Supplementary Information

1. Targeted sequencing, genotyping and quality control

1.1. Targeted sequencing

A custom SeqCap EZ Choice XL library (Roche NimbleGen) was designed to target exons and regulatory regions of 1,853 genes, as described in detail previously (1). In total, 32 Mb were targeted for sequencing. Sequencing libraries were prepared by ultrasonification of DNA from whole blood to 400 bp fragments (Covaris E220) followed by barcoding (NEXTflex-96 DNA barcode adapters, Bio Scientific). Samples were pooled in batches of 8, hybridized (Roche NimbleGen) and sequenced with 100 bp paired-end reads using Illumina HiSeq 2500 version 3 or 4 chemistry. Targeted sequencing of the samples included in the current study has been described previously (2-4).

1.2. Genotyping and quality control

Sequencing reads were mapped to the human hg19 reference using bwa mem (version 0.7.12), and duplicate reads were marked with Picard (version 1.92). Applying the GATK Best Practices workflow (GATK version 3.3.0) for variant discovery, indel realignment and base score recalibration was performed prior to variant discovery using HaplotypeCaller in gVCF mode, excluding samples with a mean target coverage < 10x. Joint genotyping on the whole study cohort of patients and healthy controls was performed using GATK GenotypeGVCFs, and bi-allelic single-nucleotide variants were next passed on for recalibration of SNV quality scores using VariantRecalibrator with a filter at tranche level 99.0. Genotype calls with read depth < 8 and genotype Phred quality score < 20 were excluded using VCFtools.

The genetic structure of the study participants was analysed with LASER using the Human Genome Diversity Project (HGDP) as reference population (5, 6). Study participants > 5 standard deviations outside of the mean of the European sub-population of the HGDP reference set were excluded, followed by recursive exclusion of subjects exceeding > 5 standard deviations of the remaining study subjects. Duplicate and first-degree related individuals were excluded based on relatedness analysed using KING (7). An extra filter on rate of missing data, heterozygosity ratio, transition-transversion ratio and singleton counts was applied to exclude extreme sample outliers (3). Finally, samples with a call rate < 80% were removed.

For genetic variants, a number of filters were applied. Heterozygous calls were kept if the allelic balance across all samples was > 0.2 and < 0.8, whereas SNVs deviating from Hardy-Weinberg equilibrium ($p_{\text{healthy control}} < 1x10^{-6}$), monomorphic sites, and variants called for < 90% of individuals were excluded. The remaining variants were tested for differential missingness between patients and healthy controls using PLINK (version 1.90) (8), and variants differing between groups (Bonferroni corrected, $p < 0.05$) were excluded.

Individuals passing genotyping QC were considered eligible for analysis of *C4* copy number and *HLA*.

2. Analysis of *C4* copy number

2.1. Structure of C4A/C4B

The human paralogous *C4* genes, *C4A* and *C4B*, are located in the *HLA* class III region on the p arm of chromosome 6, centromeric to *HLA* class I and telomeric to *HLA* class II. The two *C4* genes are both 20.6 kb long and code for 41 exons (Fig. 1). The reference sequences of the two genes differ at 18 positions (Table 1), thereby being 99.91% identical. Five nucleotide variants – leading to 4 amino acid substitutions in exon 26 (PCPVLD vs. LSPVIH) – are by definition used to distinguish *C4A* and *C4B* (Table 1) (9, 10). Some C4 genes

may contain a ~6 kb human endogenous retroviral (*HERV*) insertion between exon 9 and 10 (Fig. 1) (11), but considering that this region has not been targeted for sequencing as part of this study, copy number variation of the *HERV* insertion is not part of the current *C4* analysis.

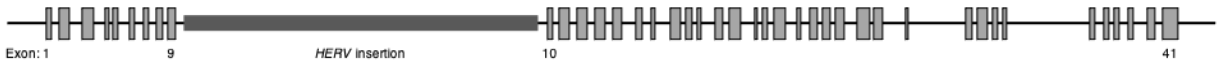

Fig. 1 Structure of the paralogous *C4* genes, *C4A* and *C4B*. The C4 genes may contain a ~6 kb human endogenous retroviral (*HERV*) insertion.

Table 1 Variants differing between the reference sequence for *C4A* and *C4B*. The 5 nucleotide variants in exon 26 (causing 4 amino acid substitutions) that per definition are used to define *C4A* and *C4B*, respectively, are marked in bold.

Position (GRCh37)		Position (GRCh38)		Allele		
C4A	*C4B*	*C4A*	*C4B*	*C4A*	*C4B*	Exon/Intron
31962174	31994912	31994397	32027135	A	G	Intron 20
31962401	31995139	31994624	32027362	G	A	Ala/Thr (exon 21)
31963559	31996297	31995782	32028520	A	G	Asp/Gly (exon 25)
31963860	**31996598**	**31996083**	**32028821**	**C**	**T**	**Pro/Leu (exon 26)**
31963863	**31996601**	**31996086**	**32028824**	**G**	**C**	**Cys/Ser (exon 26)**
31963871	**31996609**	**31996094**	**32028832**	**T**	**A**	**Leu/Ile (exon 26)**
31963874	**31996612**	**31996097**	**32028835**	**G**	**C**	**Asp/His (exon 26)**
31963876	**31996614**	**31996099**	**32028837**	**C**	**T**	
31964228	31996966	31996451	32029189	A	G	Asn/Ser (exon 28)
31964316	31997054	31996539	32029277	G	C	Ala/Ala (exon 28)
31964321	31997059	31996544	32029282	T	C	Val/Ala (exon 28)
31964330	31997068	31996553	32029291	T	G	Leu/Arg (exon 28)
31964331	31997069	31996554	32029292	C	G	
31964391	31997129	31996614	32029352	C	G	Intron 28
31964394	31997132	31996617	32029355	TC	T	Intron 28
31964785	31997522	31997008	32029745	T	G	Ser/Ala (exon 29)
31965242	31997979	31997465	32030202	T	C	Intron 30
31965383	31998120	31997606	32030343	A	G	Intron 30

2.2. Calling C4 copy number

C4 copy number was estimated using GATK GermlineCNVCaller (version 4.1.8.1), which is a read depth-based method for analysis of copy number variation in WES/targeted sequencing data using bam files as input. Prior to analysis, reads mapped to the *C4A*/*C4B* regions ± 500bp (hg19, chr6:31949334-32003695) were extracted (samtools version 1.10) and remapped (bwa mem version 0.7.17) to the reference sequence for chromosome 6 in which *C4A* ± 1,000 bp (chr6:31948834-31971457) had been masked. Next, the *C4* reads mapped to the *C4A*-masked reference were merged with chromosome 6 reads outside the *C4A*/*C4B* region ± 1,000 bp (chr6:1-31948834 and chr6:32004195-171115067). Before analysis in the GermlineCNVCaller pipeline, duplicate reads were marked using Picard (version 2.20.4).

Samples were analysed using the GATK GermlineCNVCaller pipeline in cohort mode with batches of size ~300. Forty-two samples with known *C4* copy number were included in all batches to allow for quality control and normalisation (see below). Intervals on chr6 targeted for sequencing were first split to have a maximum size of 5,000 bp. Intervals in the *C4B* region were manually defined to cover the relevant regions (chr6:31982572-31984923, chr6:31991707-31994992, chr6:31994993-31998278, chr6:31999328-32000075,

chr6:32001567-32003195), and a total number of 5,478 intervals on chromosome 6 were prepared using GATK PreprocessIntervals using default settings for targeted sequencing data. In the next step, the number of reads was analysed sample-wise for all intervals using CollectReadCounts, followed by AnnotateIntervals, FilterIntervals [--extreme-count-filter-maximum-percentile 100], DetermineGermlineContigPloidy, GermlineCNVCaller [--max-copy-number 8], and PostprocessGermlineCNVCalls (alternative settings defined in brackets).

The output from GermlineCNVCaller is a 'denoised copy ratio' for each interval across all individual samples. The total copy number of *C4* was estimated for each sample based on the average denoised copy ratio of the 5 *C4B* intervals. The copy number estimate was next normalised within each batch by linear regression using the samples with known *C4* copy number. Combining *C4* copy number estimates from all samples showed a multimodal distribution, corresponding to a *C4* copy number in the range 2-6 (Fig. 2), and the continuous estimate was rounded to the nearest integer copy number value. Two of the 42 samples with known *C4* copy number and included in all batches consistently showed a mismatch in the expected and the estimated copy number across all batches (see section 2.3). Including the two samples during normalisation had a negative impact on copy number estimation by 'pushing' the modes away from the integer copy number levels seen in Fig. 2, and therefore normalisation was performed using only the 40 samples that were consistent in the expected and estimated *C4* copy number across all analysis batches.

Two individuals had more than 6 estimated copies of *C4* (one SSA-SSB- pSS patient had a *C4* copy number of 7, and one control sample had a copy number of 8). Due to the low number of individuals with *C4* copy number ≥ 7, we excluded the two individuals from further analysis.

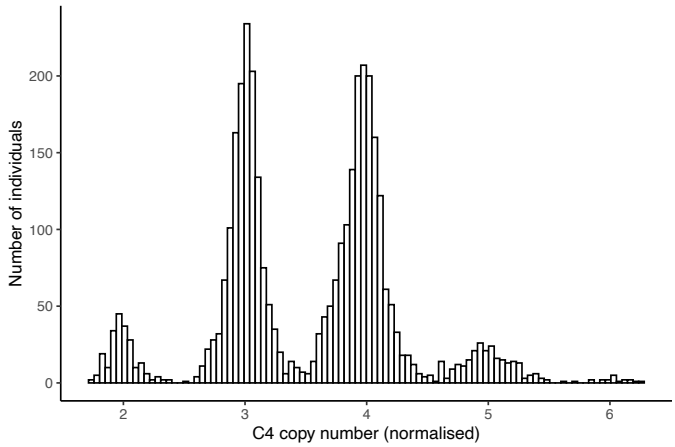

Fig. 2 *C4* copy number estimates of for healthy controls and patients with SLE, pSS and myositis (n = 3,541). Two individuals with copy number ≥ 7 have been excluded.

The proportion of *C4A* and *C4B* genes among the total number of *C4* genes was estimated based on the average read depth of the 5 paralog-specific variants (*C4B*: chr6 position 31996598T/C, 31996601C/G, 31996609A/T, 31996612C/G and 31996614T/C) analysed using GATK HaplotypeCaller. By plotting the estimate for total *C4* copy number against the read depth of *C4A*-specific variants relative to the total read depth of both *C4A*- and *C4B*-specific variants, samples generally clustered on the integer combinations of *C4A*/*C4B* copy number possible for each *C4* copy number level (Fig. 3).

Based on the total *C4* copy number, the integer copy number of *C4A* and *C4B* were calculated from their relative *C4A*-specific read depth using the relation: *C4* = *C4A* + *C4B*. *C4A*/*C4B* copy number was not estimated for

samples with a total read depth < 10 of the *C4A*-/*C4B*-defining variants, meaning that *C4A*/*C4B* copy number was not estimated for 21 individuals.

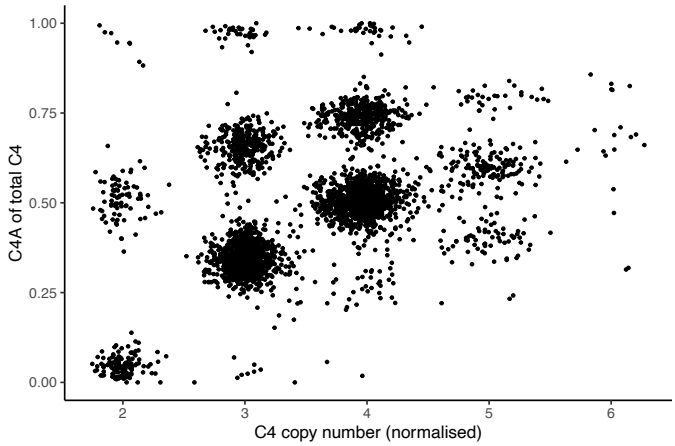

Fig. 3 *C4* copy number estimates plotted against the proportion of the read depth of *C4A*-specific variants relative to the total read depth of C4A-/C4B-specific variants (n = 3,520). 21 individuals with low read depth of *C4A*/*C4B*-defining nucleotides have been excluded.

2.3. Validation of C4 copy number calls

C4 copy number had been previously analysed for 120 SLE patients included in the current study by qPCR as described elsewhere (12, 13). Comparison of the lab-based *C4* copy number estimates and the copy number determined from sequencing data showed a high level of agreement (Fig. 4), with a match in total *C4* copy number for 115 of 120 samples (95.8%). In addition, a perfect match was found for the number of *C4A* and *C4B* copies when evaluating the 115 samples with matching *C4* copy number (Fig. 4).

For the 5 samples with a mismatch in *C4* copy number (red dots in Fig. 4), the estimate from GermlineCNVCaller were located close to a possible integer combination of *C4* vs. *C4A*/*C4B,* while being distantly located to the expected lab-based value. This suggests that the copy number value estimated from sequencing data could be the true value.

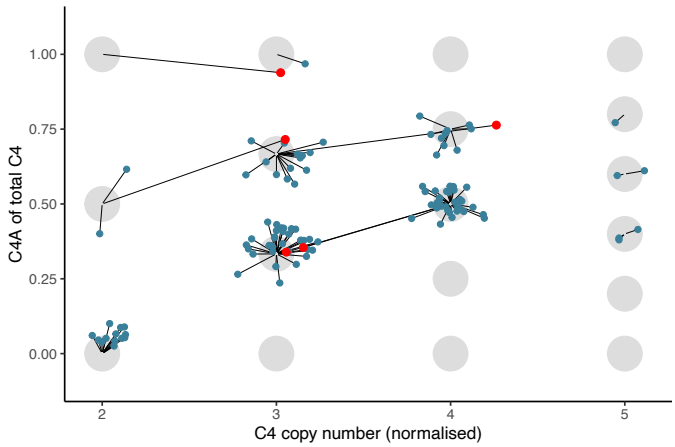

Fig. 4 *C4* copy number estimates based on calls from GATK GermlineCNVCaller in comparison to lab-based *C4* copy number detection. The x-axis denotes the total *C4* copy number, whereas the y-axis denotes the fraction of *C4A* of total *C4*. Blue and red dots denote an estimated *C4* copy number (using GermlineCNVCaller) and is connected with a line to an expected integer *C4* copy number value. Grey circles denote possible integer combinations for *C4* and *C4A*/*C4B* copy number. Five samples showed inconsistent *C4* copy number (marked in red; see description in text). n = 120.

2.4. Validation of C4 copy number from GATK GermlineCNVCaller using 1000 genomes WES data

C4 copy number has been previously determined by others using digital droplet PCR in 111 individuals from the 1000 Genomes Project (14). We tested the method for *C4* copy number analysis using GATK GermlineCNVCaller on whole exome sequencing (WES) data that was available for 90 of the 111 individuals. We selected WES data since this is targeted sequencing data that resemble the composition of the sequencing data applied in the current study.

The method largely followed the methodology applied for the DISSECT sequencing data with some minor differences. Briefly, WES data mapped to the GRCh38 reference sequence for the entire chromosome 6 and all *HLA* contigs were downloaded from the 1000 Genomes Project. Reads were remapped to GRCh38 in which the *C4A*-region ± 1kb (chr6:31981057-32003680) had been masked. After remapping, duplicate reads were marked using Picard.

Copy number was estimated with the GATK GermlineCNVCaller pipeline as described in section 2.2, analysing chromosome 6 regions targeted for sequencing in the 1000 Genomes project (intervals > 5,000 bp were split into smaller intervals). *C4* copy number was estimated based on the average denoised copy ratio of 39 *C4B* intervals targeted for sequencing in the region chr6:32014817-32035326 (note that *HERV* was not covered in WES data). The *C4* copy number estimates were normalised by linear regression using the *C4* copy number calls from digital droplet PCR. The proportion of *C4A* and *C4B* genes of the total number of *C4* genes was estimated based on the average read depth of the 5 paralog-specific variants (C4B: chr6 position 32028821T/C, 32028824C/G, 32028832A/T, 32028835C/G and 32028837T/C) analysed using GATK HaplotypeCaller after remapping to the *C4A*-masked reference.

C4 copy number was correctly called for 100% of the 1000 Genomes samples (90 of 90 samples had correct call) when comparing results from the WES data analysed by GATK GermlineCNVCaller to the lab-typed C4 copy number previously analysed by digital droplet PCR (Fig. 5).

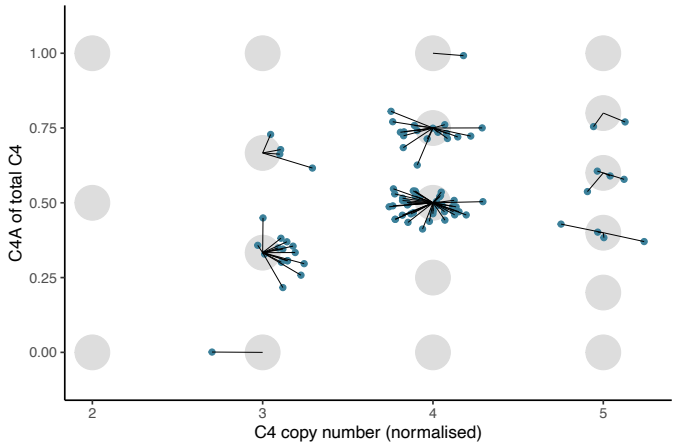

Fig. 5 *C4* copy number estimates in 90 samples from the 1000 Genomes project based on calls from GATK GermlineCNVCaller in comparison *C4* copy number analysed by droplet digital PCR by Sekar *et al.* 2016. The x-axis denoted the total *C4* copy number, whereas the y-axis denoted the fraction of *C4A* of total *C4*. Blue dots denote an estimated *C4* copy number (using GermlineCNVCaller) and is connected with a line to an expected integer *C4* copy number value as determined by digital droplet PCR (grey circles). n = 90.

2.5. Comparison of WGS calls vs. targeted sequencing calls

In a third attempt to validate the method for *C4* copy number estimation, we used WGS data from 75 parent-offspring trios, in which the offsprings were diagnosed with SLE (15). 45 of the offsprings with WGS sequencing data overlapped with the SLE patients in the current study. The analysis generally followed the

method for targeted sequencing data described in section 2.2, with some modifications. Briefly, reads mapped to a 5 Mb region of HLA (hg19; chr6:29000000-34000000), excluding duplicate reads, were extracted and remapped to the reference for chromosome 6, in which *C4A* ± 1,000 bp had been masked.

Intervals of 1000 bp size were generated for the 5 Mb HLA region using PreprocessIntervals, and intervals in *C4B* region were manually defined to ~1,000 bp intervals in the covering chr6:31982572-31984923 (*C4B* exon 1-9, 3 intervals), chr6:31991707-32003195 (*C4B* exon 10-41, 12 intervals), and chr6:31985199–31991567 (*HERV* sequence (16), 7 intervals). The residual analysis in GermlineCNVCaller was done as described in section 2.2. *C4* copy number was estimated based on the average denoised copy ratio of the 15 *C4B* intervals, and next normalised by linear regression using *C4* copy number estimates for SLE patients overlapping with samples described in section 2.2.

For the 45 samples with overlapping WGS data, a perfect match was seen for total *C4* copy number estimates. When comparing the copy number of *C4A*/*C4B*, a mismatch was seen in the *C4A*:*C4B* configuration for one individual with a copy number of 4 (2:2 vs. 1:3). Overall, a high consistency was seen for *C4* copy number calls when using different data types (WGS vs. targeted sequencing) for copy number analysis, thus corroborating the validity of our analysis approach

3. Analysis of variants in *C4*

3.1. Calling C4 variants

Due to the high sequence homology in the reference sequence for *C4A* and *C4B* (99.91% identical; section 2.1), most *C4* sequencing reads will map equally well to *C4A* and *C4B*. By default, reads that map to multiple locations are assigned a mapping quality (MAPQ) of 0 and are filtered when calling genotypes in tools such as GATK HaplotypeCaller. Consequently, genotypes from the *C4A*/*C4B* regions are generally not called from next generation sequencing data. In addition to the lack of variant discovery, the high level of copy number variation also complicates the genotyping analysis, in which diploid state is expected for autosomes.

To circumvent these challenges, we extracted the read depth of alternative variants after mapping *C4* reads to a *C4A*-masked reference (section 2.2) using GATK Haplotype Caller (version 4.1.8.1). Genetic variation at bp-resolution covering the regions chr6:31982472-31985023 (*C4B* exon 1-9) and chr6:31991607-32003295 (*C4B* exon 10-41) was analysed for individual samples. Next, variation was merged across all samples using GATK CombineGVCFs followed by GenotypeGVCFs.

Plotting the fraction of alternative alleles stratified for total *C4* copy number generally showed a pattern in which the frequency of the alternative allele was compliant with an integer copy number (Fig. 6), i.e. for a total *C4* copy number of 3, alternative allele ‘bands’ are seen at a fraction of 0.0, 0.33, 0.67 and 1.0. By using total *C4* copy number estimated for each individual, we converted the relative fraction of reads with an alternative allele into the nearest integer value defining *C4* gene copies with an alternative variant. To call a variant, we set a minimum threshold for total read depth at 14 together with a minimum alternative read depth of 7.

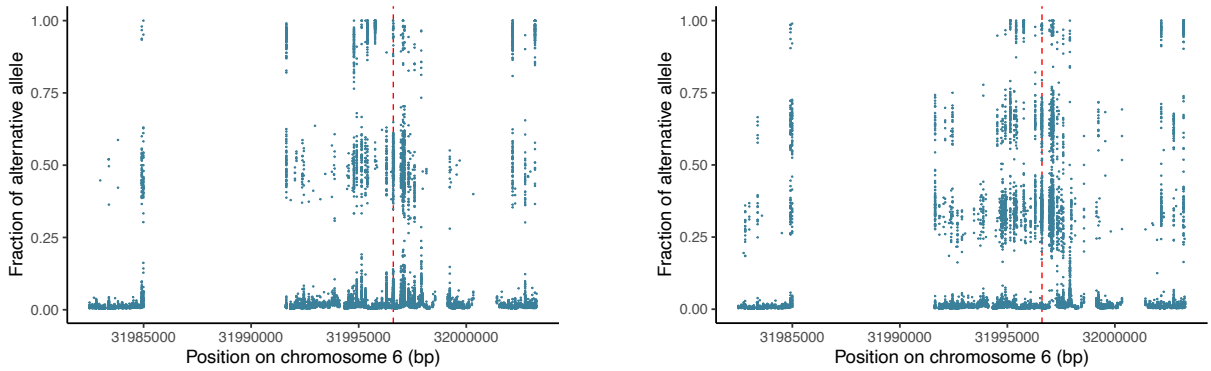

Fig. 6 Fraction of reads coding for alternative allele when stratifying for a total *C4* copy number of 2 (left panel) or 3 (right panel). Positions at which an alternative variant has been called for at least one individual have been included, and variants with fraction of alternative allele > 0 at the selected positions are been plotted. The number of points has been randomly downsampled to 10,000 points. The red dashed line indicates the position of the 5 *C4A*/*C4B*-defining nucleotide variants. The large region around 31.99 Mb without any variation is the location of the *HERV* insertion.

We used Ensembl Variant Effect Predictor version 99 (17) for annotation of variants. As exons were targeted for sequencing to a higher extent than introns, we focused the analysis on rare coding *C4* variants with a potentially stronger impact on disease involvement. Nevertheless, due to the high sequence homology between the reference sequences of *C4A* and *C4B* and the short read sequencing data, it was generally not possible to assign a variant specifically to a *C4A* or *C4B* gene, and therefore we analysed the detected variants as ambiguous *C4* variants.

3.2. Validation of rare C4 variants

Using the WGS data for 75 parent-offspring trios in which the offsprings were diagnosed with SLE, we aimed to verify the calls of the rare/semi-rare *C4* variants (alternative allele present in ≤ 10% of individuals). Comparison of samples overlapping with both targeted sequencing data and WGS data (n = 45) showed an allelic match of 98.4% (127/129 variants) in the variants called in both datasets. Further, 82 of the 86 variants called among 75 offsprings (95.3%) could be traced back to their parents in the correct allelic number.

The *C4A* loss-of-function variant rs760602547 (G/GCT) showed complete allelic match when comparing samples with overlapping targeted sequencing data and WGS data ($n_{individuals)}$ = 6), and when comparing transmission from parents to offspring in WGS trios (n_{trios} = 10).

4. Validation of *HLA* allele calls

4.1. Calling of HLA

HLA alleles of the 6 genes *HLA-A*, *-B*, *-C*, *-DPB1*, *-DQB1* and *-DRB1* were called at 2-field (i.e. 4-digit) resolution from sequencing data using xHLA (18). Prior to analysis, reads in the extended *HLA* region (chr6:29-34mb) and unmapped reads were remapped to chromosome 6 of the GRCh38 reference, and duplicate reads were discarded.

4.2. Validation using lab-typed HLA calls for DRB1

For 61 individuals, *HLA-DRB1* had been previously typed at 2-field (4-digit) resolution by direct Sanger sequencing (BGI, Shenzhen, People's Republic of China). Comparison of lab-typed *DRB1* alleles and calls from targeted sequencing data using xHLA (18) showed a 100% allelic match at 2-field resolution.

4.3. Transmission of HLA alleles from parents to offspring in WGS trios

As additional approach to validate the tool xHLA (18) for calling *HLA* alleles, we assessed whether *HLA* alleles called in offsprings could be traced back to their parents by using WGS data from 75 parent-offspring trios in which the offspring had been diagnosed with SLE (15). Briefly, read mapped to chr6:29000000-34000000 (GRCh37) and unmapped reads were remapped to GRCh38, duplicate reads discarded, and 2-field (i.e. 4-digit) alleles were called for *HLA-A*, *-B*, *-C*, *-DPB1*, *-DQB1* and *-DRB1* using the tool xHLA.

We next evaluated whether the alleles called among the 75 offsprings could be traced back to the parents for each of the 6 *HLA* genes (900 *HLA* alleles in total). The correct transmission pattern was observed for 99.6% of all alleles with 4 wrong allele calls in either the offspring or one of the parents (*HLA-A* for one trio, *-DQB1* for one trio, and *-DPB1* for two trios), therefore showing a high agreement for *HLA* calls.

4.4. Comparison of HLA calls from WGS vs. targeted sequencing data

We next compared 2-field *HLA* calls for overlapping samples with both targeted sequencing data and WGS data (n = 45 SLE patients). For the 6 *HLA* genes, allelic match between calls from targeted sequencing data and WGS data was seen for 95.0% of all genes, and for individual genes, an allelic match of 97.8% (*HLA-A*), 96.7% (*-B*), 84.8% (*-C*), 94.6% (*-DPB1*), 98.8% (*-DQB1*), and 97.8% (*-DRB1*) was seen.

5. C4b deposition on heat-aggregated human IgG

MaxiSorp plates were coated overnight at 4°C with 5 µg/ml heat-aggregated IgG in phosphate buffered saline (PBS) pH 7.4. PBS with 1% bovine serum albumin (BSA) was coated as control. Plates were blocked for 2 hours at room temperature with 1% BSA in PBS. Sera were diluted in GVB++ buffer (2.5 mM veronal buffer [pH 7.3], 150 mM NaCl, 0.1% gelatin, 1 mM $MgCl_2$, 0.15 mM $CaCl_2$) and incubated for 20 min. at room temperature. Deposited C4b was detected with a polyclonal rabbit anti-human C4c antibody (Dako cat. Q0369) followed by HRP-conjugated polyclonal swine anti-rabbit antibody (Dako cat. P0399). Plates were developed using TMB one (Kementec), and absorbance was measured at 450 nm with 620 nm as reference wavelength using a Cytation-5 multi-mode reader (BioTek). Unspecific binding to 1% BSA resulted in $Abs_{450nm-620nm} < 0.02$.

6. Consortia

6.1. The DISSECT consortium

Lars Rönnblom (Department of Medical Sciences, Rheumatology, Uppsala University, Uppsala, Sweden), Gunnel Nordmark (Department of Medical Sciences, Rheumatology, Uppsala University, Uppsala, Sweden), Ingrid E. Lundberg (Division of Rheumatology, Department of Medicine Solna, Karolinska Instituttet, Karolinska University Hospital, Stockholm, Sweden), Johanna K. Sandling (Department of Medical Sciences, Rheumatology, Uppsala University, Uppsala, Sweden), Pascal Pucholt (Department of Medical Sciences, Rheumatology, Uppsala University, Sweden), Lina Hultin Rosenberg (Science for Life Laboratory, Department of Medical Biochemistry and Microbiology, Uppsala University, Uppsala, Sweden), Sergey V. Kozyrev (Science for Life Laboratory, Department of Medical Biochemistry and Microbiology, Uppsala University, Uppsala, Sweden), Maija-Leena Eloranta (Department of Medical Sciences, Rheumatology, Uppsala University, Uppsala, Sweden), Andrei Alexsson (Department of Medical Sciences, Rheumatology, Uppsala University, Uppsala, Sweden), Matteo Bianchi (Science for Life Laboratory, Department of Medical Biochemistry and Microbiology, Uppsala University, Uppsala, Sweden), Christine Bengtsson (Department of

Public Health and Clinical Medicine/Rheumatology, Umeå University, Umeå, Sweden), Roland Jonsson (Broegelmann Research Laboratory, Department of Clinical Science, University of Bergen, Bergen, Norway), Roald Omdal (Department of Internal Medicine, Stavanger University Hospital, Stavanger, Norway), Øyvind Molberg (Department of Rheumatology, Oslo University Hospital, Oslo, Norway), Ann-Christine Syvänen (Department of Medical Sciences, Molecular Medicine and Science for Life Laboratory, Uppsala University, Uppsala, Sweden), Andreas Jönsen (Department of Clinical Sciences Lund, Rheumatology, Lund University, Skåne University Hospital, Lund, Sweden), Iva Gunnarsson (Division of Rheumatology, Department of Medicine Solna, Karolinska Institutet, Karolinska University Hospital, Stockholm, Sweden), Elisabet Svenungsson (Division of Rheumatology, Department of Medicine Solna, Karolinska Institutet, Karolinska University Hospital, Stockholm, Sweden), Solbritt Rantapää-Dahlqvist (Department of Public Health and Clinical Medicine/Rheumatology, Umeå University, Umeå, Sweden), Anders A. Bengtsson (Department of Clinical Sciences Lund, Rheumatology, Lund University, Skåne University Hospital, Lund, Sweden), Christopher Sjöwall (Department of Biomedical and Clinical Sciences, Division of Inflammation and Infection, Linköping University, Linköping, Sweden), Dag Leonard (Department of Medical Sciences, Rheumatology, Uppsala University, Uppsala, Sweden), Kerstin Lindblad-Toh (Science for Life Laboratory, Department of Medical Biochemistry and Microbiology, Uppsala University, Uppsala, Sweden and Broad Institute of MIT and Harvard, Cambridge, MA, USA), Jonas Carlsson Almlöf (Department of Medical Sciences, Molecular Medicine and Science for Life Laboratory, Uppsala University, Uppsala, Sweden), Niklas Hagberg (Department of Medical Sciences, Rheumatology, Uppsala University, Uppsala, Sweden), Jennifer R. S. Meadows (Science for Life Laboratory, Department of Medical Biochemistry and Microbiology, Uppsala University, Uppsala, Sweden), Jessika Nordin (Science for Life Laboratory, Department of Medical Biochemistry and Microbiology, Uppsala University, Uppsala, Sweden), Marie Wahren-Herlenius (Division of Rheumatology, Department of Medicine Solna, Karolinska Instituttet, Karolinska University Hospital, Stockholm, Sweden and Broegelmann Research Laboratory, Department of Clinical Science, University of Bergen, Norway), Sule Yavuz (Department of Medical Sciences, Rheumatology, Uppsala University, Uppsala, Sweden), Anna Tjärnlund (Department of Medicine, Rheumatology unit, Karolinska Institutet, Stockholm, Sweden), Antonella Notarnicola (Division of Rheumatology, Department of Medicine Solna, Karolinska Instituttet, Karolinska University Hospital, Stockholm, Sweden), Daniel Hammenfors (Department of Rheumatology, Haukeland University Hospital, Bergen, Norway), Elke Theander (Department of Rheumatology, Skåne University Hospital Malmö/Lund University, Lund, Sweden), Eva Baecklund (Department of Medical Sciences, Rheumatology, Uppsala University, Uppsala, Sweden), Guðný Ella Thorlacius (Department of Medicine, Unit for Experimental Rheumatology, Karolinska Institutet, Stockholm, Sweden), Hector Chinoy (Rheumatology Dept, Salford Royal NHS Foundation Trust, Manchester Academic Health Science Centre, Salford, UK and National Institute for Health Research Manchester Biomedical Research Centre, Manchester University NHS Foundation Trust, The University of Manchester, Manchester, UK), Helena Andersson (Department of Rheumatology, Oslo University Hospital, Oslo, Norway), Helena Enocsson (Department of Biomedical and Clinical Sciences, Division of Inflammation and Infection, Linköping University, Linköping, Sweden), Helena Forsblad-d'Elia (Department of Rheumatology and Inflammation Research, Sahlgrenska Academy at University of Gothenburg, Gothenburg, Sweden), Janine Lamb (Centre for Integrated Genomic Medical Research (CIGMR) , University of Manchester, Manchester, UK), Johan G. Brun (Department of Rheumatology, Haukeland University Hospital, University of Bergen, Bergen, Norway), Jonas Wetterö (Department of Biomedical and Clinical Sciences, Division of Inflammation and Infection, Linköping University, Linköping, Sweden), Jorge I. Ramírez Sepúlveda

(Department of Medicine, Unit for Experimental Rheumatology, Karolinska Institutet, Stockholm, Sweden), Juliana Imgenberg-Kreuz (Department of Medical Sciences, Rheumatology, Uppsala University, Uppsala, Sweden), Karin Hjorton (Department of Medical Sciences, Rheumatology, Uppsala University, Uppsala, Sweden), Karl A. Brokstad (Broegelmann Research Laboratory, Department of Clinical Science, University of Bergen, Bergen, Norway), Kathrine Skarstein (The Gade Laboratory for Pathology, Department of Clinical Medicine, University of Bergen, Norway), Katrine Brække Norheim (Department of Internal Medicine, Stavanger University Hospital, Stavanger, Norway), Lilian Vasaitis (Department of Medical Sciences, Rheumatology, Uppsala University, Uppsala, Sweden), Louise Pyndt Diederichsen (Center for Rheumatology and Spine Disease, Copenhagen University Hospital, Rigshospitalet, Copenhagen, Denmark and Department of Rheumatology, Odense University Hospital, Odense, Denmark), Malin V. Jonsson (Section for Oral and Maxillofacial Radiology, Department of Clinical Dentistry, University of Bergen, Bergen, Norway), Marika Kvarnström (Division of Rheumatology, Department of Medicine Solna, Karolinska Instituttet, Karolinska University Hospital, Stockholm, Sweden and Academic Specialist Center, Center for Rheumatology, Stockholm Health Services, Region Stockholm, Stockholm, Sweden), Maryam Dastmalchi (Center for Molecular Medicine, Karolinska Institutet, Stockholm, Sweden), Per Eriksson (Department of Biomedical and Clinical Sciences, Division of Inflammation and Infection, Linköping University, Linköping, Sweden), Robert G. Cooper (Institute of Ageing and Chronic Disease, University of Liverpool, Liverpool, UK), Sara Magnusson Bucher (Department of Rheumatology, Faculty of Medicine and Health, Örebro University, Örebro, Sweden), Silke Appel (Broegelmann Research Laboratory, Department of Clinical Science, University of Bergen, Bergen, Norway), Simon Rothwell (Institute of Inflammation and Repair, University of Manchester, Manchester, UK), Svein Joar Johnsen (Department of Internal Medicine, Stavanger University Hospital, Stavanger, Norway), Thomas Mandl (Department of Clinical Sciences Malmö, Division of Rheumatology, Lund University, Malmö, Sweden), Lara Adnan Aqrawi (Department of Oral Surgery and Oral Medicine, Institute of Clinical Odontology, University of Oslo, Oslo, Norway and Department of Health Sciences, Kristiania University College, Oslo, Norway), Janicke Liaaen Jensen (Department of Oral Surgery and Oral Medicine, Institute of Clinical Odontology, University of Oslo, Oslo, Norway), Øyvind Palm (Department of Rheumatology, Oslo University Hospital, Oslo, Norway), Maria Liden (Department of Medical Sciences, Rheumatology, Uppsala University, Uppsala, Sweden), Thomas Skogh (Department of Biomedical and Clinical Sciences, Division of Inflammation and Infection, Linköping University, Linköping, Sweden), Balsam Hanna (Department of Rheumatology, Sahlgrenska University Hospital, Gothenburg, Sweden), Christina Ståhl Hallengren (Rheumatology Unit, Helsingborg Hospital, Helsingborg, Sweden), Helena Hellström (Department of Rheumatology, Falu Hospital, Falun, Sweden), Åsa Häggström (Rheumatology clinic, Kalmar Hospital, Kalmar, Sweden), Aladdin Mohammad (Department of Clinical Sciences Lund, Rheumatology, Lund University, Skåne University Hospital, Lund, Sweden), Tomas Husmark (Department of Rheumatology, Falu Hospital, Falun, Sweden), Anna Svärd (Centre of Clinical Research Dalarna, Uppsala University, Falun, Sweden), Awat Jalal (Department of Rheumatology, Örebro University, Örebro, Sweden).

Hector Chinoy is supported by the National Institute for Health Research Biomedical Research Centre Funding Scheme. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health

6.2. The ImmunoArray development consortium

Kerstin Lindblad-Toh (Science for Life Laboratory, Department of Medical Biochemistry and Microbiology, Uppsala University, Uppsala, Sweden and Broad Institute of MIT and Harvard, Cambridge, MA, USA), Gerli Rosengren Pielberg (Science for Life Laboratory, Department of Medical Biochemistry and Microbiology, Uppsala University, Uppsala, Sweden), Anna Lobell (Office for Medicine and Pharmacy, Uppsala University, Uppsala, Sweden), Åsa Karlsson (Science for Life Laboratory, Department of Medical Biochemistry and Microbiology, Uppsala University, Uppsala, Sweden), Eva Murén (Science for Life Laboratory, Department of Medical Biochemistry and Microbiology, Uppsala University, Uppsala, Sweden), Göran Andersson (Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, Sweden), Kerstin M. Ahlgren (Department of Surgical Sciences, Uppsala University, Uppsala, Sweden), Lars Rönnblom (Department of Medical Sciences, Rheumatology, Uppsala University, Uppsala, Sweden), Maija-Leena Eloranta (Department of Medical Sciences, Rheumatology, Uppsala University, Uppsala, Sweden), Nils Landegren (Department of Medical Biochemistry and Microbiology, Uppsala University, Uppsala, Sweden and Centre for Molecular Medicine, Department of Medicine (Solna), Karolinska Institute, Stockholm, Sweden), Olle Kämpe (Department of Medicine (Solna), Center for Molecular Medicine, Karolinska Institutet, Stockholm, Sweden, Department of Endocrinology, Metabolism and Diabetes Karolinska University Hospital, Stockholm, Sweden and KG Jebsen Center for autoimmune diseases, University of Bergen, Norway), Peter Söderkvist (Division of Cell Biology, Department of Biomedical and Clinical Sciences, Linköping University, Linköping, Sweden).

7. References

1. Eriksson D, Bianchi M, Landegren N, Nordin J, Dalin F, Mathioudaki A, et al. Extended exome sequencing identifies BACH2 as a novel major risk locus for Addison's disease. Journal of Internal Medicine. 2016;280(6):595-608.
2. Thorlacius GE, Hultin-Rosenberg L, Sandling JK, Bianchi M, Imgenberg-Kreuz J, Pucholt P, et al. Genetic and clinical basis for two distinct subtypes of primary Sjögren's syndrome. Rheumatology. 2021;60(2):837-48.
3. Sandling JK, Pucholt P, Hultin Rosenberg L, Farias FHG, Kozyrev SV, Eloranta M-L, et al. Molecular pathways in patients with systemic lupus erythematosus revealed by gene-centred DNA sequencing. Annals of the Rheumatic Diseases. 2021;80(1):109-17.
4. Bianchi M, Kozyrev SV, Notarnicola A, Hultin Rosenberg L, Karlsson Å, Pucholt P, et al. Contribution of rare genetic variation to disease susceptibility in a large Scandinavian myositis cohort. Arthritis & Rheumatology. 2021.
5. Wang C, Zhan X, Bragg-Gresham J, Kang HM, Stambolian D, Chew EY, et al. Ancestry estimation and control of population stratification for sequence-based association studies. Nature Genetics. 2014;46(4):409-15.
6. Wang C, Zhan X, Liang L, Abecasis Gonçalo R, Lin X. Improved Ancestry Estimation for both Genotyping and Sequencing Data using Projection Procrustes Analysis and Genotype Imputation. The American Journal of Human Genetics. 2015;96(6):926-37.
7. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen W-M. Robust relationship inference in genome-wide association studies. Bioinformatics. 2010;26(22):2867-73.
8. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience. 2015;4(1):7.
9. Yu CY. The complete exon-intron structure of a human complement component C4A gene. DNA sequences, polymorphism, and linkage to the 21-hydroxylase gene. The Journal of Immunology. 1991;146(3):1057.
10. Yu CY, Belt KT, Giles CM, Campbell RD, Porter RR. Structural basis of the polymorphism of human complement components C4A and C4B: gene size, reactivity and antigenicity. The EMBO journal. 1986;5(11):2873-81.
11. Dangel AW, Mendoza AR, Menachery CD, Baker BJ, Daniel CM, Carroll MC, et al. The dichotomous size variation of human complement C4 genes is mediated by a novel family of endogenous retroviruses, which also establishes species-specific genomic patterns among Old World primates. Immunogenetics. 1994;40(6):425-36.
12. Wu Y, Lundstrom E, Liu C-C, Yang Y, Gunnarsson I, Svenungsson E, et al. Low copy-number of complement C4A, the presence of HLA-DR3, and the presence of HLA-DR2 are independent and additive risk factors for human systemic lupus erythematosus (SLE). The Journal of Immunology. 2010;184(1 Supplement):93.13.
13. Wu YL, Savelli SL, Yang Y, Zhou B, Rovin BH, Birmingham DJ, et al. Sensitive and Specific Real-Time Polymerase Chain Reaction Assays to Accurately Determine Copy Number Variations (CNVs) of Human Complement *C4A*, *C4B*, *C4*-Long, *C4*-Short, and RCCX Modules: Elucidation of *C4* CNVs in 50 Consanguineous Subjects with Defined HLA Genotypes. The Journal of Immunology. 2007;179(5):3012.
14. Sekar A, Bialas AR, de Rivera H, Davis A, Hammond TR, Kamitaki N, et al. Schizophrenia risk from complex variation of complement component 4. Nature. 2016;530(7589):177-83.

15. Almlöf JC, Nystedt S, Leonard D, Eloranta M-L, Grosso G, Sjöwall C, et al. Whole-genome sequencing identifies complex contributions to genetic risk by variants in genes causing monogenic systemic lupus erythematosus. Human Genetics. 2019.
16. Kamitaki N, Sekar A, Handsaker RE, de Rivera H, Tooley K, Morris DL, et al. Complement genes contribute sex-biased vulnerability in diverse disorders. Nature. 2020;582(7813):577-81.
17. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, et al. The ensembl variant effect predictor. Genome biology. 2016;17(1):1-14.
18. Xie C, Yeo ZX, Wong M, Piper J, Long T, Kirkness EF, et al. Fast and accurate HLA typing from short-read next-generation sequence data with xHLA. Proceedings of the National Academy of Sciences. 2017;114(30):8059.