk. A major challenge of WES studies, therefore, is how to differentiate between variants truly important to disease and background variation[33]. In light of recent studies which have demonstrated the limitations of large-scale exome-wide case-control studies in detecting rare variant contributions[9,34], our results support extreme-phenotype sampling and *de novo* mutation discovery to aid a hypothesis-driven search for rare variation contributing to the heritability of complex diseases.

Methods

Selection of trios for sequencing

SLE patients of European ancestry were selected from the UK SLE genetic repository assembled in the Vyse laboratory. The study cases have been subject to genome-wide genotyping as part of a GWAS[13]. The criteria for inclusion were as follows: age of onset of SLE < 25 years (median age 21 years); more marked disease phenotype as shown by either evidence for renal involvement as per standard classification criteria and/or the presence of hypocomplementemia and anti-dsDNA autoantibodies; and DNA available from both unaffected parents. Thirty trios (90 individuals) were studied by WES. Ethical approval for the research was granted by the NRES Committee London (12/LO/1273 and 06/MRE02/9).

Sequencing and alignment

Sequencing libraries were prepared from 1ug of DNA using the SureSelect XT Human All exon v4 +UTR kit (Agilent Technologies). The libraries were prepared according to the manufacturer's manual (SureSelect XT target enrichment system for Illumina paired end sequencing library, v1.4.1 Sept2012).The exome capture libraries from each individual in the trio were sequenced on Illumina HiSeq 2000 as 100bp paired-end reads. The resulting BCL files from the sequencer were processed with Illumina Casava software v1.8 to obtain paired end reads in FastQ format. The paired end FastQ reads were aligned to the human reference genome hg19 (GCRh37) using Novoalign v2.07.11 with the following parameters (-i 200 30 -o SAM -o SoftClip -k -a -g 65 -x 7). The resulting BAM file was processed to sort and remove PCR duplicates using Picard tools. Only reads uniquely aligned to the reference genome were considered for further analysis. At the end of this process one BAM file per individual was obtained totaling 90 BAM files (30 trios).

Quality control (QC) and variant annotation

Variants were retained if they passed the following criteria: (i) read depth ≥20x, (ii) located within exome-captured regions as annotated in Gencode (on-target) and NCBI RefSeq annotation[35]. Read depth estimation was performed using DepthOfCoverage from GATK tool[36], as evaluated using bedtools[37]. All but one family had 75% representation of the exome at 20X (**Supplementary Fig. 1)**. Variants passing these QC criteria were annotated using ANNOVAR[38]. Any variant observed in either ExAC, 1000 genomes, ESP6500, or in-house databases was considered polymorphic.

De novo variant calling

To screen for *de novo* genetic variants in the affected offspring, three different bioinformatics tools were used: BCFtools[39], DeNovoGear[40] and DeNovoCheck[41]. They are based on the SAMtools algorithm for genotype calling and call *de novo* variants in the data from parents and offspring in each family trio, and have previously been used in *de novo* variant identification studies[41,42]. For each identified *de novo* variant, BCFtools assigns a combined likelihood ratio score (CLR, range 1-255)) and DeNovoGear assigns a posterior probability score (PP_dnm, range 0.0 – 1.0), while DeNovoCheck flags a variant as 'denovo' without providing a score. Conservative threshold scores for ascertainment of *de novo* variation were applied: CLR ≥ 80 for BCFtools and PP_dnm ≥ 0.8 for DeNovoGear (**Supplementary Fig. 5**). 454 variants were identified at these thresholds. Eight additional variants that were identified by DeNovoCheck and validated by IGV, resulting in a total of 462 variants, which map to 257 genes. The variants were next filtered in the following sequential steps (**Supplementary Figure 2**):

1. Removal of NGS error prone genes (NEPG): genes previously reported as probable false positive signals from NGS studies due to high frequency of rearrangements, polymorphisms or present in multiple copies[43]
2. Fulfil a *de novo* pattern of inheritance: any variant that supports Het:Ref:Ref for Child:Father:Mother, respectively, was considered a potential *de novo* variant. We also further selected variants that did not contain any trace of alternate allele in any of the parents. IGVtools was used to count the number of non-reference bases at each identified variant position from both father and mother. Only variants with a zero count of non-reference bases in both father and mother were considered as very high quality variants and retained for further analysis.
3. Variant annotation: only non-silent variants (Missense, Nonsense, splicing and insertions/deletions) was retained for further analysis.

This process resulted in a total of 17 variants in 17 genes (**Supplementary Table 1**).

Analysis of whole exome sequencing (WES) in cases only

To quantify the advantage of using parent-offspring trios, the whole exome samples of the 30 probands were analysed alone. 584,798 variants with ≥20X coverage depth and within Gencode capture regions were identified. All stringent filters to help in refinement of variants that could be potentially causal were applied. These include two filters previously described in the *de novo* filtering approach (Non-NEPG and variant annotation) along with selection of heterozygous variants (based on the pattern of zygosity expected for *de novo* mutations) and non-polymorphic filters (variants not observed in control datasets). To note, this is unlike the trio analysis where variants were not filtered for non-polymorphic. Filters were applied sequentially 1) Non-NEPG 2) Variant annotation 3) Heterozygosity 4) Non-polymorphic and resulted in 1194 variants in 1067 genes (**Supplementary Figure 3**).

Sanger Sequencing confirmation

Primers were designed using Primer 3 to target the exon containing the *de novo* mutation. Primers and PCR conditions available on request. 10ng of DNA from SLE probands, any unaffected siblings and both parents was amplified with Hot Start Taq polymerase on a G-Storm Thermocycler. PCR products were first cleaned with EXO-SAP before BigDye labelling in a linear PCR. Samples were sequenced on an ABI 3300XL.

Imputation

Genotype data from 10,995 individuals of matched European ancestry, including 4,036 SLE cases, genotyped on the Illumina Chip Illumina HumanOmni1 BeadChip for a previous study[13] were used. These data had undergone quality control as previously described[13], including Principal Component Analysis (PCA) to account for population structure. The UK10K (REL-2012-06-02) plus 1000 Genomes Project Phase3 data (release 20131101.v5)

merged reference panel (UK10K-1000GP3) was accessed through the European Genome-phenome Archive (EGAD00001000776). The genotype data were imputed using the UK10K-1000GP3 reference panel across the coding regions of the 14 genes with *de novo* mutations plus a 2Mb flanking region. To increase the accuracy of imputed genotype calls, a full imputation without pre-phasing was conducted using IMPUTE2[44,45]. Imputed genotypes were filtered for confidence using an info score (IMPUTE2) threshold of 0.3 (**Supplementary Figure 6**). The most likely genotype from IMPUTE2 was taken if its probability was > 0.5. If the probability fell below this threshold, it was set as missing. Variants with >10% missing genotype calls were removed for further analysis. All individuals had <8% missing genotype data.

Rare Variant Burden Tests

Imputed data were filtered, using Plink v1.9, to include only variants mapping to coding exons of hg19 RefSeq transcripts. Plink/SEQv1.0[16] was used to run gene-wise burden testing with a MAF<1% threshold. A 5% false discovery rate was used for multiple testing correction.

Common variant association tests

Imputed data were filtered to include variants with MAF>1% and SNPTEST 2.5.2[46] was used to test for associations across the region spanning the encoded gene. Bonferroni correction was used for 3,000 tests across the loci (q=1.66E-5).

Plasmids

Myc-Flag-tagged *C1QTNF4* on the pCMV6 vector and the empty pCMV6 vector were purchased from OriGene. The mutant pCMV6-*C1QTNF4 C594G* (p.His198Gln) was

generated by site-directed mutagenesis (Quikchange II XL; Stratagene) according the manufacturer instructions: mutagenic primer: 5'-GCGAGTGGTTGCTGCCGCGGCCC-3' (Sigma Aldrich). The plasmids production was carried on in XL10-Gold Ultracompetent cells, isolated and purified using EndoFree Maxi Prep kit (Qiagen). All the plasmid ORFs were confirmed by full Sanger sequencing (GATC-Biotech). The expression and secretion of the flagged proteins was confirmed by western blot on cell lysates and supernatants with monoclonal anti-FLAG antibody (clone M2; Sigma-Aldrich).

Luciferase assays and TNF-induced programmed cell death

GloResponse NF-κB-RE-luc2P HEK293 cell line (Promega) and TNF-sensitive L929 fibrosarcoma cell line (ATCC) were cultured in Dulbecco's Modified Eagle Medium (DMEM) enriched with 10% fetal bovine serum (FBS) and 1% Penicillin/Streptomycin (complete DMEM) at 37°C, 5% CO_2. HEK293 were seeded 24 hours before transfection in antibiotic free DMEM in 96 wells plate (2×10^4 cells/well), transfected with either *C1QTNF4*, *C1QTNF4 C594G* or Empty Vector via Fugene HD (Promega). 48 hours after transfection the cell were left unstimulated or stimulated with TNFα 5 ng/ml (PeproTech) for 4 hours. Luciferase activity was assayed by One-Glo (Promega) on Berthold Orion luminometer, the values were normalized to cell viability measured by CellTiter Glo (Promega). L929 were challenged with TNFα 0.45 ng/ml and Actinomycin D 1µg/ml (R&D) for 24 hours in presence of C1QTNF4 or C1QTNF4 p.His198Gln containing media, cell viability was measured by CellTiter Glo.

Size exclusion chromatography

Supernatants (750 µl) of HEK293 producing C1QTNF4 or C1QTNF4 p.His198Gln were buffer exchanged in PBS on Zeba Spin Desalting Columns (Thermo Fisher) and 0.5 mL loaded on an AKTA FPLC with a Superdex 200 10/300 GL column (GE Healthcare). Absorbance was normalized to the maximum peak of each sample.

In silico protein structure prediction

The web-based service Protein Homology/AnalogY Recognition Engine (Phyre2) [47] was used for the protein structure prediction of C1QTNF4 and C1QTNF4 p.His198Gln. The PBD file produced was loaded on Pymol software for visualization (Schrödinger, LLC).

Data availability

WES data on 90 individuals – 30 parent-offspring trios – will be deposited at the European Genome-phenome Archive.

Acknowledgements

The work leading to these results received funding from the European Union FP7 programme (grant agreement nº 262055) via the European Sequencing and Genotyping Infrastructure (ESGI). Sequencing was performed by the SNP&SEQ Technology Platform in Uppsala, which is part of the National Genomics Infrastructure (NGI) hosted by Science for Life Laboratory in Sweden. This work was supported in part by the Swedish Research Council for Medicine and Health (grant nº E0226301) and by the Knut and Alice Wallenberg Foundation (KAW 2011.0073). We thank Johanna Lagensjö and Olof Karlberg for assistance with sequencing.

The research was funded/supported by the National Institute for Health Research (NIHR) Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust and King's College London

1. Bamshad, M. J. *et al.* Exome sequencing as a tool for Mendelian disease gene discovery. *Nat. Rev. Genet.* **12,** 745–755 (2011).

2. Chong, J. X. *et al.* The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. *Am. J. Hum. Genet.* **97,** 199–215 (2015).

3. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461,** 747–753 (2009).

4. Rivas, M. a *et al.* Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nat. Genet.* **43,** 1066–73 (2011).

5. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536,** 285–291 (2016).

6. Jordan, C. T. *et al.* Rare and common variants in CARD14, encoding an epidermal regulator of NF-kappaB, in psoriasis. *Am. J. Hum. Genet.* **90,** 796–808 (2012).

7. Diogo, D. *et al.* Rare, low-frequency, and common variants in the protein-coding sequence of biological candidate genes from GWASs contribute to risk of rheumatoid arthritis. *Am. J. Hum. Genet.* **92,** 15–27 (2013).

8. Hunt, K. a *et al.* Negligible impact of rare autoimmune-locus coding-region variants on missing heritability. *Nature* **498,** 232–5 (2013).

9. Fuchsberger, C. *et al.* The genetic architecture of type 2 diabetes. *Nature* **536,** 41–47 (2016).

10. Moutsianas, L. *et al.* The power of gene-based rare variant methods to detect disease-associated variation and test hypotheses about complex disease. *PLoS Genet.* **11,** e1005165 (2015).

11. Veltman, J. a. & Brunner, H. G. De novo mutations in human genetic disease. *Nat. Rev. Genet.* **13,** 565–575 (2012).

12. Rahbari, R. *et al.* Timing, rates and spectra of human germline mutation. *Nat. Genet.* **48,** 1–11 (2015).

13. Bentham, J. *et al.* Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. *Nat. Genet.* **47,** 1457–1464 (2015).

14. Morris, D. L. *et al.* Genome-wide association meta-analysis in Chinese and European individuals identifies ten new loci associated with systemic lupus erythematosus. *Nat. Genet.* **48,** (2016).

15. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46,** 310–315 (2014).

16. Purcell, S. M. PLINK/SEQ: A library for the analysis of genetic variation data. at <https://atgu.mgh.harvard.edu/plinkseq/>

17. Okada, Y. *et al.* Integration of sequence data from a consanguineous family with genetic data from an outbred population identifies PLB1 as a candidate rheumatoid arthritis risk gene. *PLoS One* **9,** 1–12 (2014).

18. Piotrowski, P., Grobelna, M. K., Wudarski, M., Olesinska, M. & Jagodzinski, P. P. Genetic variants of DNMT3A and systemic lupus erythematosus susceptibility. *Mod Rheumatol* **25,** 96–99 (2015).

19. Roberts, A. L. *et al.* Resequencing the susceptibility gene, ITGAM, identifies two functionally deleterious rare variants in systemic lupus erythematosus cases. *Arthritis Res. Ther.* **16,** R114 (2014).

20. Li, Q. *et al.* Identification of C1qTNF-related protein 4 as a potential cytokine that stimulates the STAT3 and NF-κB pathways and promotes cell survival in human cancer cells. *Cancer Lett.* **308,** 203–214 (2011).

21. Beigel, F. *et al.* Formation of antinuclear and double-strand DNA antibodies and frequency of lupus-like syndrome in anti-TNF-α antibody-treated patients with inflammatory bowel disease. *Inflamm. Bowel Dis.* **17,** 91–98 (2011).

22. Eriksson, C., Engstrand, S., Sundqvist, K.-G. & Rantapää-Dahlqvist, S. Autoantibody formation in patients with rheumatoid arthritis treated with anti-TNF alpha. *Ann. Rheum. Dis.* **64,** 403–7 (2005).

23. Pink, A. E., Fonia, A., Allen, M. H., Smith, C. H. & Barker, J. N. W. N. Antinuclear antibodies associate with loss of response to antitumour necrosis factor-alpha therapy in psoriasis: a retrospective, observational study. *Br. J. Dermatol.* **162,** 780–5 (2010).

24. Franke, A. *et al.* Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat. Genet.* **42,** 1118–25 (2010).

25. Jostins, L. *et al.* Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491,** 119–24 (2012).

26. Belot, A. *et al.* Protein kinase Cδ deficiency causes mendelian systemic lupus erythematosus with B cell-defective apoptosis and hyperproliferation. *Arthritis Rheum.* **65,** 2161–2171 (2013).

27. Sheng, Y.-J. *et al.* Follow-up study identifies two novel susceptibility loci PRKCB and 8p11.21 for systemic lupus erythematosus. *Rheumatology (Oxford).* **50,** 682–688 (2011).

28. Ballestar, E. Epigenetic alterations in autoimmune rheumatic diseases. *Nat. Rev. Rheumatol.* **7,** 263–71 (2011).

29. Wen, Z. K. *et al.* DNA hypomethylation is crucial for apoptotic DNA to induce systemic lupus erythematosus-like autoimmune disease in SLE-non-susceptible mice. *Rheumatology* **46,** 1796–803 (2007).

30. Javierre, B. M. *et al.* Changes in the pattern of DNA methylation associate with twin discordance in systemic lupus erythematosus. *Genome Res.* **20,** 170–9 (2010).

31. Onengut-Gumuscu, S. *et al.* Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. *Nat. Genet.* **47,** 381–6 (2015).

32. Murayama, M. A. *et al.* CTRP6 is an endogenous complement regulator that can effectively treat induced arthritis. *Nat Commun* **6,** 8483 (2015).

33. Cooper, G. M. & Shendure, J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat. Rev. Genet.* **12,** 628–40 (2011).

34. Luo, Y. *et al.* Exploring the genetic architecture of inflammatory bowel disease by whole genome sequencing identifies association at ADCY7. *Nat. Genet.* **49,** 186–192 (2016).

35. Frankish, A. *et al.* Comparison of GENCODE and RefSeq gene annotation and the impact of reference geneset on variant effect prediction. *BMC Genomics* **16(Suppl 8),** S2 (2015).

36. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43,** 491–8 (2011).

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

38. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38,** e164 (2010).

39. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).

40. Ramu, A. *et al.* DeNovoGear: de novo indel and point mutation discovery and phasing. *Nat. Methods* **10,** 985–987 (2013).

41. de Ligt, J. *et al.* Diagnostic exome sequencing in persons with severe intellectual disability. *N. Engl. J. Med.* **367,** 1921–9 (2012).

42. Nava, C. *et al.* De novo mutations in HCN1 cause early infantile epileptic encephalopathy. *Nat. Genet.* **46,** 640–5 (2014).

43. Fuentes Fajardo, K. V. *et al.* Detecting false-positive signals in exome sequencing. *Hum. Mutat.* **33,** 609–613 (2012).

44. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* **44,** 955–959 (2012).

45. Roshyara, N. R. *et al.* Comparing performance of modern genotype imputation methods in different ethnicities. *Sci. Rep.* **6,** 34386 (2016).

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

47. Kelley, L. A. & Sternberg, M. J. E. Protein structure prediction on the Web: a case study using the Phyre server. *Nat. Protoc.* **4,** 363–371 (2009).

Table 1: *De novo* mutations in SLE probands with extreme phenotypes

Family	Chr	Position (hg19)	Ref	Alt	Gene	Gene Description	Exon	Amino acid	MAF in ExAC[a]	PhyloP	CADD Phred	Mutation Type[b]
SLE0751	22	38336799	C	T	*MICALL1*	MICAL-like 1	16	Arg852Cys	1.5E-04	2.31	35	Ti CpG
SLE0496	3	53223122	G	A	*PRKCD*	protein kinase C, delta	16	Gly535Arg	-	2.85	34	Ti CpG
SLE0679	12	57588368	C	T	*LRP1*	Low-density lipoprotein receptor-related protein 1	50	Arg2693Cys	8.3E-06	1.34	34	Ti CpG
SLE0592	6	36260896	G	A	*PNPLA1*	patatin-like phospholipase domain containing 1	3	Arg166His	5.8E-05	2.56	33	Ti CpG
SLE0296	2	25457236	G	A	*DNMT3A*	DNA (cytosine-5-)-methyltransferase 3 alpha	19	Ala695Val	-	2.75	32	Ti CpG
SLE0571	4	79512728	G	T	*ANXA3*	annexin A3	7	Ser145Ile	-	2.56	25.2	Tv
SLE0679	3	171431716	G	A	*PLD1*	phospholipase D1, phosphatidylcholine-specific	9	Thr293Met	5.8E-05	2.68	25.1	Ti CpG
SLE0411	5	179743769	C	T	*GFPT2*	glutamine-fructose-6-phosphate transaminase 2	12	Val383Met	2.6E-05	1.48	23.4	Ti CpG
SLE0679	7	138968784	C	A	*UBN2*	ubinuclein 2	15	Pro1045Thr	-	2.75	18.46	Tv
SLE0080	16	2812426	C	T	*SRRM2*	serine/arginine repetitive matrix 2	11	Arg633Cys	-	0.77	14.32	Ti CpG
SLE0852	11	47611769	G	C	*C1QTNF4*	C1q and tumor necrosis factor related protein 4	2	His198Gln	-	2.09	12.29	Tv
SLE0321	18	61621642	G	A	*HMSD*	histocompatibility (minor) serpin domain containing	3	Ala25Thr	-	1.26	9.732	Ti
SLE0390	12	32369376	G	C	*BICD1*	bicaudal D homolog 1 (Drosophila)	2	Val137Leu	-	0.72	8.673	Tv
SLE0321	1	35251125	C	G	*GJB3*	gap junction protein, beta 3	2	Asp254Glu	-	-0.25	0.002	Tv

The mutations are ordered by level of severity, from most to least, predicted by CADD score

[a]Frequencies are presented from all 61,468 multiethnic individuals in ExAC because the *de novo* mutations observed in ExAC are likely to be identity-by-state not identity-by-descent.

[b]Tv = Transversion; Ti = Transition; Ti CpG = Transition within a CpG dinucleotide

Table 2: Evidence for role of *de novo* mutation gene in autoimmunity

Gene	Functional Candidate[a]	Association with SLE[b]	Associations with other AID[b]	Immune cell type with highest expression[c]	Missense Constraint[d]
PRKCD	B cell signaling and self-antigen induced B cell tolerance induction	Monogenic forms[26]	IBD, UC, CD[24]	Dendritic	**3.75**
DNMT3A	DNA methyltransferase	Candidate gene study[18]	CD[25]	-	**4.31**
C1QTNF4	Pro-inflammatory cytokine	-	-	CD34+	**3.17**
SRRM2	Spliceosome-associated pre-mRNA splicing	-	-	CD8+	No data
LRP1	Endo/Phagocytosis of apoptotic cells	-	-	-	**10.60**
HMSD	Minor histocompatibility antigen	-	-	n/a	0.25
UBN2	DNA binding	-	-	-	0.01
ANXA3	-	-	RA[17]	-	-0.37
PLD1	-	-	-	Lymphoblasts	-0.73
PNPLA1	-	-	-	-	0.27
GFTP2	-	-	-	-	1.59
BICD1	-	-	-	-	2.12
GJB3	-	-	-	-	-0.81
MICALL1	-	-	-	-	0.50

Genes appear in descending order of supporting evidence. UC=ulcerative colitis, CD=Crohn's Disease, IBD=inflammatory bowel disease, RA=Rheumatoid Arthritis

[a] See Supplementary Table 4

[b] See Supplementary Table 5

[c] See Supplementary Figure 4. Data from BioGPS. If gene expression is highest in immune cells compared to all other cells, the immune cell type with highest expression is listed.

[d] Gene-wise ExAC Constraint Z-scores. Genes with significant restraint against missense variants are highlighted in bold.

Table 3: Gene-based rare variant burden analyses

			SLE		Anti-dsDNA		Renal with hypocomplementemia	
Locus	**# variants**	**# minor alleles controls**	**# minor alleles cases**	**p-value**	**# minor alleles cases**	**p-value**	**# minor alleles cases**	**p-value**
LRP1	84	927	514	1.00	143	1.00	25	0.57
BICD1	68	673	397	0.38	147	0.50	22	0.45
UBN2	63	338	214	0.26	77	0.07	17	0.11
PLD1	55	910	530	0.60	153	0.24	23	1.00
SRRM2	37	380	188	1.00	57	1.00	12	0.29
MICALL1	29	146	75	1.00	16	1.00	6	0.22
GFTP2	25	350	212	0.35	72	0.26	13	0.14
DNMT3A	24	110	91	0.0075	38	**0.0005**	9	**0.0033**
PRKCD	13	69	69	**0.0028**	20	0.06	2	0.71
ANXA3	12	311	155	1.00	47	1.00	4	1.00
GJB3	11	145	71	1.00	19	1.00	1	1.00
PNPLA1	11	194	139	0.04	51	0.02	3	1.00
C1QTNF4	9	186	100	1.00	31	1.00	4	1.00
HMSD	5	10	8	0.71	6	0.03	0	1.00
Grouped	446	4749	2763	0.57	877	0.31	141	0.23

Individual genes are ordered by descending number of observed rare variants. Significant p-values (burden test) at 5% FDR are highlighted in bold (q=0.003)