Multimodal repertoire analysis unveils B cell biology in health and immune-mediated diseases

Mineto Ota[1,2*], Masahiro Nakano[1,3], Yasuo Nagafuchi[1,2], Satomi Kobayashi[1], Hiroaki Hatano[1,4], Ryochi Yoshida[1], Yuko Akutsu[1], Takahiro Itamiya[1], Atsushi Matsuo[5], Yumi Tsuchida[1], Hirofumi Shoda[1], Kazuhiko Yamamoto[1,3], Kazuyoshi Ishigaki[4], Tomohisa Okamura[2], Keishi Fujio[1*]

[1]Department of Allergy and Rheumatology, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan
[2]Department of Functional Genomics and Immunological Diseases, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan
[3]Laboratory for Autoimmune Diseases, RIKEN Center for Integrative Medical Sciences, Yokohama, Kanagawa, Japan
[4]Laboratory for Human Immunogenetics, RIKEN Center for Integrative Medical Sciences, Yokohama, Kanagawa, Japan
[5]Research Division, Chugai Pharmaceutical Co., Ltd., Kamakura, Kanagawa, Japan

* Corresponding authors. Email: otam@m.u-tokyo.ac.jp (M.O.), fujiok-int@h.u-tokyo.ac.jp (K.F.)

Abstract

Despite involvement of B cells in the pathogenesis of immune-mediated diseases, biological mechanisms underlying their function are scarcely understood. To overcome this gap, comprehensive analysis of the B cell repertoire is essential. Here, we cataloged and investigated the repertoire of five B cell subsets from 595 cases under immune-mediated diseases and health. CDR-H3 length among naïve B cells was shortened among autoimmune diseases in an interferon signature-dependent manner. VDJ gene usage was skewed especially in plasmablasts and unswitched-memory B cells of systemic lupus erythematosus patients with frequent usage of VDJ genes used mainly in naïve B cells and not unswitched-memory B cells of healthy controls. We developed a scoring system for this skewing, and it correlated with peripheral helper T cell transcriptomic signatures and disease activity and decreased after belimumab treatment. Moreover, genetic association analysis identified three molecules possibly involved in somatic hyper-mutation processes in humans. Our multimodal repertoire analysis brings new insights to B cell biology.

Main

B cells play critical roles in adaptive immunity, and their importance in immune-mediated disease (IMD) pathogenesis is suggested by the overlap of genetic risk variants of IMDs to regulatory regions in B cells[1, 2] and by the clinical efficacy of B cell-targeted therapies in IMDs[3, 4]. Thus, deeper understanding of B cell biology could clarify IMD pathology.

To increase their diversity, B cell receptors (BCRs) are created by intrinsic genome-editing. In the bone marrow, one V, D and J gene for one clonotype is selected (VDJ recombination), and random deletions and insertions occur in the junctional regions. In the germinal center (GC), class switch recombination and loading of somatic hypermutation (SHM) further increases BCR diversity. These processes generate a vast variety of BCRs in an individual, a collection known as the "BCR repertoire". Heavy-chain complementarity-determining region 3 (CDR-H3) is the most variable region and central to antigen specificity of BCRs.

For securing prompt defensive responses against diverse foreign pathogens, humans have evolved a distinct immune system compared to other organisms[5]. Genetic adaptation in human populations has resulted in strong immune responses against pathogens, but it was a trade-off against susceptibility to autoimmunity[5]. The B cell immune system in humans is complex and has been elusive. It is now clear that genetic determinants of individual difference among BCR repertoires as well as analysis of actual differences in BCR repertoires among IMDs are essential for deeper understanding of B cell biology.

IMDs consist of autoimmune and autoinflammatory diseases[6]. Autoimmune diseases are characterized by two prominent features: the emergence of autoantibodies and hyperactivity of interferons (IFNs)[6], although the relationship between these features remains elusive. For creation of memory B cells and antibody secreting cells, affinity maturation through GC is a critical step. On the other hand, previous studies revealed that those cell types can be differentiated in a GC-independent manner[7], a process termed "extrafollicular" maturation. Pathogenic autoantibodies are generated

through B cell extrafollicular reactions in systemic lupus erythematosus (SLE) and flares in SLE is characterized by the expansion of naive-derived activated effector B cells of extrafollicular phenotype[8, 9]. The disease relevance of these features of abnormal B cell maturation in IMDs could be better understood if placed in the context of a large-scale dataset combined with other features.

We recently created a gene expression and genome sequence database containing a large variety of immune cells from 10 IMD patients as well as healthy volunteers[2]. In this follow-up study, we focused on BCR features of 5 sorted B cell subsets that could be identified from RNA-seq data. We also examined their associations with transcriptomic, genetic, and clinical information. This largest-ever functional genomics dataset of immune cells from IMD patients enabled us to comprehensively explore BCR repertoire abnormalities in IMDs. We developed a scoring system of repertoire abnormalities in IMDs based on VDJ gene usage, the "Repertoire Naïveness (RN) Score". The RN score showed significant association with the transcriptomes of specific T cell subsets and disease activity. Also, genetic association analysis of >500 plasmablast samples suggested previously unknown biology in nucleotide substitution mechanisms in SHM. Our multi-omics approach to a patient-derived, large-scale dataset generated new insights into B cell biology in both health and disease states.

Results

Profiling of BCR repertoires of IMD patients and healthy volunteers

We mapped BCR regions from RNA-seq reads of 5 sorted $CD19^+$ B cell subsets: $CD27^-IgD^+$ naïve B cells, $CD27^+IgD^+$ unswitched memory B cells (USM B), $CD27^+IgD^-$ switched memory B cells (SM B), $CD27^-IgD^-$ double negative B cells (DN B) and $IgD^-CD27^{++}CD38^+$ plasmablasts (Fig. 1a). We extended our cohort from the previous report[2] and samples as follows: 136 patients with SLE, 90 patients with systemic sclerosis (SSc), 85 patients with idiopathic inflammatory myopathy (IIM), 147

patients with other IMDs and 137 healthy controls (HC), giving a total of 595 cases (Fig. 1a). For 22 SLE patients, samples were also obtained after a half-year treatment with belimumab, a biologic targeting B-cell activating factor (BAFF). After mapping and quality control (Methods), 5.8 million productive unique CDR-H3 clonotypes were identified from 2,893 samples. On average, 2,801 distinct productive clonotypes were identified in each sample (Extended data Fig. 1a). More than 5,000 BCR-derived reads were identified per million RNA-seq reads in each subset (median), and in plasmablasts it reached 1.1×10^5 reads (Fig. 1b), reflecting their specialized role in secreting antibodies.

As naïve B cells are expected to have few SHM, the mutation frequency can reflect the error rate in PCR or sequencing steps. The SHM frequency in V genes in naïve B cells was well controlled at 0.15% (median), in clear contrast to 6.7% (median) in plasmablasts (Fig. 1c). The number of shared CDR-H3 sequences was substantially larger in comparisons of the same individuals than in the comparisons of different individuals (Fig. 1d). The greatest degree of sharing was observed among the samples from the same individual and the same subset at different time points, indicating the valid reproducibility of our BCR repertoire dataset (Fig. 1d). Equivalent to the previous single cell BCR analysis that classified more than 99% of IgG+ and 98% of IgA+ cells as memory cells[10], IgG or IgA were rarely detected (0.4%) in naïve B cells in our dataset (Fig. 1e). Together, our dataset captured biologically reasonable characteristics of BCR repertoires and enabled deep phenotyping of repertoires among IMDs.

In order to evaluate the influence of the number of reads for BCR repertoire characterization, we down-sampled the number of reads from plasmablast RNA-seq data as well as the BCR-seq data of peripheral blood mononuclear cells from the previous study[11] and mapped BCR regions in the same way. These analyses revealed the strong influence of the number of reads on the number of identified CDR-H3 clonotypes, in clear contrast to stable quantification of CDR-H3 length, V gene usage and isotype frequency even with shallow depth (Extended data Fig. 1b). Thus, in our following RNA-seq based BCR repertoire analysis, we focused on these relative features, rather than analyzing whole

clonotypes.

In autoimmune diseases, CDR-H3 length is shortened in naïve B cells in an interferon signal strength-dependent manner

CDR-H3 loop length is one of the classic indices for the assessment of repertoire bias[12]. When we compared CDR-H3 length with HC, marked shortening of CDR-H3 lengths in naïve B cells (by more than 1 amino acid) and DN B cells among autoimmune diseases were observed (Fig. 2a; Extended data Fig. 2a). Conversely, CDR-H3 lengths in USM B cells and plasmablasts of SLE patients were longer than those in HC (Fig. 2a; Extended data Fig. 2a). Among the sections of the CDR-H3 region, lengthening in plasmablasts was observed in germline-coded regions but not in junction regions, in contrast to shortening in overall sections in naïve B cells (Fig. 2b; Extended data Fig. 2b).

Observations so far were only from productive sequences. Non-productive sequences, which refer to the sequences with a frameshift or stop codon, can be used as a means of studying the pre-selection repertoire[13]. In our dataset, shortening of CDR-H3 lengths among autoimmune diseases was observed in non-productive sequences of naïve and memory B cells (Extended data Fig. 2c). Thus, together with the observed shortening of overall CDR-H3 sections in productive sequences, short CDR-H3 clonotypes may be generated before selection of mature naïve B cells (e.g., during the process of VDJ recombination and nucleotide insertion/deletion at the junctions). This observation is analogous to previous reports of shortening of CDR3 length in pre-selection T cell receptors among type 1 diabetes patients[13]. Moreover, CDR-H3 lengthening in germline regions of plasmablasts and USM B in SLE may be the consequence of skewed VDJ gene usage as described in the next section.

CDR-H3 shortening was especially prominent in naïve B cells from diseases with high IFN signatures (Fig. 2a; Extended data Fig. 2d), although CDR3 length was not as divergent among light chains (Extended data Fig. 2e). To investigate further, we tested the association of gene expression and

CDR-H3 length in naïve B cells in each disease and subsequently performed meta-analysis in order to avoid the confounding of disease effects (Methods). Genes associated with a short CDR-H3 length were significantly enriched in IFN signature genes (ISG, Fig. 2c, d), and the association was observed in each autoimmune disease (Extended data Fig. 2f). CDR-H3 length in naïve B cells was also significantly associated with ISG expression in 4 other B cell subsets, indicating the association of generalized IFN activity itself with CDR-H3 length (Extended data Fig. 2g). Strong negative correlation between CDR-H3 length and IFN activity was characteristic of naïve B cells among 5 B cell subsets (Fig. 2e). These results suggested a role of IFN signaling at the early stage of B cell development, which is consistent with previous findings of IFN-induced abnormal B cell maturation in bone marrow of SLE[14].

Skewed gene usage in autoimmune diseases suggested an abnormal repertoire maturation process

It was previously suggested that there was differential usage of VDJ genes according to B cell subsets[15, 16]. As these reports were limited to one healthy female[15] and inbred strain of mice, C57BL/6[16], the generalizability of this finding in a large population has been elusive. In our cohort, gene usage preference among 5 B cell subsets was also observed (Extended data Fig. 3a). More surprisingly, by principal components analysis (PCA) of VDJ gene usage, samples were clearly separated according to cell types (Fig. 3a). The separation pattern reflected the lineage relationships among the subsets, and they were arranged similarly to the flow-cytometric plots according to CD27 positivity and IgD positivity (Fig. 3a). This observation is analogous to previously reported clear separation of 3 B cell subsets by PCA of VDJ use in C57BL/6 mice[16]. Those results strongly suggested that B cell selection was based on VDJ gene segments independent of the sequence of their antigen-binding regions in humans.

It has been shown that there is skewed usage of the IGHV in IMDs[17] with dependence on the stage of disease[18], although the sample size has been limited. Notably, in the PCA space, the gene usage signature was drastically skewed in the negative direction of the CD27 axis in SLE among USM B and plasmablasts (Fig. 3a; Extended data Fig. 3b). With regard to IIM and SSc, the repertoire skewed in the same direction in the PCA space in USM B, but not obviously in plasmablasts (Extended data Fig. 3b). In case-control comparisons, many differentially used genes were identified especially in autoimmune diseases (Fig. 3b). In parallel with the findings in the PCA space, VDJ gene usage differences between SLE and HC in USM B, as well as plasmablasts, showed clear agreement with gene usage differences between naïve B and USM B in HC (Fig. 3c). These results indicated that the majority of observed differential gene usage in SLE was associated with the global abnormality in B cell maturation process, rather than the consequence of specific antigens.

Based on the observed skewing in VDJ gene usage in autoimmune diseases toward a CD27 negative, "naïve like" repertoire, we next developed a score to quantify such skewing based on its VDJ gene usage. We termed this factor the "Repertoire Naïveness (RN) Score". Briefly, the RN score is calculated as the sum of the product of normalized VDJ gene usage and the size of its effect, defined by naïve B versus USM B comparisons in HC (Methods, Supplementary table 1). As we defined the effect sizes of each gene only using HC samples, this score can be applied to inter-disease comparison without bias. Consistent with the appearance in PCA, our RN score was the highest among naïve B cells followed by DN B cells, and the highest among SLE patients in USM B or plasmablasts (Extended data Fig. 3c, d). To validate our approach, we adopted the same scoring system to the BCR-seq data derived from sorted B cells from human tonsils in a previous study[19]. The RN score was clearly high in naïve B cells and low in memory B cells (Extended data Fig. 3e).

As reported previously[20], the use of each VDJ gene was associated with different CDR-H3 lengths (Extended data Fig 3f). The genes associated with an extended CDR-H3 were preferentially used by

plasmablasts in SLE (Extended data Fig. 3g). Thus, the extended CDR-H3 length in SLE in this subset is associated with abnormal VDJ gene usage, at least in part.

The repertoire abnormality in autoimmune diseases may be associated with the extrafollicular pathway and specific subpopulations of Th1 cell

Extrafollicular maturation of plasmablasts has been reported as a hallmark of SLE[8]. Next, we assessed the association of the observed "naïve like" repertoire and the extrafollicular maturation. Although extrafollicular maturation-derived antibody secreting cells undergo SHMs, they are expected to have less SHMs[7, 9, 21] compared to GC-derived cells. In line with this notion, the frequency of SHMs in plasmablasts was significantly lower in SLE patients than in HC (Extended data Fig. 4a). Furthermore, it showed a significant negative correlation with the RN score (Fig. 4a). To examine the relationship in clonotype levels, we divided BCR clonotypes into bins according to the frequencies of SHMs and calculated RN scores in each bin. In this comparison, we observed that clonotypes with a lower number of mutations showed higher RN scores (Fig. 4b). Taken together, naïve-like gene usage in plasmablasts is associated with a low load of SHM. Reflecting the accumulation of mutations with aging[11], SHM frequencies in plasmablasts showed a significant positive correlation with aging ($p < 2 \times 10^{-16}$), although the RN score was not affected by aging ($p > 0.05$, Extended data Fig. 4b). The RN score distinguished SLE patients from HC more accurately than did the SHM frequency (Extended data Fig. 4c). Thus, although well correlated, the RN score might reflect disease-associated pathways better than the SHM frequency.

We next compared the RN scores of 136 SLE patients focusing on the gene expression of 27 immune cell subsets that were obtained from the same patients. The number of genes that showed a significant correlation with the RN score was the largest in Th1 cells for both plasmablasts and USM B (Fig. 4c). The top associated genes in Th1 cells included *PDCD1*, *TOX* and *ICOS*, which are hallmarks of

recently identified peripheral helper T cells (Tph)[22] (Fig. 4d). Indeed, the Th1 genes associated with RN score showed significant concordance with the direction of change in the differential expression analysis between PD-1^{hi} versus PD-1^{-} T cells from the original report[22] (Fig. 4e). Thus, the Th1 genes associated with the RN score may reflect the expansion of Tph-like subpopulation among Th1 cells. The Tph gene expression score in Th1 cells (Tph1 score, Methods) was high in SLE (Extended data Fig. 4d), consistent with the reported expansion of Tph population[23], particularly with Th1 phenotype[24], in SLE. The Tph1 score showed a strong correlation with the RN score in SLE, both in USM B and plasmablasts (Extended data Fig. 4e). Correlation of these two scores was also prominent in USM B in IIM, although the correlation in plasmablasts was lower than that in SLE. These results may reflect the shared and distinct pathogenicities among autoimmune diseases. The association of Tph1 score and RN score was also validated in our independent SLE cohort[25] (Extended data Fig. 4f). Although IFN signal strength was also correlated with RN scores (Extended data Fig. 4g), this association was mostly mediated by the Tph1 score (Fig. 4f). The direct effect of IFN signal strength on RN score was not significant (p = 0.21) in the mediation analysis (Fig. 4f). This result indicated the role of IFN on Tph cell expansion. The proposed function of Tph cells is that they provide help to B cells outside the GC[22, 23]. Collectively, associations of RN score with low SHM frequencies and high Tph1 score support the notion that the naïve-like repertoire in IMDs is a consequence of excess extrafollicular maturation.

The RN score of USM B correlated significantly and positively with the expression of *TBX21* in USM B, which codes for T-bet (Extended data Fig .4h). As T-bet^{+} B cells reportedly have distinct functions[26] and are associated with autoimmunity[27], this naïvenesss-associated USM B population might play a role in autoimmune pathogenesis.

Repertoire abnormalities are associated with disease activity in SLE

We next tested the association of BCR repertoire abnormalities and disease activity in SLE. RN scores in plasmablasts and USM B showed significant positive correlations with SLEDAI-2K, a measure of SLE disease activity (Fig. 5a), and these associations were still significant after adjusting for Tph1 scores ($p = 1.3 \times 10^{-3}$, 1.7×10^{-3}) and ISG scores ($p = 5.3 \times 10^{-4}$, 2.3×10^{-5}), indicating the clinical relevance of RN scores. In addition, RN score in plasmablasts significantly correlated with total IgG concentrations in sera (Extended data Fig. 5a).

We next asked whether specific V, D or J gene usage was associated with disease activity. In SLE, among the 5 B cell subsets, IGHV4-34 usage in USM B showed by far the most significant positive association with SLEDAI-2K ($p = 3.9 \times 10^{-10}$, Fig. 5b, Extended data Fig. 5b). Although the expansion of antibody secreting cells with IGHV4-34 in acute SLE has been reported[9], the clinical relevance of its usage in USM B has not been clarified. IGHV4-34 is a V gene that is predominantly used in naïve B cells (Extended data Fig. 3a) and its usage explained 28% of the variance of RN score in USM B (Extended data Fig. 5c). Mediation analysis indicated that over one-half of the association between RN scores in USM B and SLEDAI-2K was mediated by IGHV4-34 usage. This contrasted with the independent direct association of RN scores in plasmablasts with SLEDAI-2K (Extended data Fig. 5d). Thus, repertoire naïveness in SLE may be associated with disease activity partly via high usage of IGHV4-34 in USM B.

It has been reported that IGHV4-34 has AVY and NHS motifs that are associated with autoimmunity[28]. Mutations in these motifs abrogate its autoreactivity and only such clones without autoreactivity are selected for maturation under healthy conditions[28, 29]. IgG B cells without mutations in these motifs from IRAK4- and MYD88-deficient patients demonstrated reactivity to autoantigens as well as binding to commensal bacterial antigens[29]. In our dataset, NHS and AVY motifs were significantly more conserved without mutations in SLE USM B compared to HC as well as in

247 plasmablasts (Fig. 5c). Thus, IGHV4-34 clonotypes in SLE might feature binding to autoantigens.

248

249 **Belimumab attenuated abnormal repertoire in SLE**

250 For SLE, the BAFF inhibitor belimumab is one of the two biologics that has been approved by the
251 United States Food and Drug Administration[3]. In SLE, BAFF is induced by type I IFN stimulation and
252 supports the survival and differentiation of autoreactive B cells[30, 31]. To assess the effect of BCR
253 modifications, we tested the BCR repertoire before and after treatment with belimumab in 22 cases of
254 SLE. Belimumab effectively reduced SLEDAI-2K and increased complement levels in our cohort (Fig.
255 5d). In order to assess the overall trend, we projected cases before and after belimumab treatment on
256 the same PCA space as Fig. 3a. Belimumab treatment significantly changed the repertoire of USM B
257 against the naiveness-associated direction, but not in the other subsets (Fig. 5e, f). Belimumab
258 significantly reduced the RN scores in USM B, although we detected no effects on Tph1 scores or IFN
259 scores (Fig. 5d). As to IGHV4-34, belimumab significantly reduced the fraction of nonmutated
260 AVY/NHS motifs in USM B and plasmablasts (Extended data Fig. 5e). Thus, together with its major
261 effect on the USM B repertoire, belimumab also reduced the autoimmune-associated fraction of BCRs
262 in plasmablasts.

263 We next divided all cases in our cohort, including healthy controls, into 6 clusters based solely on
264 the usage of VDJ genes of USM B and plasmablasts that were utilized for RN score calculation
265 (Extended data Fig. 5f). The largest cluster (cluster 1) mostly consisted of autoimmune disease
266 patients; these subjects showed higher RN scores, Tph1 scores and IFN scores than the others. Among
267 SLE patients, subjects in this group responded better to belimumab than did others when assessed by
268 delta SLEDAI-2K after a half-year treatment (p = 0.007, Extended data Fig. 5g). These results
269 indicated that the BCR repertoire abnormality reflected clinical heterogeneity within and between
270 diseases.

Altered isotype usage in SLE is partly influenced by abnormal repertoire maturation

With regard to isotype frequency, significant skewing was also observed in IMDs compared with HC (Extended data Fig. 5h). As the frequencies of many isotypes were altered in SLE plasmablasts, we assessed their association with RN scores in plasmablasts. The association between isotype frequencies and RN scores differed according to class (Extended data Fig. 5i). A substantial proportion of the decrease in IgA2 frequency and the increase in IgG1 frequency in SLE could be explained by the high RN score in SLE. That finding might indicate an association with the extrafollicular pathway, although the increase in IgA1 frequency in SLE seemed to have no association with this pathway (Extended data Fig. 5i). Thus, the increase in IgA1 frequency in SLE might be due to other pathways, such as involvement of mucosal immunity as suggested previously[17, 29].

Analysis of genetic association reveals molecules associated with somatic hyper-mutation biology

Finally, we tested the association of BCR repertoire features with genetic polymorphisms. For VDJ gene usage, we observed significant cis-regulatory effects (Fig. 6a). This result is concordant with a classical twin study that showed significant heritability in V gene usage[32]. Significant genetic association was observed only in V gene usage, and not in D and J gene usage (Extended data Fig. 6a). The genetic association with V gene usage was highly concordant among 5 B cell subsets (Extended data Fig. 6b), indicating that polymorphisms have effects on gene preference at the step of VDJ recombination.

SHM is likely a random process. However, the unevenness of nucleotide substitution patterns indicates complex biological processes in the substitution process[33]. To assess the association of SHM substitution patterns and genetic polymorphisms, we developed a new framework, i.e., SHM-QTL analysis. In this analysis, we quantified nucleotide substitution pattern frequencies in plasmablasts of

each subject and assessed its genetic association (Methods). In this way, we identified two loci that showed significant genome-wide association with SHM patterns (Fig. 6b). Of note, in the *NUGGC* locus, 3 conditionally independent lead variants were identified, and they were associated differently with nucleotide substitution patterns (Fig. 6c, Extended data Fig. 6c). Two of the 3 lead variants were missense variants and the other one was also in tight LD ($r^2 = 0.96$) with a missense variant. These missense variants were located in different exons of *NUGGC*, and 2 of the 3 variants were estimated to be located on the surface of the protein (Extended data Fig. 6d). Thus, these variants could change the interaction of *NUGGC* with other molecules. In the *RRM2B* locus, the lead variant was in high LD with a missense variant of *RRM2B* (Fig. 6d, rs1037699, $r^2 = 0.87$), and conditioning by the missense variant cancelled the association of the lead variant (Fig. 6e), thus indicating the association of this missense variant. In the *RRM2B* locus, conditioning analysis revealed the other associated variant, rs892503. Although this variant was not in LD with any missense variant, the residual association showed nearly perfect co-localization with eQTL effects in plasmablasts for *AP002852.2* (Fig. 6e, coloc[34] *PP. H4* = 0.985), long non-coding RNA (lncRNA) expressed almost exclusively in memory B cells and plasmablasts (Extended data Fig. 6e). The identified 5 independent variants contributed to substitution patterns differently (Extended data Fig. 6f), and they explained 21.1% of SHM substitution diversity in total (Methods). To our knowledge, this is the first assessment of genetic effects on SHM substitution patterns and highlighted three molecules that may be associated with SHM biology in humans.

Discussion

In this report, we characterized BCR repertoire abnormalities in IMDs using a large cohort of 595 cases. Our multi-layer dataset enabled us to link the BCR repertoire to other layer information, including transcriptome of multiple immune cell types, genetic polymorphisms and clinical

information.

We identified marked shortening of CDR-H3 length in naïve B cells of autoimmune disease patients in an IFN signal strength-dependent manner, as well as lengthening in plasmablasts of SLE patients. Shortening of BCR CDR-H3 length in peripheral blood mononuclear cells of SLE patients had been reported[35]. Antibodies with long CDR-H3 length were associated with autoreactivity [36] and long CDR-H3 length in class switched clonotypes in SLE was also reported[17]. Our results are consistent with all of these seemingly contradicting reports and reinforce the importance of using sorted B cells for repertoire analysis[15].

In gene usage analysis, we observed a global pattern of skewness in specific subsets of IMDs as quantified with the RN score. This score can measure extrafollicular maturation based on the observation of its association with naïve-like gene usage, low SHM load and its correlation with Tph signature gene expression in Th1 cells. Another possible explanation includes the defects in GC selection by formation of spontaneous GCs as suggested in lupus model mice and patients[37, 38], which might result in a similar BCR phenotype. Although the elucidation of the precise intermediate process is a future problem, the observed significant association of these scores with clinical parameters and their responses to clinically effective treatment suggested the disease relevance of these BCR repertoire-based measures and their possible utility for clinical practice such as patient stratification. Interestingly, in SLE patients, the repertoires of USM B and plasmablasts seemed to be skewed similarly in response to Tph expansion. This was in contrast to the observation of similar associations in USM B but less so in plasmablasts in IIM. This result may suggest the existence of a “gate-keeper” for plasmablast maturation that is defective in SLE. Together with CDR-H3 length analysis, our results support the notion that the break of tolerance occurs both in central and peripheral checkpoints in SLE[39, 40]. Also, our analysis supported direct and indirect associations of IFN signaling with both of these checkpoints.

In SHM-QTL analysis, we identified 3 molecules possibly associated with SHM biology. SHM is triggered by activation-induced cytidine deaminase (AID)[41]. Although multiple molecules are involved in subsequent DNA damage response[42], the roles of 3 identified molecules in this study has not been well known. *NUGGC*, Nuclear GTPase, Germinal Center Associated, was first identified as a GTPase expressed in germinal center B cells. It inhibits the function of AID in cell lines[43]. Interestingly, NUGGC-deficient mice reportedly have a substantial increase in mutations at G:C base pairs[44] and experience altered patterns of SHM. These data, together with our findings, strongly suggest an association of this molecule in DNA repair in SHM processes in humans. *RRM2B* has been reported to repair DNA damage in a TP53-dependent manner[45], although its function in B cells has not been known. RRM2B may also be associated with repair of DNA damage induced by AID.

In summary, our large-scale BCR repertoire analysis provided an overview of B cell abnormalities in IMDs and revealed the presence of new molecules associated with SHM biology in health. This analysis exemplifies the usefulness of multi-omic human data for finding new features in cellular biology.

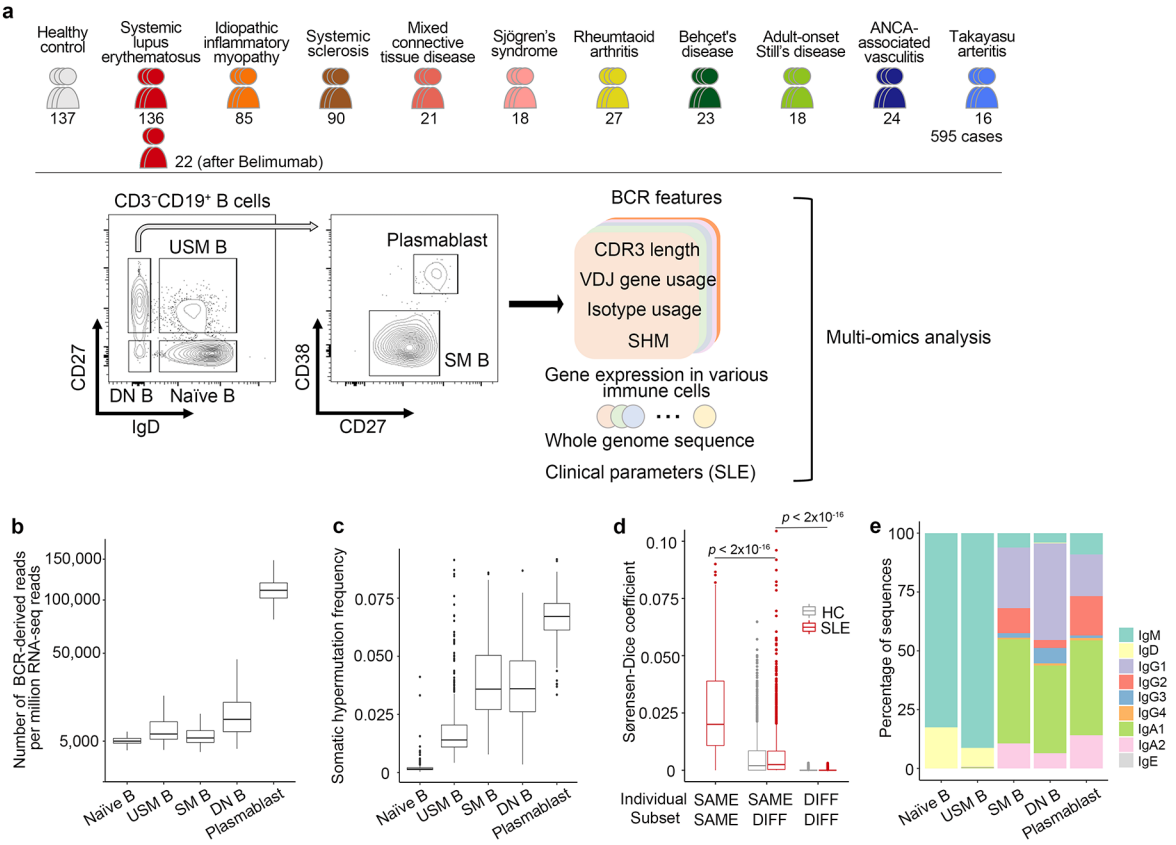


Fig. 1: **Study design and quality features of our dataset.**

a, Study design used in this report. Sample numbers after quality control are presented under each disease. **b,** Number of RNA-seq reads aligned to BCR regions per million reads in 5 B cell subsets. **c,** Comparison of the per base mutation frequency in the V gene region of 5 B cell subsets. **d,** Comparison of clonotype overlap frequencies between samples according to the individual or the subset being the same or different. "The same individual, the same subset" comparison was performed with SLE samples before and after treatment with belimumab 6 months apart. *P* values were calculated with two-sided Mann–Whitney *U*-test. **e,** Comparison of median isotype frequencies in each subset among healthy controls.

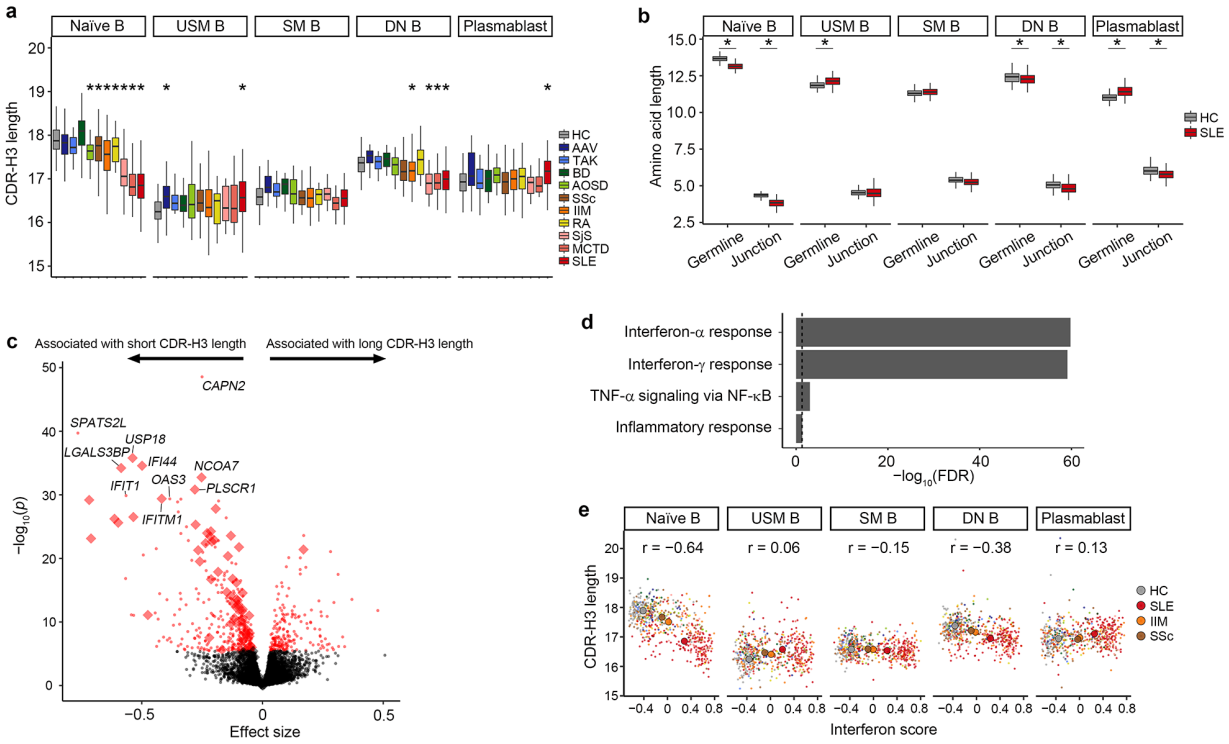


Fig. 2: In autoimmune diseases, CDR-H3 length is shortened in naïve B cells in an interferon signal strength-dependent manner.

a, Comparison of CDR-H3 length between IMDs and HC. The difference between HC and each disease was tested with linear regression, adjusting for age and batch effects. *, $p < 0.05/50$. **b,** Comparison of germline and junction part length of CDR-H3 region (see also Extended data Fig. 2b). The differences between HC and SLE were tested with linear regression, adjusting for age and batch effects. *, $p < 0.05/10$. **c,** Association of naïve B cell gene expression with naïve B cell CDR-H3 length. Fixed effect meta-analysis p values and effect sizes are plotted (Methods). Genes that attained Bonferroni-adjusted $p < 0.05$ are marked red. Genes in "HALLMARK_INTERFERON_ALPHA_RESPONSE" gene set from MSigDB are marked with diamonds. **d,** Gene set enrichment analysis. Results of genes significantly negatively associated with CDR-H3 length in naïve B cells. FDR, false discovery rate. **e,** Correlation between IFN score and CDR-H3 length in 5 B cell subsets. Mean values among HC, SLE, SSc, and IIM are marked with large points. The color of each point indicates the clinical diagnosis as illustrated in Fig. 1a. Pearson's r is provided.

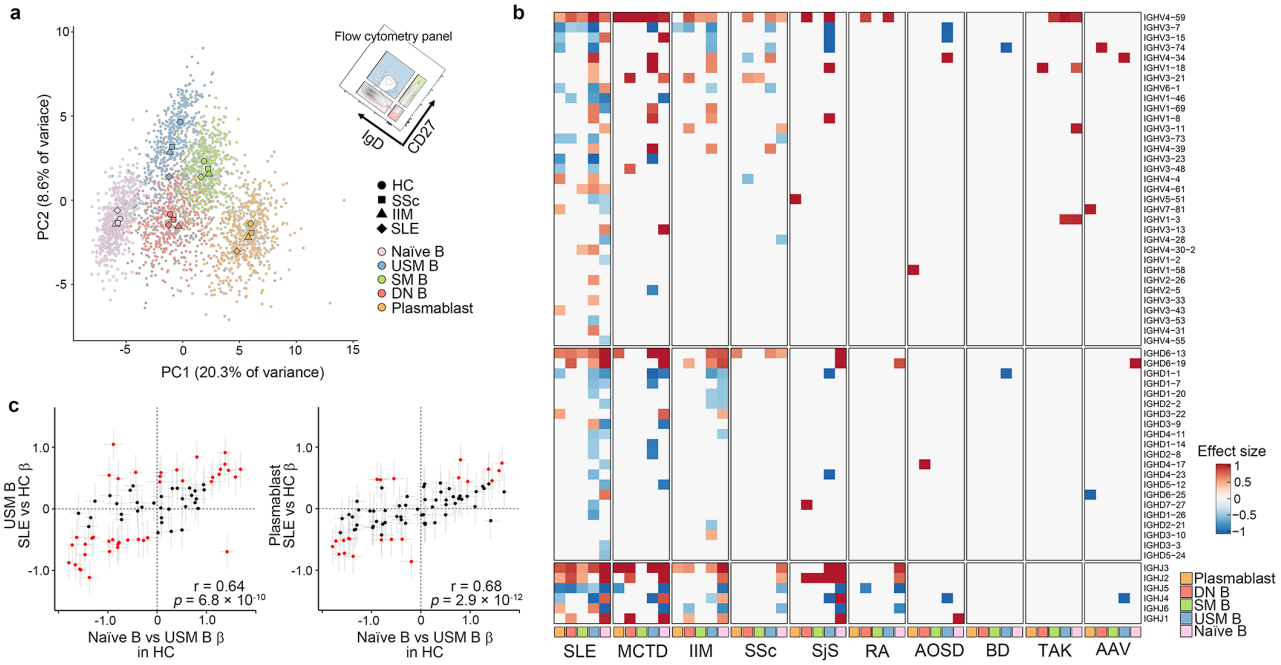


Fig. 3: Skewed gene usage in autoimmune diseases suggested an abnormal repertoire maturation process.

a, Principal component analysis plot of all samples based on VDJ gene usage. Centroids of HC, SSc, IIM, and SLE in each subset are marked with large points. **b,** Skewed gene usage in IMDs compared to HC. In each disease-subset (column), gene usage (row) was compared with HC after adjusting for age and batch effects. Effect sizes of linear model comparison are plotted if the difference was statistically significant (FDR < 0.05). Only genes that attained FDR < 0.05 in at least one comparison are shown. **c,** Comparison of β estimates in usage comparison between naïve B vs USM B in HC (y axis) and SLE vs HC (x axis) in USM B (left) or plasmablasts (right). Each plot corresponds to each VDJ gene. Genes that attained FDR < 0.05 in SLE vs HC comparison are marked red**.** Grey bars indicate 95%CI. Pearson's r and its significance are provided.

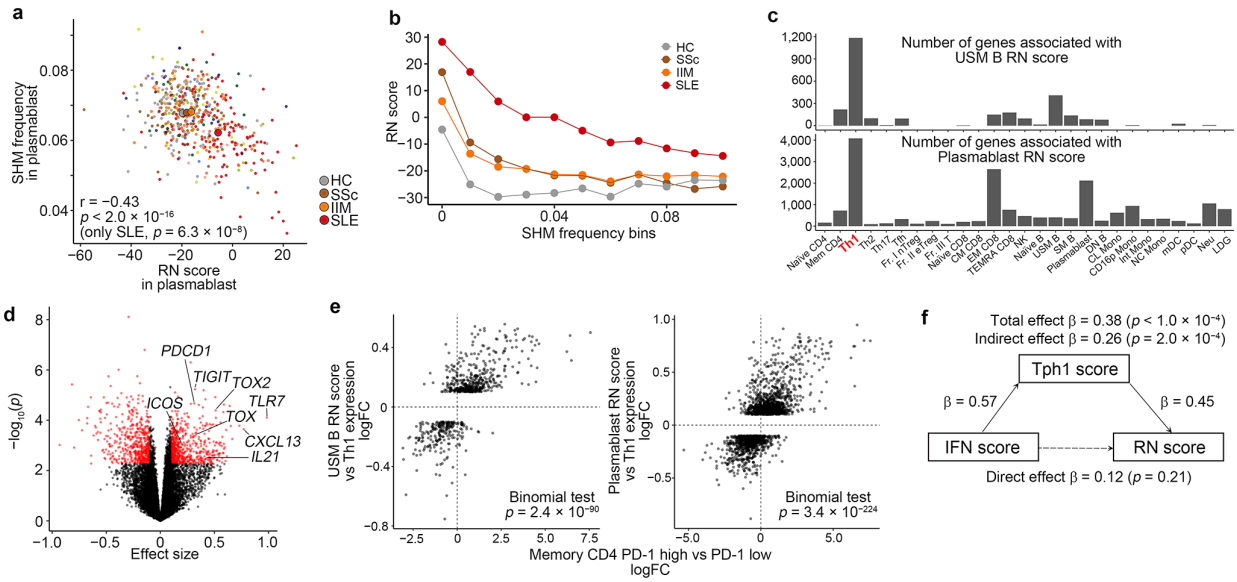


Fig. 4: Association of the repertoire abnormality in autoimmune diseases with a low load of mutations and specific subpopulations in Th1 cells.

a, Correlation of RN score and SHM frequency in the V gene region of plasmablasts. *P* values were calculated with linear regression adjusting for diagnosis, age and batch effects. Mean values among HC, SLE, SSc, and IIM are marked with large points. The color of each point indicates the clinical diagnosis as illustrated in Fig. 1a. **b,** RN scores of clonotypes stratified by SHM frequency. In each disease, clonotypes for which ≥ 200 nucleotides in the V gene were sequenced were aggregated and divided into bins according to mutation frequency in the V gene region. In each bin, RN score was calculated based on VDJ gene usage. **c,** Number of significantly associated genes with RN score in USM B (top) and plasmablasts (bottom) among expressed genes in each immune cell. Full description of subset names is in Table S3. **d,** Association of gene expression in Th1 cells and RN score in USM B. Significantly associated genes are colored red. **e,** Comparison of the effect size estimates of Th1 expression versus RN score comparison in our cohort (Y-axis) and PD-1^{hi} versus PD-1^{-} T cells comparison in the data by Rao, *et al*[22] (X-axis). Only genes significantly associated with RN scores in our cohort were included for the analysis. *P* values were calculated by the binomial test for concordance of the direction of effect size estimates. logFC, log fold change. **f,** Mediation model representing the relationships among IFN score, Tph1 score and RN score in plasmablasts among SLE patients. The bootstrapped indirect effect was 0.26 (95% CI, 0.14-0.39; $p = 2 \times 10^{-4}$). Thus, the indirect effect was statistically significant.

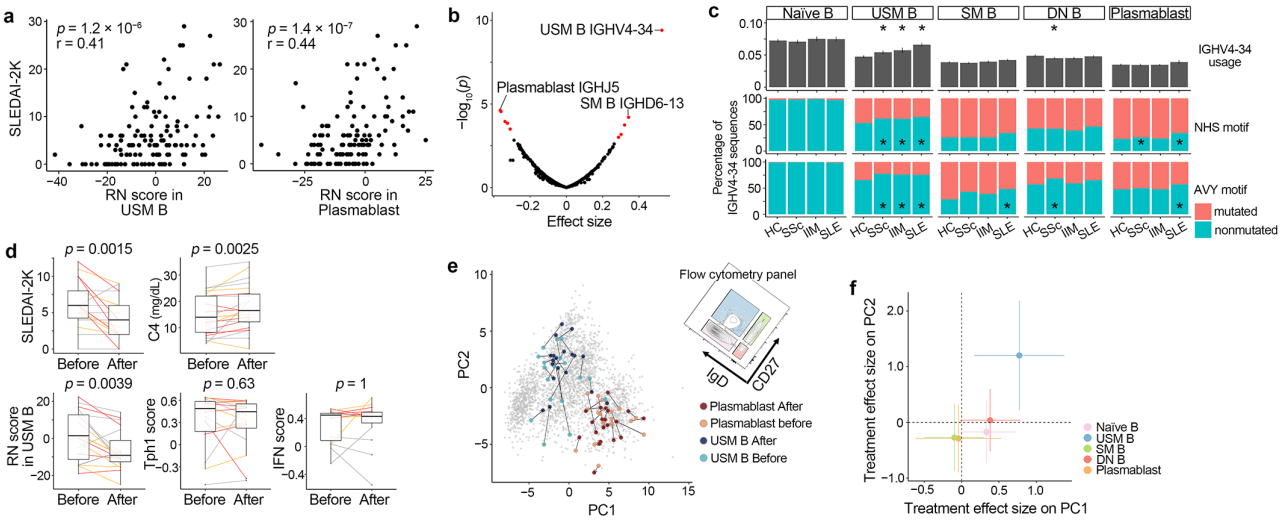


Fig. 5: Repertoire abnormalities are associated with disease activity in SLE and attenuated by belimumab treatment.

a, Correlation between RN scores and SLEDAI-2K among SLE patients. Pearson's r and its significance are provided. **b,** Association of VDJ gene usages among 5 B cell subsets and SLEDAI-2K. Significantly associated genes (FDR < 0.05) are red. **c,** Comparison of mean IGHV4-34 usage (top) and percentage of clonotypes with mutations in the NHS motif (middle) or the AVY motif (bottom) among IGHV4-34 clonotypes. For usage comparison, the differences between HC and each disease were tested with linear regression adjusting for age and batch effects. Bars indicate 2SE. For mutation comparison, Fisher's exact test comparing the number of mutated/nonmutated clonotypes between HC and each disease was performed in each subset. *, $p < 0.05/15$. **d,** Comparison of clinical parameters and scores before and after treatment with belimumab. Patients are colored according to improvement in SLEDAI-2K after treatment (red, ≥ 4; orange, 1-3; grey, ≤ 0). Paired Wilcoxon signed-rank test. **e,** Change in BCR repertoire after belimumab treatment. Paired samples before and after treatment were connected on the same PCA space as Fig. 3a. **f,** Belimumab treatment effects on PC scores by fitting the linear mixed model. Bars indicate 95% CI.

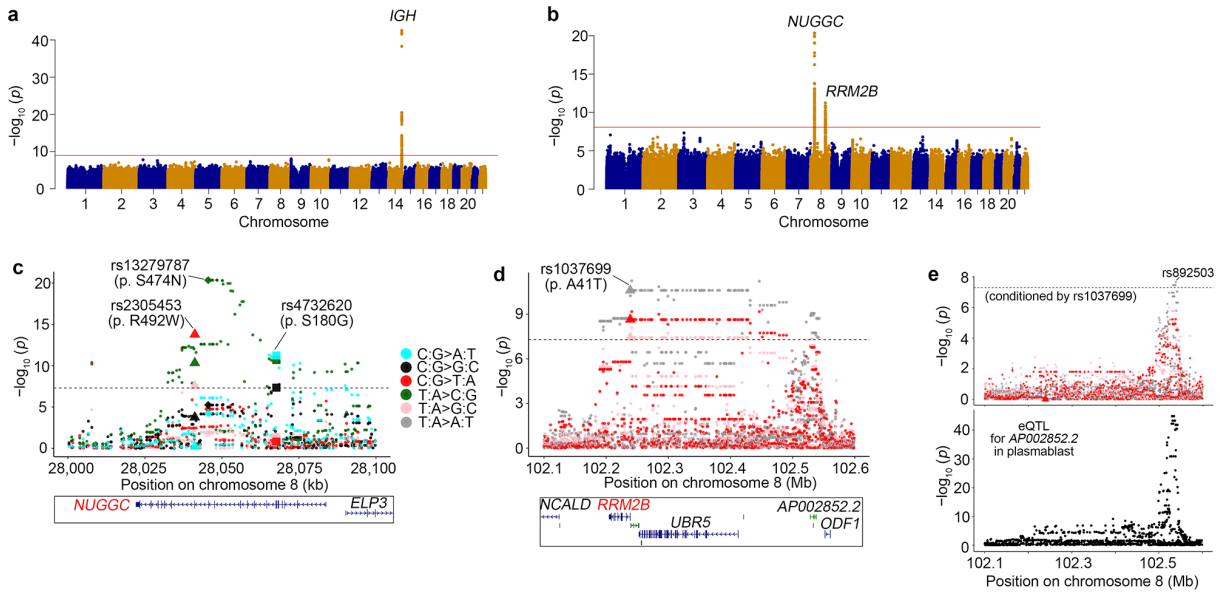


Fig. 6: Analysis of genetic association reveals molecules associated with somatic hyper-mutation biology.

a, *P* values of V gene usage-QTL analysis in naïve B cells. *P* values of QTL analysis for 43 V genes are plotted together. Red line indicates the significance threshold ($p < 5.0 \times 10^{-8}/43$). **b,** *P* values of SHM-QTL analysis. *P* values for QTL analysis for 6 patterns of mutations are plotted together. Red line indicates the significance threshold ($p < 5.0 \times 10^{-8}/6$). **c,** Regional association plot of *NUGGC* locus. Associations of 5 mutation patterns that attained genome-wide significance in this region are plotted. Large marks correspond to 3 independent missense variants that are also (or in tight LD with) lead variants. **d,** Regional association plot of *RRM2B* locus. Associations of 3 mutation patterns that attained genome-wide significance in this region are plotted. A large mark corresponds to the missense variant for *RRM2B* that is in high LD with the lead variant. **e,** Residual association in *RRM2B* locus after conditioning by rs1037699 (top) and eQTL *p* values for *AP002852.2* expression in plasmablasts (bottom).

Methods

Criteria for patient inclusion/exclusion

The SLE cohort was comprised of patients who had attended the Department of Allergy and Rheumatology at the University of Tokyo Hospital, Division of Rheumatic Diseases at National Center for Global Health and Medicine or Immuno-Rheumatology Center at St. Luke's International Hospital, who met the 1997 revised version of ACR SLE criteria[46]. Of the enrolled 136 patients, 88 cases had active disease with SLEDAI-2K ≥ 4. The other cases were inactive at the time of blood withdrawal. Patients administered cyclophosphamide or rituximab within one year or taking ≥ 21 mg/day prednisolone were excluded from this analysis. For 22 patients, belimumab was initiated after the collection of the samples, and samples after a half-year treatment were also collected. To assess the drug's effect, cases with concomitant prednisolone increase ≥ 5 mg/day or concomitant addition of other immunosuppressants were excluded.

The SSc cohort was comprised of patients who had attended the Department of Allergy and Rheumatology at the University of Tokyo Hospital or Department of Rheumatology at Tokyo Metropolitan Komagome Hospital, who met the 2013 American College of Rheumatology (ACR)/ European League Against Rheumatism (EULAR) classification criteria for Systemic Sclerosis[47]. Patients taking ≥ 21 mg/day of prednisolone were excluded from this analysis.

The IIM cohort was comprised of patients who had attended to the Department of Allergy and Rheumatology at the University of Tokyo Hospital or Division of Rheumatology at the Jikei University Hospital. They met standards defined by the Bohan and Peter criteria[48, 49], or the European Neuromuscular Center criteria[50], or the Sontheimer criteria[51] or the Griggs criteria[52]. Patients taking ≥ 12 mg/day prednisolone were excluded. Of the 85 patients, 36 cases had active disease and required subsequent initiation or an increase of immunomodulatory drugs.

The inclusion criteria for HC were people with no apparent co-morbidities, no direct family history

of autoimmune diseases and no use of prescription drugs or supplements. Age and sex were matched with the patient cohort as much as possible. For 30 HC cases, only naïve B cells among 5 B cell subsets were collected and utilized for the analysis. For the remaining 107 cases, 5 B cell subsets were collected.

Patient inclusion/exclusion criteria for the other diseases were described in detail in our first report[2].

Samples were collected from the University of Tokyo, the Jikei University School of Medicine, St. Luke's International Hospital, National Center for Global Health, and Medicine and Tokyo Metropolitan Komagome Hospital with approval by the Ethics Committees at each site.

RNA-seq experiments

We followed the same protocol that we reported previously[2]. Briefly, peripheral blood mononuclear cells (PBMCs) were isolated by density gradient separation with Ficoll-Paque (GE Healthcare) immediately after blood drawing. After lysis of erythrocytes, up to 26 immune cell subsets were collected with purity > 99% using a MoFlo XDP instrument (Beckman Coulter) or a 14-colour cell sorter BD FACSAria Fusion (BD Biosciences). We intended to collect a minimum of 1,000 cells (up to 5,000 cells) per subset. Libraries for RNA-seq were prepared using SMART-seq v4 Ultra Low Input RNA Kits (Takara Bio). Samples failing any of the quality control steps (RNA quality and quantity, PCR amplification, library fragment size) were eliminated from further downstream steps. Libraries were sequenced on the HiSeq 2500 Illumina platform to obtain 100-bp paired-end reads with HiSeq SBS Kit v4 (Illumina).

Mapping of the transcriptome

We followed the same protocol we reported previously[2]. Briefly, adaptor sequences were trimmed using cutadapt (v1.16) and 3′- ends with low-quality bases (Phred quality score < 20) were trimmed

using the fastx-toolkit (v0.0.14). Reads containing more than 20% low-quality bases were removed. Subsequently, reads were aligned against the GRCh38 reference sequence using STAR (v2.5.3)[53] in two-pass mode with GENCODE, version 27 annotations[54]. A median of 10.1 million paired fragments were uniquely mapped per sample and utilized for the subsequent analysis. Expression was quantified using HTSeq (v 0.11.2)[55].

Mapping and QC of BCR sequence

We utilized MiXCR software (version 3.0.13)[56] for BCR mapping. With FASTQ files after adapter and quality trimming, we utilized the "analyze shotgun" option with parameters "-s hsa", "--starting-material rna" and "--receptor-type bcr". After alignment and assembly of clonotypes based on the CDR3 sequence, the full BCR sequence was assembled with the "--contig-assembly" parameter. We set "specificMutationProbability" parameter to 2×10^{-3}. Only productive heavy-chain sequences were kept for the analysis unless specified. For all analyses, samples with more than 500 unique CDR-H3 sequences were used for the analysis. For each sample, we treated each unique BCR sequence as a single event, regardless of read depth. For mapping non-productive sequences, we raised the "specificMutationProbability" parameter to 8×10^{-3} in order to reduce the amount of productive sequences with error or mutations counted as non-productive sequences.

Mapping of whole genomic sequences

We followed the same protocol that we reported previously[2]. Genomic DNA was isolated from peripheral blood using QIAmp DNA Blood Midi kit (Qiagen) and libraries for whole genome sequencing were prepared using TruSeq DNA PCR-Free Library prep kit (Illumina). Whole genomes were sequenced on Illumina's HiSeq X with 151-bp pair-end reads. Data processing was based on the standardized best-practice method proposed by GATK (v 4.0.6.0)[57] with additional recalculation of

518 genotype quality (GQ) scores based on allele frequencies in the 1000 Genomes Project, phase 3, East
519 Asian (EAS) population using CalculateGenotypePosteriors module from GATK. Samples with
520 genotyping call rates < 99% were removed and multi-allelic sites were split into bi-allelic. Calls with
521 GQ < 20 and/or DP < 5 were assigned to missing. Variants with genotyping call rates < 85% and/or
522 HWE P-value < 1.0×10^{-6} were removed. Finally, we used BEAGLE (v 5.1)[58] to impute missing
523 genotypes. Variants with minor allele frequency < 1% and variants on actionable 59 genes by ACMG[59]
524 were excluded from our analysis because they were out of the scope of this research.
525
526 **Principal component analysis based on VDJ gene usage**
527 All B cell samples used in this study except those samples after belimumab treatment were utilized for
528 PCA analysis. For each sample, each VDJ gene usage was calculated by dividing the number of
529 clonotypes using that gene by the total number of clonotypes in that sample. We retained genes that
530 were used more than 0.5% by 10% or more of the samples. Each gene usage was scaled to mean = 0
531 and variance = 1, and PCA was performed with prcomp function in R.
532 To project samples after belimumab treatment to the same PCA space described above, for each
533 VDJ gene, we centered the mean and standardized the variance with regards to those calculated among
534 all the other B cell samples above. Then, this normalized VDJ gene usage matrix and the PC loadings
535 of each gene calculated in the above analysis were utilized for calculating PC scores for samples after
536 belimumab treatment.
537 We assessed the disease effects on PC scores by linear regression analysis comparing each disease
538 sample versus HC samples in each subset. Treatment effects on PC scores were assessed by the linear
539 mixed model comparing samples before and after treatment, with individuals treated as random effects.
540 The linear mixed model was fit with lme4 package.
541

Calculation of the Repertoire Naïveness Score

For scoring, we only included VDJ genes that were used by more than an average of 0.1% of the clonotypes in each of the 5 subsets. At first, each VDJ gene usage was scaled to mean = 0 and variance = 1 among the 5 B cell subsets of HCs. Then, the effect sizes of VDJ genes were estimated by comparing their usage between naïve B and USM B among HC samples with a linear mixed model, with individuals treated as random effects. Only genes that attained Bonferroni-adjusted p values < 0.05 in this comparison were utilized for the subsequent scoring. In addition, genes whose usage showed substantial correlation among 5 B cell subsets of HCs ($|\beta| > 0.4$ in the linear mixed model) were pruned to keep the most significantly different ones in the comparison of USM B and naïve B. As a result, 55 genes (35 V genes, 16 D genes and 4 J genes) were retained for the scoring (Supplementary table 1).

For each sample, VDJ gene usage was mean-centered and variance-standardized with regards to those calculated among the 5 B cell subsets of HCs. Then, the following formula was applied for calculating the RN score:

$$RNS_i = \sum_{k=1}^{35} \widehat{\beta_{V_k}} \times U_{V_{ik}} + \sum_{k=1}^{16} \widehat{\beta_{D_k}} \times U_{D_{ik}} + \sum_{k=1}^{4} \widehat{\beta_{J_k}} \times U_{J_{ik}}$$

where $\widehat{\beta_{v_k}}$ corresponded to the effect size of the V_k gene usage comparison between USM B and naïve B among HCs, and U_{Vik} corresponded to the normalized usage of the V_k gene in individual *i*.

Calculation of gene set enrichment score

We utilized the GSVA package[60] to calculate the gene set enrichment score for samples. After TMM normalization[61] of count data, conversion to log count per million values and batch effect removal with the ComBat function from the sva package[62], the "gsva" function was utilized for calculating the gene set enrichment score for the gene sets in Supplementary table 2. The Tph1 score was calculated using the 54 genes that were significantly upregulated in PD-1^{hi} CD4^{+} T cells compared with PD-1^{-}

566 populations in the original article[22] and normalized expression values of Th1 cells. The IFN score was
567 calculated using 97 genes in HALLMARK_INTERFERON_ALPHA_RESPONSE gene set in
568 MSigDB[63].
569
570 **Association analysis with gene expression**
571 The association of naïve B cell CDR-H3 length with gene expression in naïve B cells was assessed in
572 each disease at first for reducing the confounding of inter-disease differences in gene expression-CDR-
573 H3 length association. For this step, gene expression data after filtering of low expressed genes was
574 assessed for the association with CDR-H3 length using the limma[64] package with voom transformation.
575 Known batch effects and age were treated as covariates. Then, resultant effect sizes of each comparison
576 and their standard deviations were utilized for meta-analysis with Metasoft (version 2.0.1) software[65].
577 For the assessment of association of RN score and gene expression in each immune cell subset, we
578 also utilized limma with voom transformation. The RN score was scaled to mean = 0 and variance =
579 1, and known batch effects and age were treated as covariates. In this comparison, genes fulfilling FDR
580 < 0.05 and $|\log FC| > 0.1$ were treated as significantly associated genes.
581
582 **Gene set enrichment analysis**
583 We tested the enrichment of CDR-H3 length-associated genes that passed a Bonferroni significance
584 threshold in fixed-effect meta-analysis to hallmark gene sets by MsigDB[63] using the clusterProfiler
585 package[66].
586
587 **Differential usage analysis**
588 For the assessment of disease effects on VDJ gene usage in each subset, we kept genes that were used
589 more than 0.5% by 10% or more of the samples. For each disease, gene usage of the disease and HC

590 samples were scaled to mean = 0 and variance = 1 and tested for the disease effect on gene usage by
591 linear regression with age and known batch effects as covariates.
592
593 **Mediation analysis**
594 The relationships among IFN scores, Tph1 scores and RN scores in plasmablasts were tested among
595 SLE patients by utilizing the classical mediation analysis framework[67]. The IFN score was treated as
596 the independent variable, the Tph1 score was treated as the mediator and the RN score was treated as
597 the outcome variable. P values were calculated via 10000-time bootstrapping using mediation package
598 in R.
599
600 **Calculation of V gene mutation frequency**
601 For the assessment of gene mutation frequency, we quantified per-base mutation frequency in the
602 heavy-chain V gene except for the CDR3 region, that is, in FR1, CDR1, FR2, CDR2 and FR3 regions.
603 As MiXCR assembles the sequence of these regions after alignment and assembly of clonotypes based
604 on CDR3 sequence, only a portion of the nucleotides is actually sequenced in these regions. We
605 counted the number of mutations in these regions in each individual and divided it by the total number
606 of sequenced nucleotides in these regions in the individual to infer the V gene mutation frequency.
607
608 **Comparison with the data from the other cohort**
609 The BCR repertoire sequencing data of peripheral blood mononuclear cells from healthy individual[11]
610 was downloaded from the NCBI Sequencing Read Archive under BioProject number PRJNA527941.
611 The BCR repertoire sequencing data of sorted B cells from human tonsil[19] was downloaded from
612 ArrayExpress with the accession number E-MTAB-8999.
613 Gene expression data of PD-1^{hi} and PD-1^{-} T cells[22] were downloaded from the ImmPort repository

with the accession number SDY939. We performed the differential expression analysis between (memoryCD4^{+}PD-1^{hi}CXCR5^{-}MHCII^{-}ICOS^{+} + memoryCD4^{+}PD-1^{hi}CXCR5^{-}MHCII^{+}ICOS^{+}) versus memoryCD4^{+}PD-1^{-}CXCR5^{-} populations to obtain "PD-1^{hi} versus PD-1^{-} logFC".

To validate the reproducibility of the correlation between Tph1 score and RN score in plasmablasts, we utilized peripheral blood Th1 cells and plasmablast RNA-seq data of Japanese SLE patients and HCs that we had obtained previously[25], which have no sample overlap with this study.

Comparison of the frequency of IGHV4-34 clonotypes with nonmutated AVY/NHS motifs

We counted the number of IGHV4-34 clonotypes with or without mutations in the AVY motif in the FWR1 region or the NHS motif in the CDR2 region. Only clonotypes sequenced at these positions were included in the analysis. As the number of IGHV4-34 clonotypes was not large in one sample, we aggregated the IGHV4-34^{+} clonotypes in each comparison group (e.g., SLE plasmablasts, HC plasmablasts, etc.), and compared the cumulative number of mutated and nonmutated clonotypes between groups by Fisher's exact test.

SHM-QTL analysis

In each individual, all substitutions in V genes in plasmablasts were collected from "mutationsDetailed" columns for FR1, CDR1, FR2, CDR2, and FR3 regions from MiXCR output. The frequency of each of 6 nucleotide substitution pattern (i.e., C:G>A:T, C:G>T:A, C:G>G:C, T:A>C:G, T:A>G:C and T:A>A:T) was calculated in each individual. Nucleotide substitution pattern frequencies were normalized across samples using a rank based inverse normal transformation with the GenABEL package[68].

Association tests of these normalized substitution pattern frequencies and genetic variants were performed using QTLtools (v1.3.1)[69] with known batch effects, age, sex, disease and the top 2 genetic

638 principal components included as covariates. In total, 509 cases with whole genome information and
639 plasmablast repertoire information were included in this analysis.

640 Variance of substitution patterns explained by 5 lead SHM-QTL variants was estimated by
641 MANOVA analysis using the MVLM package. We calculated the variance of (6 -1) substitution
642 patterns explained by full model (including the 5 lead variants, known batch effects, age and disease
643 term) and a null model (including known batch effects, age and disease term) separately, and the
644 difference between the two values was defined as the variance explained by 5 variants.

645

646 **Data visualization and statistical test**

647 The statistical test performed are indicated in the figure legends or Methods. Throughout the analyses,
648 multiple test correction was performed with Benjamini-Hochberg procedure to obtain corrected q-
649 values unless otherwise indicated. In boxplots, the center lines indicate the median, the box limits
650 indicate the interquartile range (IQR), the whiskers reflect the maximum and minimum values no
651 further than 1.5 × IQR from the hinge and the points indicate outliers.

652

653 **Acknowledgements**
654 The super-computing resource was provided by Human Genome Center, Institute of Medical Sciences,
655 The University of Tokyo (http://sc.hgc.jp/shirokane.html). This study was supported by Chugai
656 Pharmaceutical Co., Ltd., Tokyo, Japan, the Center of Innovation Program from Japan Science and
657 Technology Agency (JST) (JPMJCE1304), the Ministry of Education, Culture, Sports, and the Japan
658 Agency for Medical Research and Development (AMED) (JP17ek0109103, 19ek0410047 and
659 JP20ek0410074).

660 **Author contributions**
661 M.O. conducted bioinformatics analysis with the help of Y.N. and K.I. M.N. contributed to the
662 collection and management of clinical information. M.O., M.N., Y.N., S.K., H.H., R.Y., T.I., Y.A., and
663 H.S. managed and contributed to sample collection, cell sorting, RNA sequencing and whole genome
664 sequencing. H.H., Y.A., and T.I. contributed to QC of the RNA-seq data. A.M. performed protein
665 structure prediction. T.O., K.Y. and K.F designed and managed the project. Y.T. contributed to critical

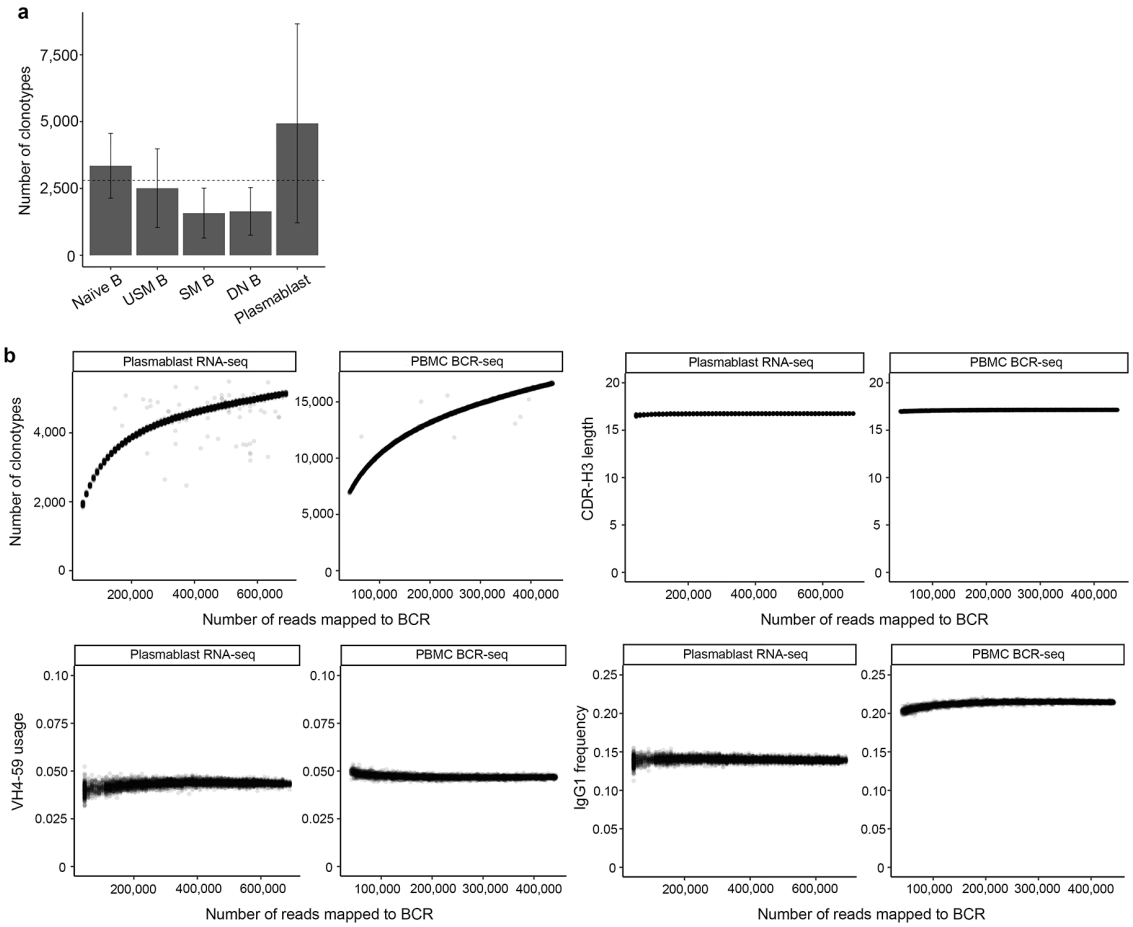
reading and revision of the manuscript. M.O. and K.F designed the study and wrote the manuscript with contributions from all authors on the final version of the manuscript.

Competing interests

M.O., Y.N., and T.O. belong to the Social Cooperation Program, Department of functional genomics and immunological diseases, supported by Chugai Pharmaceutical. A.M. is an employee of Chugai Pharmaceutical. K.F. receives consulting honoraria and research support from Chugai Pharmaceutical.

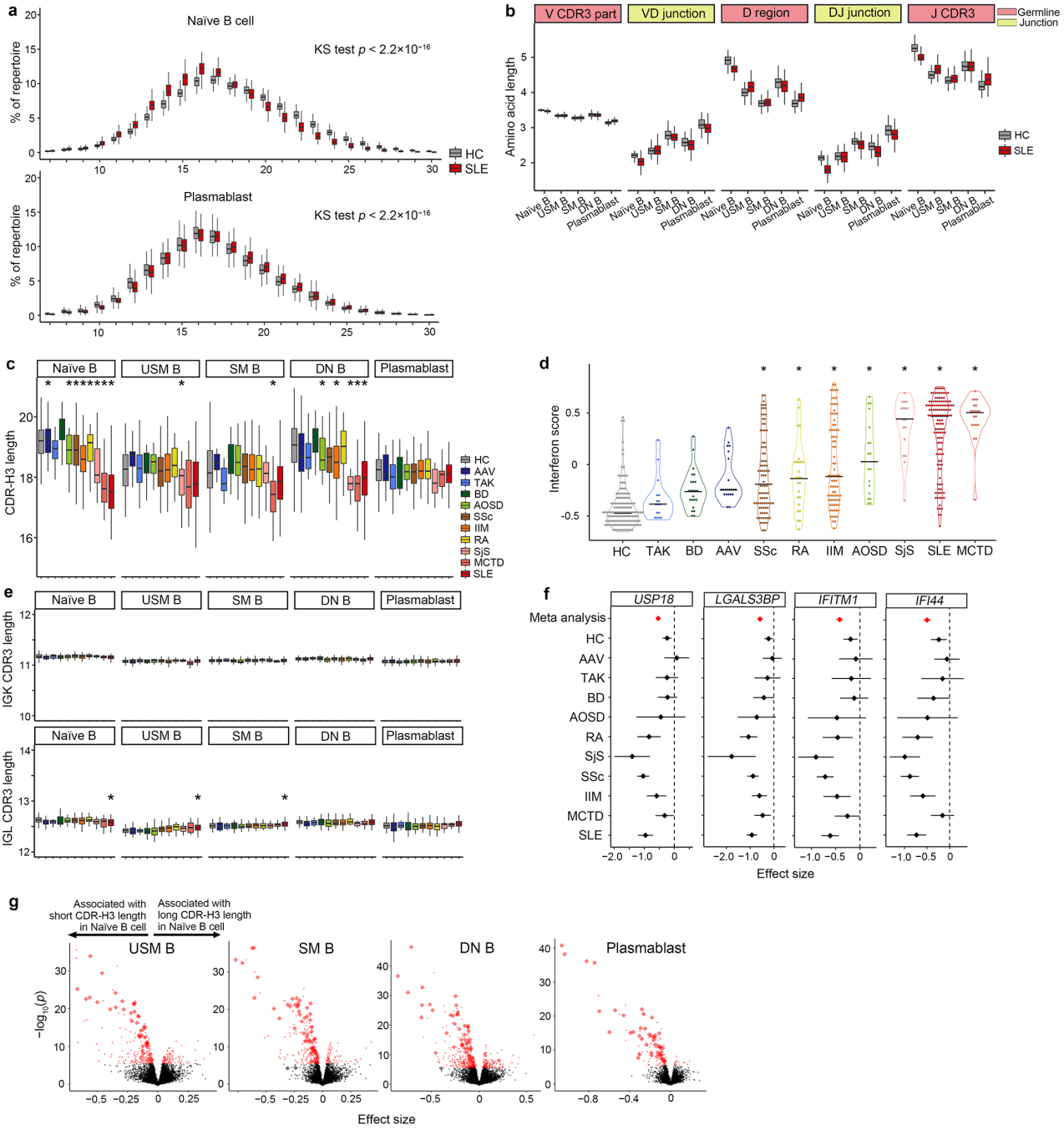
Data availability

BCR data used in this study will be available at the National Bioscience Database Center (NBDC) Human Database (https://humandbs.biosciencedbc.jp/en/) at the time of publication. We used publicly available software for the analyses. Custom code is available from the corresponding authors upon reasonable request.



Extended data Fig. 1: BCR repertoire characteristics and its association with read depth.

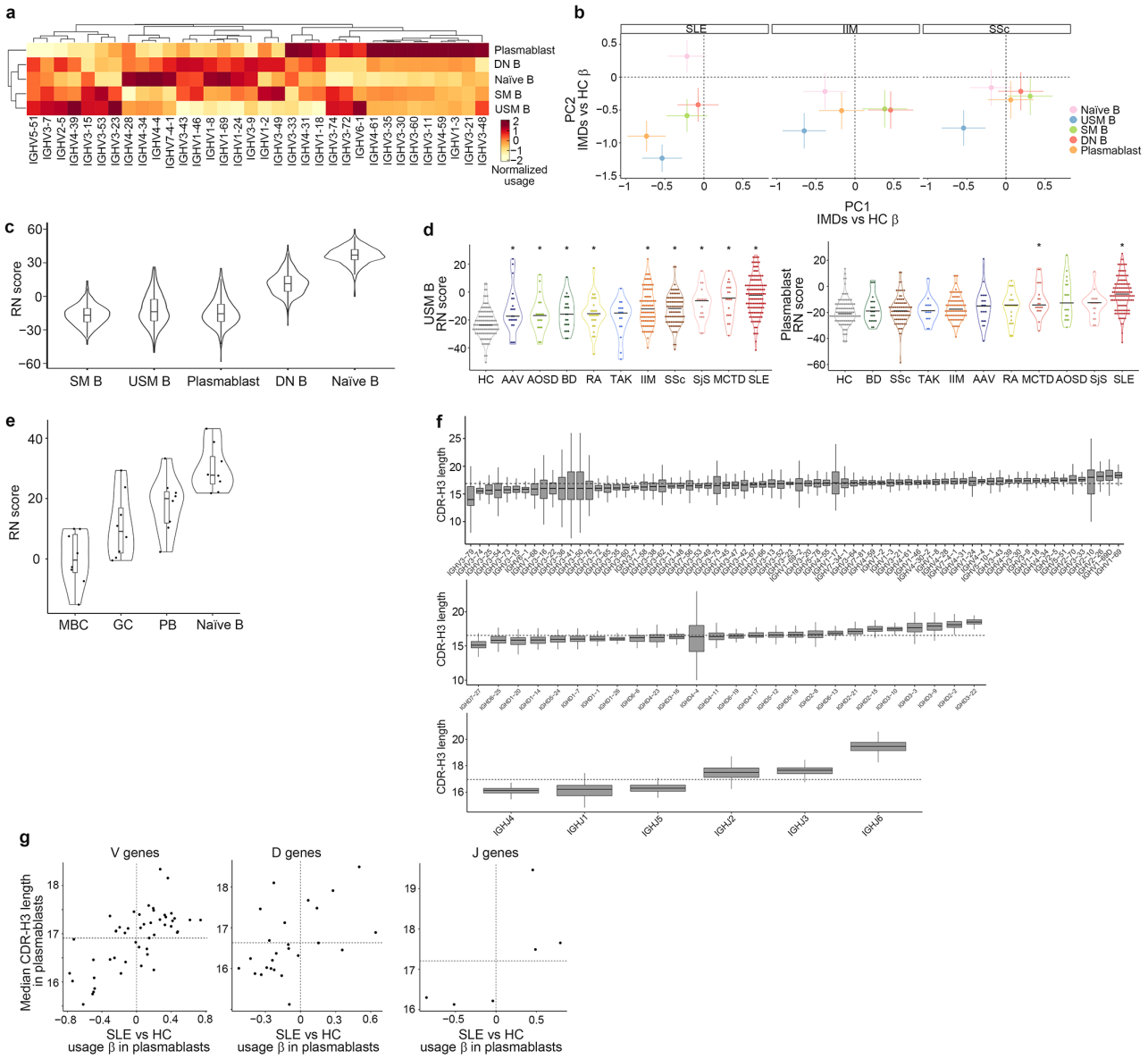
a, Mean number of clonotypes identified in each subset in our cohort. Bars indicate 2SD. Dotted line indicates the mean number of identified clonotypes among 5 subsets. **b,** Influence of the number of BCR-mapped reads to the BCR characteristics. The number of reads was down-sampled randomly from the FASTQ files of one plasmablast RNA-seq readout and one peripheral blood mononuclear cell BCR-seq readout[11]. Changes in the identified number of clonotypes, CDR-H3 length, V gene usage (represented by IGHV 4-59 usage) and isotype frequency (represented by IgG1 frequency) according to the number of reads are plotted.



688 **Extended data Fig. 2: Association of CDR-H3 length with IMDs and gene expression.**

689 **a,** Comparison of distribution of CDR-H3 lengths in SLE and HC in naïve B cells and plasmablasts. KS test,
690 Kolmogorov-Smirnov test. **b,** Comparison of lengths of sections that constitute the CDR-H3 region,
691 related to Fig. 2b. **c,** Comparison of CDR-H3 lengths between IMDs and HC among non-productive
692 clonotypes. Differences between HC and each disease were tested with linear regression adjusting for
693 age and batch effects. *, $p < 0.05/50$. **d,** Comparison of IFN scores in naïve B cells among IMD and

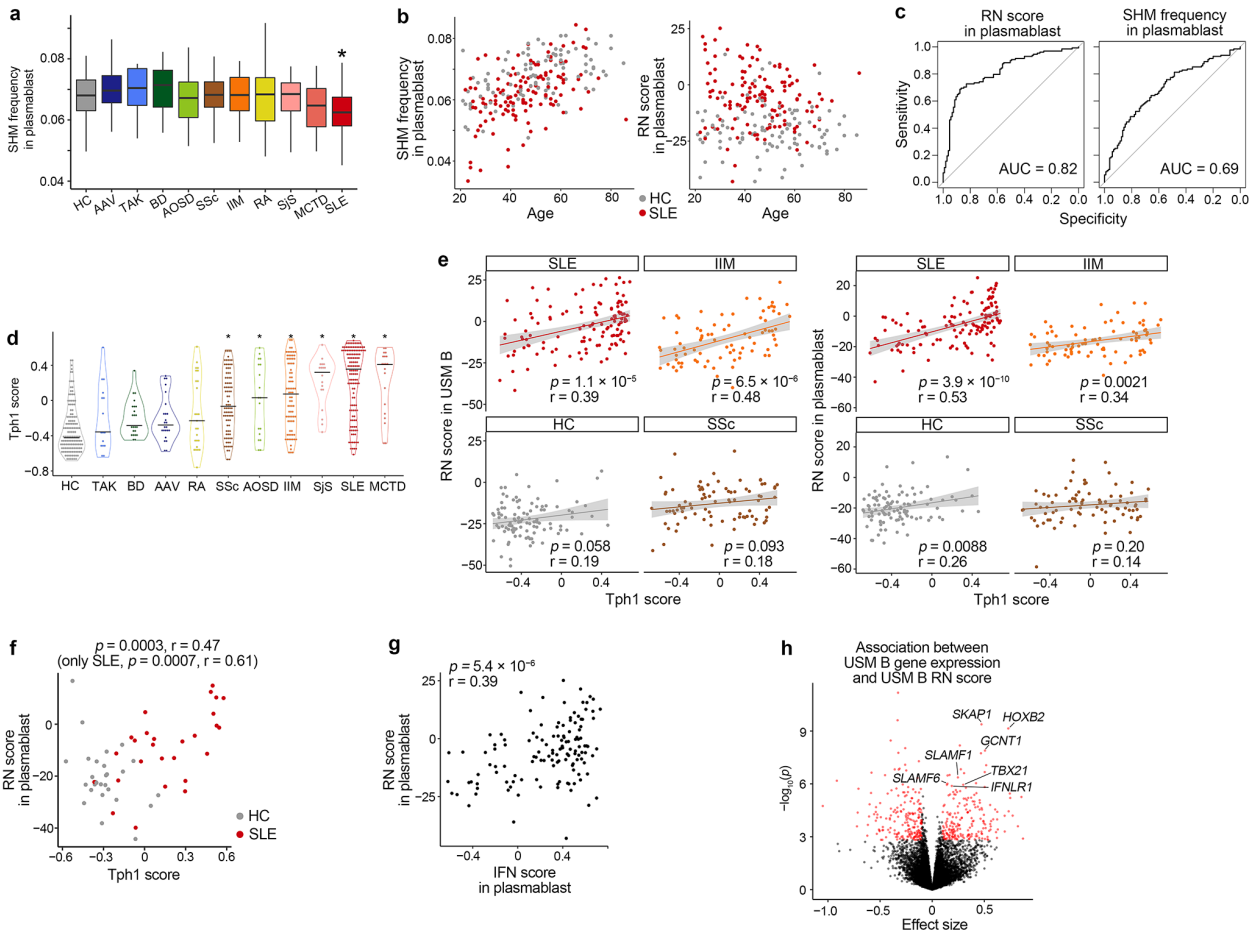
694 HC. Median values are marked with black line. Differences between HC and each disease were tested
695 with linear regression adjusting for age and batch effects. *, $p < 0.05/10$. **e,** Comparison of light chain
696 CDR3 length between IMD and HC. Differences between HC and each disease were tested with linear
697 regression adjusting for age and batch effects. *, $p < 0.05/50$. **f,** Effect sizes for the associations of
698 labeled genes and CDR-H3 lengths in each disease as well as fixed effect meta-analysis. Bars indicate
699 95% CI. **g,** Association of gene expression in 4 B cell subsets with naïve B CDR-H3 length, related to
700 Fig. 2c. Fixed effect meta-analysis *p* values and effect sizes are plotted. Genes attained Bonferroni-
701 adjusted $p < 0.05$ are marked red. Genes in "HALLMARK_INTERFERON_ALPHA_RESPONSE"
702 gene set are marked as diamonds.
703



Extended data Fig. 3: Comparison of gene usage patterns between IMDs and evaluation of the Repertoire Naïveness score.

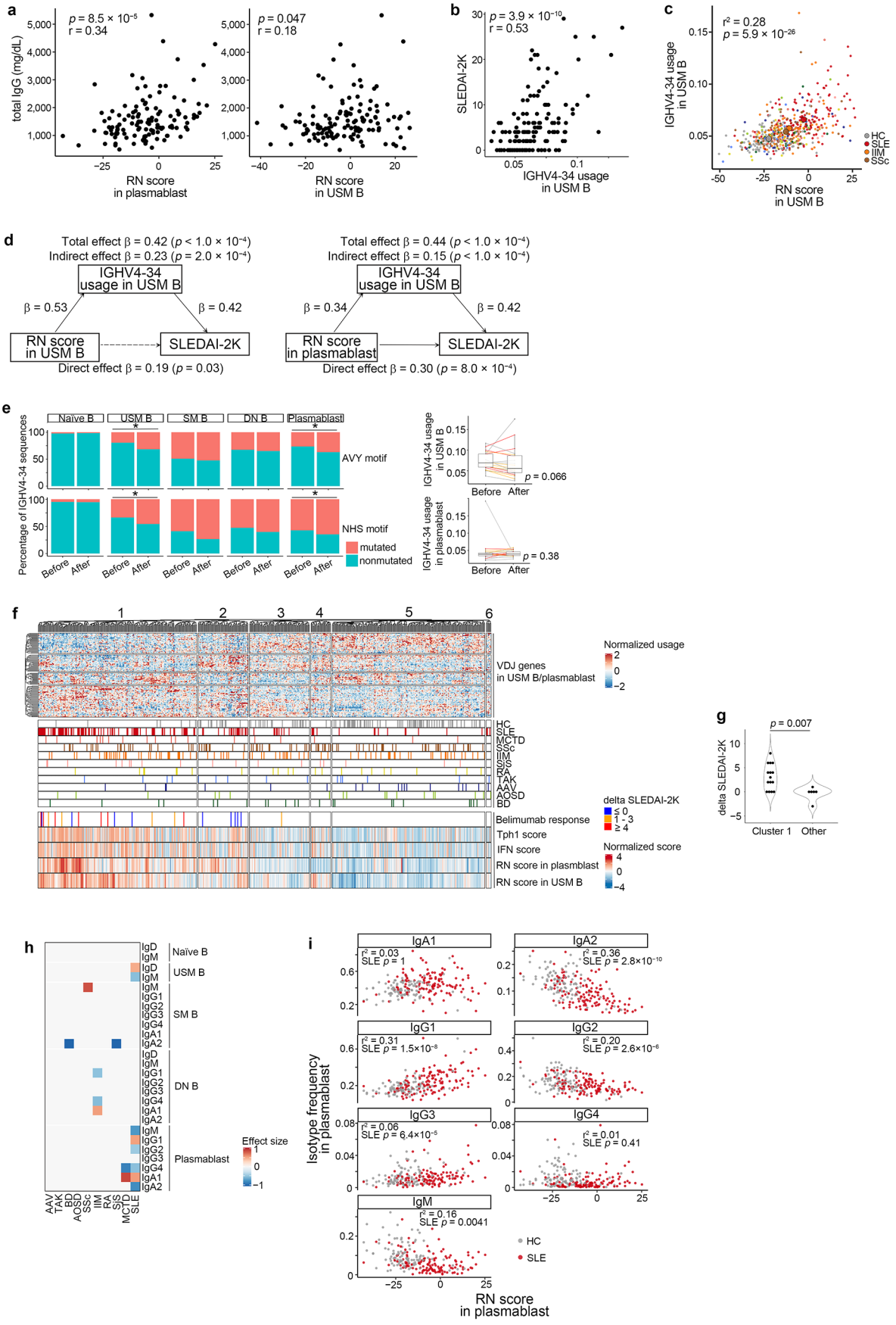
a, Comparison of mean V gene usage among 5 B cell subsets in HC. Genes that are utilized at least 1% in any of 5 subsets are shown. For comparison, gene usages were Z score normalized column-wise and hierarchically clustered. **b,** Disease effects on PC scores generated by linear regression analysis: comparison of each disease sample versus HC samples. Bars indicate 95% CI. **c,** Comparison of RN scores determined for 5 B cell subsets among all participants in our cohort. **d,** Comparison of RN scores among IMDs in USM B (left) and plasmablasts (right). Black bars indicate median values. Differences between HC and each disease were tested with linear regression adjusting for age and batch effects. *, $p < 0.05/10$. **e,** Comparison of RN scores among previously reported BCR-seq data of

714 sorted B cells from human tonsil. MBC, memory B cell. GC, germinal center B cell. PB, plasmablast.
715 Naïve, naïve B cell. **f,** Distribution of median CDR-H3 lengths of clonotypes utilizing each V (top), D
716 (middle), and J (bottom) gene in plasmablasts among HC. Dotted lines indicate the median CDR-H3
717 length among all clonotypes. **g,** Comparison of the effect size estimates of SLE versus HC gene usage
718 comparison in plasmablasts (x-axis) and median CDR-H3 length of each gene among HC in
719 plasmablasts (y-axis).
720



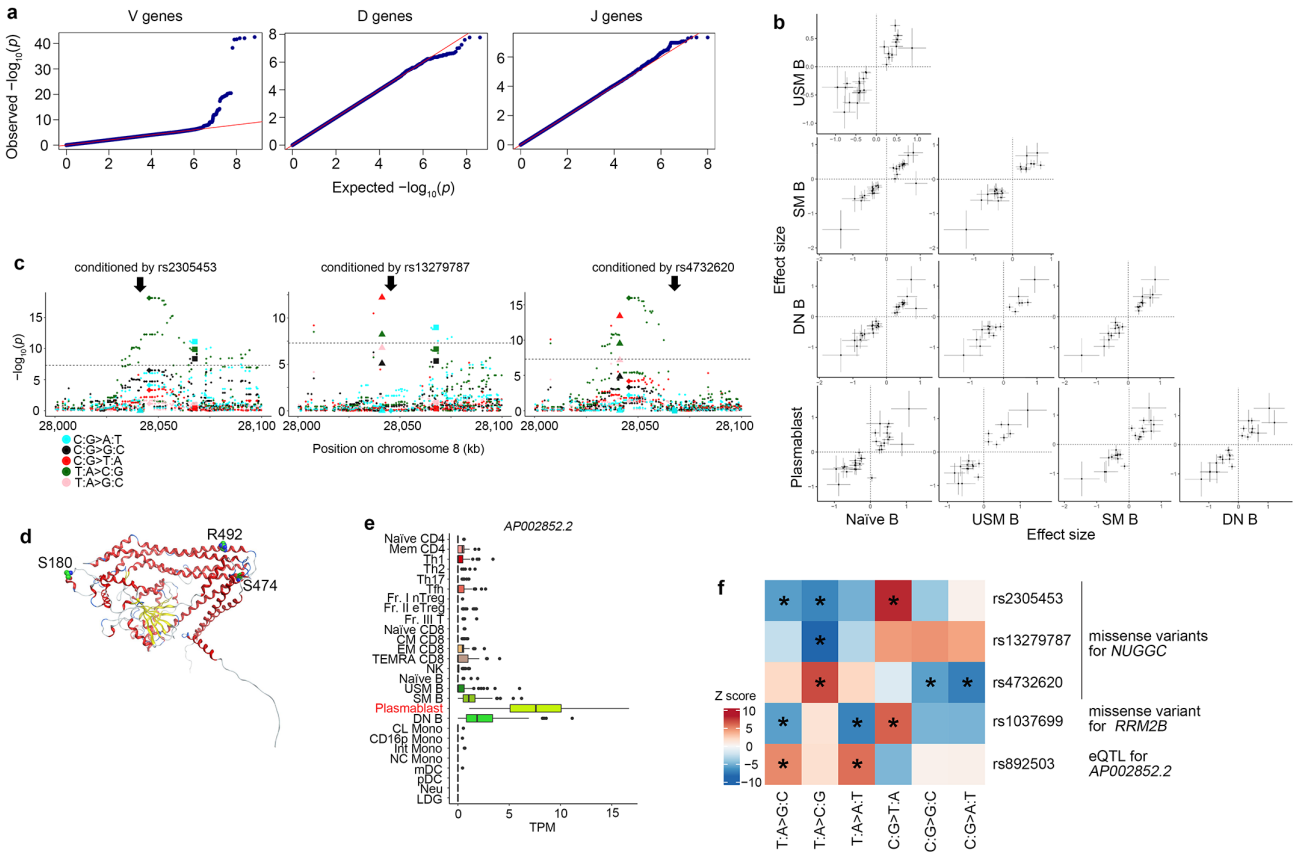
Extended data Fig. 4: Comparison of RN score-associated gene expression profiles.

a, Comparison of SHM frequencies in the V gene region of plasmablasts among IMDs. Differences between HC and each disease were tested with linear regression adjusting for age and batch effects. *, $p < 0.05/10$. **b,** Association of SHM frequencies (left) and RN scores of plasmablasts (right) with age. **c,** ROC curve for classifying SLE patients from HC using RN scores (left) or SHM frequencies (right) in plasmablasts. **d,** Comparison of Tph1 scores. Differences between HC and each disease were tested with linear regression adjusting for age and batch effects. *, $p < 0.05/10$. **e,** Correlation between Tph1 scores and RN scores in USM B (left) or plasmablasts (right). Pearson's r and its significance are provided. **f,** Correlation between Tph1 scores and RN scores in plasmablasts in an independent cohort of SLE patients and HC. **g,** Correlation between IFN scores and RN scores in plasmablasts among SLE patients. **h,** Association of gene expression in USM B cells and RN scores in USM B. Significantly associated genes are marked red.



734 **Extended data Fig. 5: Comparison of RN score, VH4-34 motifs and isotype frequencies.**

a, Correlation between RN scores and serum total IgG concentrations. Pearson's r and its significance are provided. **b,** Correlation between IGHV4-34 usage in USM B and SLEDAI-2K. Pearson's r and its significance are provided. **c,** Correlation between RN scores and IGHV4-34 usage in USM B. R^2 and *p* values were calculated with linear regression model. Mean values among HC, SLE, SSc, and IIM are marked with large points. The color of each point indicates the clinical diagnosis as illustrated in Fig. 1a. **d,** Mediation model representing the relationships among RN scores in USM B (left) or plasmablasts (right), IGHV4-34 usage in USM B and SLEDAI-2K among SLE patients. **e,** Comparison of the frequencies of clonotypes with mutations in AVY/NHS motifs among IGHV4-34 clonotypes (left) and IGHV4-34 usage (right) before and after treatment with belimumab. For motif comparison, Fisher's exact test comparing the number of mutated/nonmutated clonotypes before and after treatment samples was performed in each subset; *, $p < 0.01$. **f,** Clustering of cases with VDJ gene usage of USM B and plasmablasts that were utilized for RN score calculation. All the participants except for SLE patients after belimumab treatment were hierarchically clustered using average-linkage method based on Pearson's correlation. **g,** Comparison of changes in SLEDAI-2K after belimumab treatment (delta SLEDAI-2K) between SLE patients in cluster 1 and others. **h,** Skewed isotype frequencies in IMDs compared to HC. In each disease (column), isotype frequency (row) was compared with HC after adjusting for age and batch effects. Effect sizes of linear model comparison are plotted if the difference was statistically significant (FDR<0.05). **i,** Correlation between RN scores in plasmablasts and isotype frequency in plasmablast. *P* values and r^2 values were calculated with linear regression.



Extended data Fig. 6: Genetic association analysis with VDJ gene usage and somatic hypermutation.

a, QQ-plot of V, D, and J gene usage association with genome-wide genetic variants in naïve B cells. **b,** Comparison of V gene usage-QTL effect sizes among 5 B cell subsets. For each comparison of 2 subsets, the effect sizes were compared for significantly associated V gene-top variant pairs in either of the subsets. Bars indicate 95%CI. **c,** Regional association plot of the *NUGGC* locus after conditioning on rs2305453 (left), rs13279787 (middle), and rs4732620 (right). **d,** Estimated protein structure of NUGGC using the AlphaFold2 algorithm[70]. Three missense variants identified in SHM-QTL analysis are marked, two of which, Ser180 and R492, are believed to be located on the surface of the protein. **e,** Expression of *AP002852.2* in immune cells among HCs. **f,** Association of 5 independent SHM-QTL variants with 6 kinds of mutation patterns. *, $p < 5 \times 10^{-8}$.

Supplementary table legends

Table S1. Repertoire Naïveness Score calculation.

Table S2. Gene sets used for calculating gene set enrichment score.

Table S3. Abbreviations for immune cell subsets.

References

1. Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-330 (2015).

2. Ota, M. *et al.* Dynamic landscape of immune cell-specific gene regulation in immune-mediated diseases. *Cell* **184**, 3006-3021.e3017 (2021).

3. Mitka, M. Treatment for lupus, first in 50 years, offers modest benefits, hope to patients. *Jama* **305**, 1754-1755 (2011).

4. Edwards, J.C. *et al.* Efficacy of B-cell-targeted therapy with rituximab in patients with rheumatoid arthritis. *The New England journal of medicine* **350**, 2572-2581 (2004).

5. Liston, A., Humblet-Baron, S., Duffy, D. & Goris, A. Human immune diversity: from evolution to modernity. *Nat Immunol* **22**, 1479-1489 (2021).

6. van Kempen, T.S., Wenink, M.H., Leijten, E.F., Radstake, T.R. & Boes, M. Perception of self: distinguishing autoimmunity from autoinflammation. *Nat Rev Rheumatol* **11**, 483-492 (2015).

7. Elsner, R.A. & Shlomchik, M.J. Germinal Center and Extrafollicular B Cell Responses in Vaccination, Immunity, and Autoimmunity. *Immunity* **53**, 1136-1150 (2020).

8. Jenks, S.A. *et al.* Distinct Effector B Cells Induced by Unregulated Toll-like Receptor 7 Contribute to Pathogenic Responses in Systemic Lupus Erythematosus. *Immunity* **49**, 725-739.e726 (2018).

9. Tipton, C.M. *et al.* Diversity, cellular origin and autoreactivity of antibody-secreting cell population expansions in acute systemic lupus erythematosus. *Nat Immunol* **16**, 755-765 (2015).

10. Glass, D.R. *et al.* An Integrated Multi-omic Single-Cell Atlas of Human B Cell Identity. *Immunity* **53**, 217-232.e215 (2020).

11. Ghraichy, M. *et al.* Maturation of the Human Immunoglobulin Heavy Chain Repertoire With Age. *Front Immunol* **11**, 1734 (2020).

12. Miqueu, P. *et al.* Statistical analysis of CDR3 length distributions for the assessment of T and B cell repertoire biases. *Molecular immunology* **44**, 1057-1064 (2007).

13. Gomez-Tourino, I., Kamra, Y., Baptista, R., Lorenc, A. & Peakman, M. T cell receptor β-chains display abnormal shortening and repertoire sharing in type 1 diabetes. *Nat Commun* **8**, 1792 (2017).

14. Palanichamy, A. *et al.* Neutrophil-mediated IFN activation in the bone marrow alters B cell development in human and murine systemic lupus erythematosus. *Journal of immunology (Baltimore, Md. : 1950)* **192**, 906-918 (2014).

15. Mroczek, E.S. *et al.* Differences in the composition of the human antibody repertoire by B cell subsets in the blood. *Front Immunol* **5**, 96 (2014).

16. Kaplinsky, J. *et al.* Antibody repertoire deep sequencing reveals antigen-independent selection in maturing B cells. *Proceedings of the National Academy of Sciences of the United States of America* **111**, E2622-2629 (2014).

17. Bashford-Rogers, R.J.M. *et al.* Analysis of the B cell receptor repertoire in six immune-mediated diseases. *Nature* **574**, 122-126 (2019).

18. Demaison, C., David, D., Fautrel, B. & Theze, J. V(H) gene-family representation in peripheral activated B cells from systemic lupus erythematosus (SLE) patients. *Clinical and experimental immunology* **104**, 439-445 (1996).

19. King, H.W. *et al.* Single-cell analysis of human B cell maturation predicts how antibody class switching shapes selection dynamics. *Sci Immunol* **6**, eabe6291 (2021).

20. Sankar, K., Hoi, K.H. & Hötzel, I. Dynamics of heavy chain junctional length biases in antibody repertoires. *Commun Biol* **3**, 207 (2020).

21. William, J., Euler, C., Christensen, S. & Shlomchik, M.J. Evolution of autoantibody responses via somatic hypermutation outside of germinal centers. *Science (New York, N.Y.)* **297**, 2066-2070 (2002).

22. Rao, D.A. *et al.* Pathologically expanded peripheral T helper cell subset drives B cells in rheumatoid arthritis. *Nature* **542**, 110-114 (2017).

23. Bocharnikov, A.V. *et al.* PD-1hiCXCR5- T peripheral helper cells promote B cell responses in lupus via MAF and IL-21. *JCI Insight* **4**, e130062 (2019).

24. Makiyama, A. *et al.* Expanded circulating peripheral helper T cells in systemic lupus erythematosus: association with disease activity and B cell differentiation. *Rheumatology (Oxford, England)* **58**, 1861-1869 (2019).

25. Takeshima, Y. *et al.* Immune cell multi-omics analysis reveals contribution of oxidative phosphorylation to B cell functions and organ damage of lupus. *bioRxiv*, 2021.10.08.463629 (2021).

26. Kenderes, K.J. *et al.* T-Bet(+) IgM Memory Cells Generate Multi-lineage Effector B Cells. *Cell Rep* **24**, 824-837.e823 (2018).

27. Rubtsova, K. *et al.* B cells expressing the transcription factor T-bet drive lupus-like autoimmunity. *J Clin Invest* **127**, 1392-1404 (2017).

28. Reed, J.H., Jackson, J., Christ, D. & Goodnow, C.C. Clonal redemption of autoantibodies by somatic hypermutation away from self-reactivity during human immunization. *The Journal of experimental medicine* **213**, 1255-1265 (2016).

29. Schickel, J.-N. *et al.* Self-reactive VH4-34-expressing IgG B cells recognize commensal bacteria. *The Journal of experimental medicine* **214**, 1991-2003 (2017).

30. Banchereau, J. & Pascual, V. Type I interferon in systemic lupus erythematosus and other autoimmune diseases. *Immunity* **25**, 383-392 (2006).

31. Tobón, G.J., Izquierdo, J.H. & Cañas, C.A. B lymphocytes: development, tolerance, and their role in autoimmunity-focus on systemic lupus erythematosus. *Autoimmune Dis* **2013**, 827254

(2013).

32. Rubelt, F. *et al.* Individual heritable differences result in unique cell lymphocyte receptor repertoires of naïve and antigen-experienced cells. *Nat Commun* **7**, 11112 (2016).

33. Schramm, C.A. & Douek, D.C. Beyond Hot Spots: Biases in Antibody Somatic Hypermutation and Implications for Vaccine Design. *Front Immunol* **9**, 1876 (2018).

34. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS genetics* **10**, e1004383 (2014).

35. Liu, S. *et al.* Direct measurement of B-cell receptor repertoire's composition and variation in systemic lupus erythematosus. *Genes Immun* **18**, 22-27 (2017).

36. Meffre, E. *et al.* Immunoglobulin heavy chain expression shapes the B cell receptor repertoire in human B cell development. *J Clin Invest* **108**, 879-886 (2001).

37. Wong, E.B., Khan, T.N., Mohan, C. & Rahman, Z.S. The lupus-prone NZM2410/NZW strain-derived Sle1b sublocus alters the germinal center checkpoint in female mice in a B cell-intrinsic manner. *Journal of immunology (Baltimore, Md. : 1950)* **189**, 5667-5681 (2012).

38. Woods, M., Zou, Y.R. & Davidson, A. Defects in Germinal Center Selection in SLE. *Front Immunol* **6**, 425 (2015).

39. Yurasov, S. *et al.* Defective B cell tolerance checkpoints in systemic lupus erythematosus. *The Journal of experimental medicine* **201**, 703-711 (2005).

40. Cappione, A., 3rd *et al.* Germinal center exclusion of autoreactive B cells is defective in human systemic lupus erythematosus. *J Clin Invest* **115**, 3205-3216 (2005).

41. Muramatsu, M. *et al.* Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. *Cell* **102**, 553-563 (2000).

42. Pilzecker, B. & Jacobs, H. Mutating for Good: DNA Damage Responses During Somatic Hypermutation. *Front Immunol* **10**, 438 (2019).

43. Richter, K. *et al.* Speckled-like pattern in the germinal center (SLIP-GC), a nuclear GTPase

expressed in activation-induced deaminase-expressing lymphomas and germinal center B cells. *J Biol Chem* **284**, 30652-30661 (2009).

44. Richter, K. *et al.* Altered pattern of immunoglobulin hypermutation in mice deficient in Slip-GC protein. *J Biol Chem* **287**, 31856-31865 (2012).

45. Tanaka, H. *et al.* A ribonucleotide reductase gene involved in a p53-dependent cell-cycle checkpoint for DNA damage. *Nature* **404**, 42-49 (2000).

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

47. van den Hoogen, F. *et al.* 2013 classification criteria for systemic sclerosis: an American College of Rheumatology/European League against Rheumatism collaborative initiative. *Arthritis Rheum* **65**, 2737-2747 (2013).

48. Bohan, A. & Peter, J.B. Polymyositis and dermatomyositis (first of two parts). *The New England journal of medicine* **292**, 344-347 (1975).

49. Bohan, A. & Peter, J.B. Polymyositis and dermatomyositis (second of two parts). *The New England journal of medicine* **292**, 403-407 (1975).

50. Hoogendijk, J.E. *et al.* 119th ENMC international workshop: trial design in adult idiopathic inflammatory myopathies, with the exception of inclusion body myositis, 10-12 October 2003, Naarden, The Netherlands. *Neuromuscul Disord*, vol. 14: England, 2004, pp 337-345.

51. Sontheimer, R.D. Would a new name hasten the acceptance of amyopathic dermatomyositis (dermatomyositis sine myositis) as a distinctive subset within the idiopathic inflammatory dermatomyopathies spectrum of clinical illness? *J Am Acad Dermatol* **46**, 626-636 (2002).

52. Griggs, R.C. *et al.* Inclusion body myositis and myopathies. *Ann Neurol* **38**, 705-713 (1995).

53. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)* **29**, 15-21 (2013).

54. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic acids research* **47**, D766-d773 (2019).

55. Anders, S., Pyl, P.T. & Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics (Oxford, England)* **31**, 166-169 (2015).

56. Bolotin, D.A. *et al.* MiXCR: software for comprehensive adaptive immunity profiling. *Nat Methods* **12**, 380-381 (2015).

57. DePristo, M.A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics* **43**, 491-498 (2011).

58. Browning, B.L., Zhou, Y. & Browning, S.R. A One-Penny Imputed Genome from Next-Generation Reference Panels. *Am J Hum Genet* **103**, 338-348 (2018).

59. Kalia, S.S. *et al.* Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics. *Genetics in Medicine* **19**, 249-255 (2017).

60. Hänzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* **14**, 7 (2013).

61. Robinson, M.D., McCarthy, D.J. & Smyth, G.K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)* **26**, 139-140 (2010).

62. Leek, J.T., Johnson, W.E., Parker, H.S., Jaffe, A.E. & Storey, J.D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics (Oxford, England)* **28**, 882-883 (2012).

63. Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* **1**, 417-425 (2015).

64. Ritchie, M.E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research* **43**, e47 (2015).

65. Han, B. & Eskin, E. Interpreting meta-analyses of genome-wide association studies. *PLoS genetics* **8**, e1002555 (2012).

66. Yu, G., Wang, L.G., Han, Y. & He, Q.Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics* **16**, 284-287 (2012).

67. Baron, R.M. & Kenny, D.A. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *J Pers Soc Psychol* **51**, 1173-1182 (1986).

68. Aulchenko, Y.S., Ripke, S., Isaacs, A. & van Duijn, C.M. GenABEL: an R library for genome-wide association analysis. *Bioinformatics (Oxford, England)* **23**, 1294-1296 (2007).

69. Delaneau, O. *et al.* A complete tool set for molecular QTL discovery and analysis. *Nat Commun* **8**, 15452 (2017).

70. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589 (2021).