LETTER

https://doi.org/10.1038/s41586-019-1595-3

Analysis of the B cell receptor repertoire in six immune-mediated diseases

R. J. M. Bashford-Rogers[1,2*], L. Bergamaschi[1,3], E. F. McKinney[1,3], D. C. Pombal[1,3], F. Mescia[1,3], J. C. Lee[1,3], D. C. Thomas[1], S. M. Flint[1,5], P. Kellam[4], D. R. W. Jayne[1], P. A. Lyons[1,3] & K. G. C. Smith[1,3*]

B cells are important in the pathogenesis of many, and perhaps all, immune-mediated diseases. Each B cell expresses a single B cell receptor (BCR)[1], and the diverse range of BCRs expressed by the total B cell population of an individual is termed the 'BCR repertoire'. Our understanding of the BCR repertoire in the context of immune-mediated diseases is incomplete, and defining this could provide new insights into pathogenesis and therapy. Here, we compared the BCR repertoire in systemic lupus erythematosus, anti-neutrophil cytoplasmic antibody (ANCA)-associated vasculitis, Crohn's disease, Behçet's disease, eosinophilic granulomatosis with polyangiitis, and immunoglobulin A (IgA) vasculitis by analysing BCR clonality, use of immunoglobulin heavy-chain variable region (IGHV) genes and—in particular—isotype use. An increase in clonality in systemic lupus erythematosus and Crohn's disease that was dominated by the IgA isotype, together with skewed use of the IGHV genes in these and other diseases, suggested a microbial contribution to pathogenesis. Different immunosuppressive treatments had specific and distinct effects on the repertoire; B cells that persisted after treatment with rituximab were predominately isotype-switched and clonally expanded, whereas the inverse was true for B cells that persisted after treatment with mycophenolate mofetil. Our comparative analysis of the BCR repertoire in immune-mediated disease reveals a complex B cell architecture, providing a platform for understanding pathological mechanisms and designing treatment strategies.

During the development of B cells in the bone marrow (or fetal liver)[2], recombination of immunoglobulin genes forms the 'naive' BCR repertoire, which is modified by the removal or suppression of self-reactive B cells to reduce the chance of autoimmune disease[3] (although 20–40% of B cells remain autoreactive[4]). Further diversification of the repertoire occurs after B cells respond to antigen. Many B cells undergo 'isotype switching', in which stepwise DNA deletion and recombination from immunoglobulin M (IgM) generates downstream isotypes (IgG1, IgG2, IgG3 and IgG4, IgA1 and IgA2, IgD and IgE) that confer distinct functional characteristics and roles in disease[5,6]. Delineation of isotypes is thus vital for a full analysis of the BCR repertoire. Further diversification of BCRs occurs in specialized germinal centres, in which somatic hypermutation (SHM) of genes that encode the variable (V) regions of antibodies may enhance BCR affinity and specificity[7]. This diversification of B cell clones after exposure to antigen is tempered by tolerance checkpoints to reduce the risk of autoimmunity[8]. The peripheral BCR repertoire is thus a composite of both the naive repertoire and that generated by antigenic encounter.

Features of the BCR repertoire have been correlated with both microbial interactions and immune-mediated diseases (IMDs); specific IGHV regions recognize commensal and/or pathogenic microorganisms or are associated with IMDs (Supplementary Table 1). We analysed the BCR repertoire in 209 individuals across 6 IMDs (Extended Data Fig. 1a, Supplementary Tables 2, 3), comparing (i) IMDs that are characterized by autoantibody responses against either single dominant autoantigens (ANCA-associated vasculitis (AAV)) or multiple autoantigens (systemic lupus erythematosus (SLE)); (ii) IMDs that are not thought to be autoimmune (Crohn's disease, Behçet's disease); and (iii) IMDs for which there is incomplete evidence of B cell involvement or autoimmunity (eosinophilic granulomatosis with polyangiitis (EGPA; formerly Churg–Strauss syndrome), IgA vasculitis (IgAV; formerly Henoch–Schönlein purpura)). For disease descriptions, see Supplementary Discussion.

We developed a method to barcode, amplify and sequence BCR repertoires from RNA that encodes the antigen-binding (IgH (VDJ)) and constant regions of the BCR heavy chain. This method facilitates analysis of isotype class and subclass, as well as allowing the quantitation of clone frequency and correction of PCR- or sequencing-based error[9] (Extended Data Fig. 1b). We then analysed the BCR repertoire in sorted B cells from 19 healthy control individuals (Supplementary Discussion, Extended Data Figs. 1, 2) to develop methods to control for the effects of age and different cellular RNA content (Methods, Extended Data Figs. 2, 3a–c, Supplementary Table 4). We define the 'normalized' isotype use—that is, the percentage of unique VDJ sequences per isotype—thus counting the contribution of each B cell to the repertoire only once.

Comparative studies in IMDs have often been confounded by differences in disease duration, activity and treatment. We therefore specifically recruited patients for whom there was objective evidence of active disease and who had not yet commenced treatment (although stable doses of low-level therapy, which are known not to affect BRC repertoire, were permitted; Methods, Supplementary Discussion). The majority of the patients were newly diagnosed. For all patients, the number of B cells sampled was higher than the number of unique BCR sequences detected (Supplementary Table 3). We compared isotype use in repertoires from unseparated peripheral blood mononuclear cells (PBMCs) in healthy controls and patients with IMDs (Fig. 1a, b, Extended Data Fig. 3d). Compared to the control samples, IgA was over-represented in all diseases except AAV and EGPA, and particularly so in SLE and Crohn's disease. This corresponded with an increase in the levels of serum IgA, which was most pronounced in SLE (Fig. 1c). Expression of IgE was increased in SLE, Crohn's disease and—in particular—EGPA (Fig. 1b, Extended Data Fig. 3d, e), which also exhibited an increase in IgG3. Isotype use in AAV was similar to healthy controls. There is therefore marked variation in isotype use in IMDs, and IgA is the dominant isotype in diseases such as SLE and Behçet's disease.

The diversity of the BCR repertoire is driven in part by differential use of IGHV genes, as well as non-template additions and deletions of nucleotides. Some individual genes, and IGHV subgroups (defined by structural similarity[10]), preferentially bind microbial antigens and/or have been associated with autoimmunity (Supplementary Table 1). We examined the frequency of IGHV genes in naive and antigen-experienced B cells across IMDs (Fig. 1d, Extended Data Fig. 4a, Supplementary Data 2, 3). Expression of genes of the IGHV4 family was increased in Crohn's disease, SLE and EGPA, as was that of

[1]Department of Medicine, University of Cambridge, Cambridge, UK. [2]Wellcome Centre for Human Genetics, University of Oxford, Oxford, UK. [3]Cambridge Institute for Therapeutic Immunology and Infectious Disease, Jeffrey Cheah Biomedical Centre, University of Cambridge, Cambridge, UK. [4]Department of Medicine, Division of Infectious Diseases, Imperial College, London, UK. [5]Present address: ImmunoInflammation Therapy Area Unit, GlaxoSmithKline, Stevenage, UK. *e-mail: rbr1@well.ox.ac.uk; kgcs2@cam.ac.uk

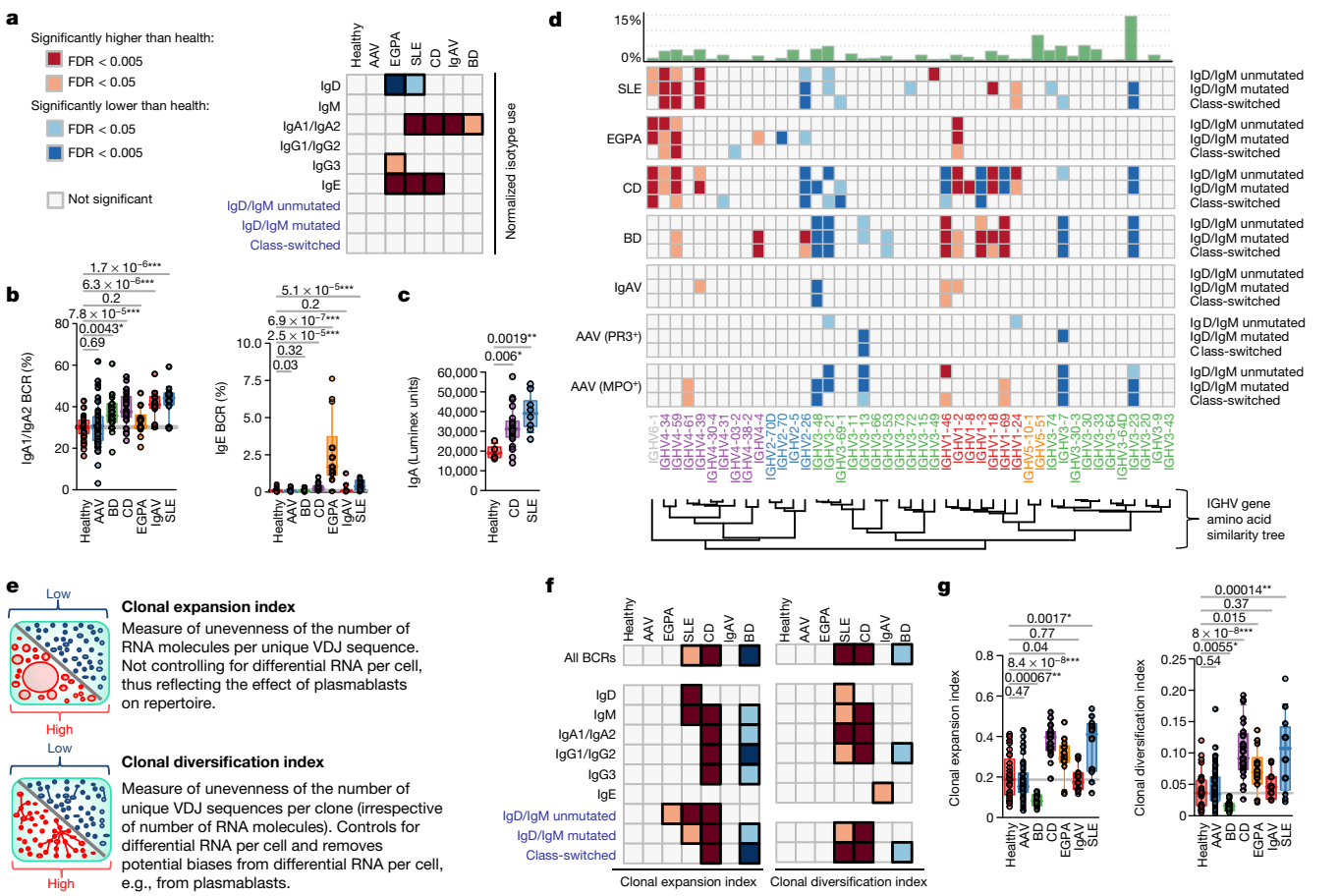


Fig. 1 | Differences in isotype use, IGHV gene use and clonality between IMDs. a, Heat map of normalized isotype use per disease. BD, Behçet's disease; CD, Crohn's disease. The light and dark orange squares indicate significantly higher, and light and dark blue squares significantly lower, isotype use in disease compared to health. **b**, Normalized percentage use of the IgA1 or IgA2 and IgE BCR per disease. **c**, IgA titre in healthy individuals ($n = 4$), patients with Crohn's disease ($n = 20$) and patients with SLE ($n = 8$). **d**, Heat map of IGHV gene frequency and BCR subtypes in health and disease: $IgM^+D^+SHM^-$ BCR sequences (over 78% derived from naive B cells); $IgM^+D^+SHM^+$ BCR sequences (SHM is evidence of antigenic stimulation); and IgM^-D^- BCR sequences (all isotype-switched and therefore post-antigenic). Light and dark orange squares indicate significantly higher, and light and dark blue squares significantly lower, gene frequency in disease than health. Only genes that occurred at a higher frequency than 0.1% are shown. Relative mean gene frequencies in healthy individuals are indicated at the top (full heat map in Supplementary Data 3). IGHV genes are ordered according to amino acid similarity, as indicated by the IGHV gene amino acid similarity tree (see Methods). **e**, Explanations of clonality measures and network representations of BCR repertoires. **f**, Heat map showing the clonal expansion index (left) and clonal diversification index (right) of each isotype per disease from total PBMC B cells. **g**, Clonal expansion index (left) and clonal diversification index (right) for BCR repertoires in PBMCs per disease. For **a**, **b**, **d**, **f**, **g**, $n = 32$ for healthy individuals and $n = 20$, $n = 34$, $n = 12$, $n = 10$, $n = 23$, $n = 10$ and $n = 13$ for patients with AAV (MPO^+), AAV ($PR3^+$), EGPA, SLE, Crohn's disease (CD), IgAV and Behçet's disease (BD), respectively. For **a**–**c**, *P* values were calculated by two-sided analysis of variance (ANOVA); *denotes a false discovery rate (FDR) <0.05, **FDR < 0.005, ***FDR < 0.0005 (determined by the Šidák method). For all box plots, box lines show the 25th, 50th and 75th percentiles; whiskers show the upper and lower quartiles.

IGHV6-1. Notably, IGHV4-34 binds both autoantigens[11] and commensal bacteria[12], and has been associated with SLE[13]. Our data extend this association of *IGHV4-34* (and its 9G4 idiotype) with SLE to EGPA and Crohn's disease (Extended Data Fig. 4b). Both *IGHV6-1* and *IGHV4-59* have been associated with autoreactivity (Supplementary Table 4). The associations between IGHV genes and these IMDs were seen in both the predominantly 'naive' and the 'post-antigenic' compartments, and in both non-expanded and expanded clones (Extended Data Fig. 4a), raising the possibility that they are not purely a consequence of selective expansion after disease development (except in Crohn's disease, in which the differences in IGHV genes were predominantly 'pre-antigenic'). Genes of the V1 family were over-represented in IMDs, particularly Crohn's disease and Behçet's disease. The most striking association was that of Behçet's disease with *IGHV1-46*, *IGHV1-3* and *IGHV1-69*—all of which have been previously associated with infection in both the naive and post-antigenic repertoires (previous studies are listed in Supplementary Table 1). Reduced representation of IGHV genes was also seen in some diseases, reflecting either a proportional reduction due to the increased frequency of other IGHV genes, or real disease associations. Levels of SHM did not vary between diseases (Extended Data Fig. 4c).

An increase in the length of complementarity-determining region 3 (CDR3) of the BCR is associated with antibody polyreactivity and autoimmunity[14]. Building on previous work[9], we found an association between the length of CDR3 and the use of IGHV genes in healthy individuals (Extended Data Figs. 2, 4d). In the case of patients with IMDs, increased CDR3 length was found in SLE (IgG and IgA) and Crohn's disease (unswitched B cells) (Extended Data Fig. 4c).

B cell clones are defined by sharing a unique VDJ rearrangement, and can be characterized by size (clonal expansion) and diversification (owing to SHM and isotype switching). Using a clone-sampling method to visualize the BCR repertoire (Supplementary Discussion, Extended Data Fig. 5), we found no differences between samples from healthy controls and patients with AAV or IgAV, reduced clonality in Behçet's disease, but increased clonal expansion and complexity in Crohn's disease, EGPA and SLE (Supplementary Data 5, 6). We

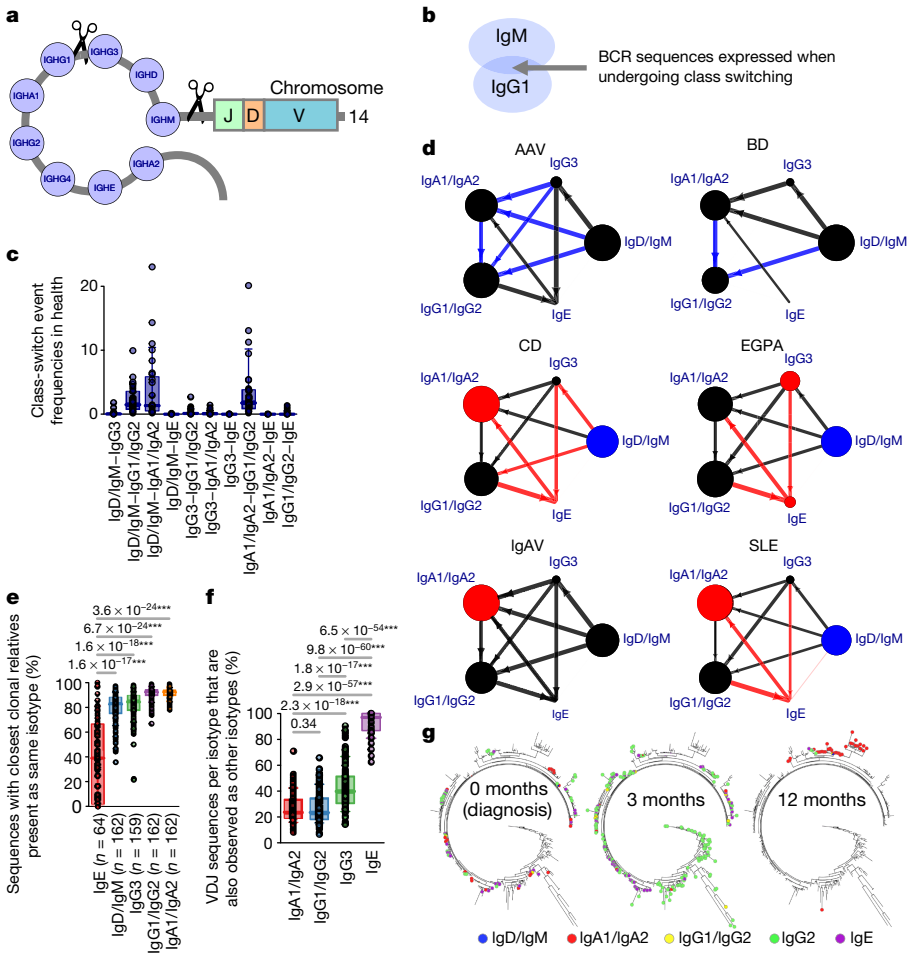


Fig. 2 | Class switching in IMDs. a, Schematic diagram of CSR. **b**, Relative frequencies of CSR between different constant regions may be determined through the frequency of unique VDJ regions expressed as two isotypes normalized for read depth. **c**, Relative CSR event frequencies in healthy individuals ($n = 32$). **d**, CSR frequencies across autoimmune diseases. Each circle represents an isotype class per disease; the size of the circle is proportional to the percentage of unique BCRs that correspond to that isotype, and circles are coloured according to whether the percentage is significantly higher than (red), lower than (blue) or not different from (black) that of healthy individuals. Arrows indicate class switching between isotypes; arrow thickness is proportional to the relative frequency of CSR events for each disease, and arrows are coloured according to whether the frequency is significantly higher than (red), lower than (blue) or not different from (black) healthy individuals. *P* values were calculated by two-way ANOVA and the FDR was determined by the Šidák method; FDR < 0.05 was used as the threshold for significance. **e**, Proportion of VDJ sequences for which the closest clonal relatives are also present as the same isotype (across health and IMDs; $n = 149$). **f**, Proportion of VDJ sequences per isotype that are also observed as other isotypes (across health and IMDs; $n = 149$). **g**, Representative phylogenetic tree of an IgE-associated clone maintained over the course of therapy from a patient with EGPA who was negative for ANCA (patient 145). Colours indicate isotype use and time point for each BCR. All nodes are scaled to unitary size. For **d**–**f**, *P* values were calculated by two-sided Wilcoxon test; *FDR < 0.05, **FDR < 0.005, ***FDR < 0.0005 (determined by the Šidák method). For all box plots, box lines show the 25th, 50th and 75th percentiles; whiskers show the upper and lower quartiles.

extended this analysis by determining the clonal expansion index (a measure of the 'unevenness' of the number of RNA molecules per unique VDJ region sequence, defined by the Gini index[15]) and the clonal diversification index, defined by Renyi entropy (a measure of the unevenness of unique VDJ region sequences per clone) (Methods, Fig. 1e–g, Extended Data Fig. 6). Patients with Crohn's disease had increased clonal expansion and diversification across many isotypes, particularly IgA, IgG and IgM. SLE showed a similar pattern—although with increased clonality primarily in unswitched cells and with greater variation between patients—as did EGPA (but with IgE predominant for the latter). Differences in the maximum clone size were consistent with these data (Extended Data Figs. 6c, d, 7a). By contrast, patients with active AAV or IgAV showed no gross difference in clonal expansion or diversification, and in Behçet's disease both were reduced compared to controls. We then used a multivariate comparison to assess 'clonal normality' (see Methods), and found significant dissimilarity between the repertoires of patients with Crohn's disease, EGPA and SLE, compared to those of healthy individuals and patients with AAV and Behçet's disease (Extended Data Fig. 7b)—reinforcing the concept that although some diseases are associated with broad abnormalities of the BCR repertoire, others show a BCR repertoire that is comparatively normal.

Class-switch recombination (CSR) is a deletional DNA recombination process, so the order of constant regions on the chromosome defines the possible isotypes to which any given B cell can switch (Fig. 2a, Extended Data Fig. 7c, d). The progression of CSR between each possible constant region (switch events) may be assessed by quantifying the frequency of unique VDJ regions that share two isotypes (suggesting their common clonal origin; Fig. 2b) after normalizing for read depth (Extended Data Fig. 7e, f). The number of class-switching types that are detectable in this analysis is reduced by the isotype ambiguity between IgA1 and IgA2, and IgG1 and IgG2, in the isotype-specific sequencing, and by alternative splicing of IgD from transcripts that contain IgM (Extended Data Fig. 8a, b). We confirmed reported frequencies of switch events in healthy individuals[16] (Fig. 2c). Switching differences between isotypes in IMDs usually corresponded

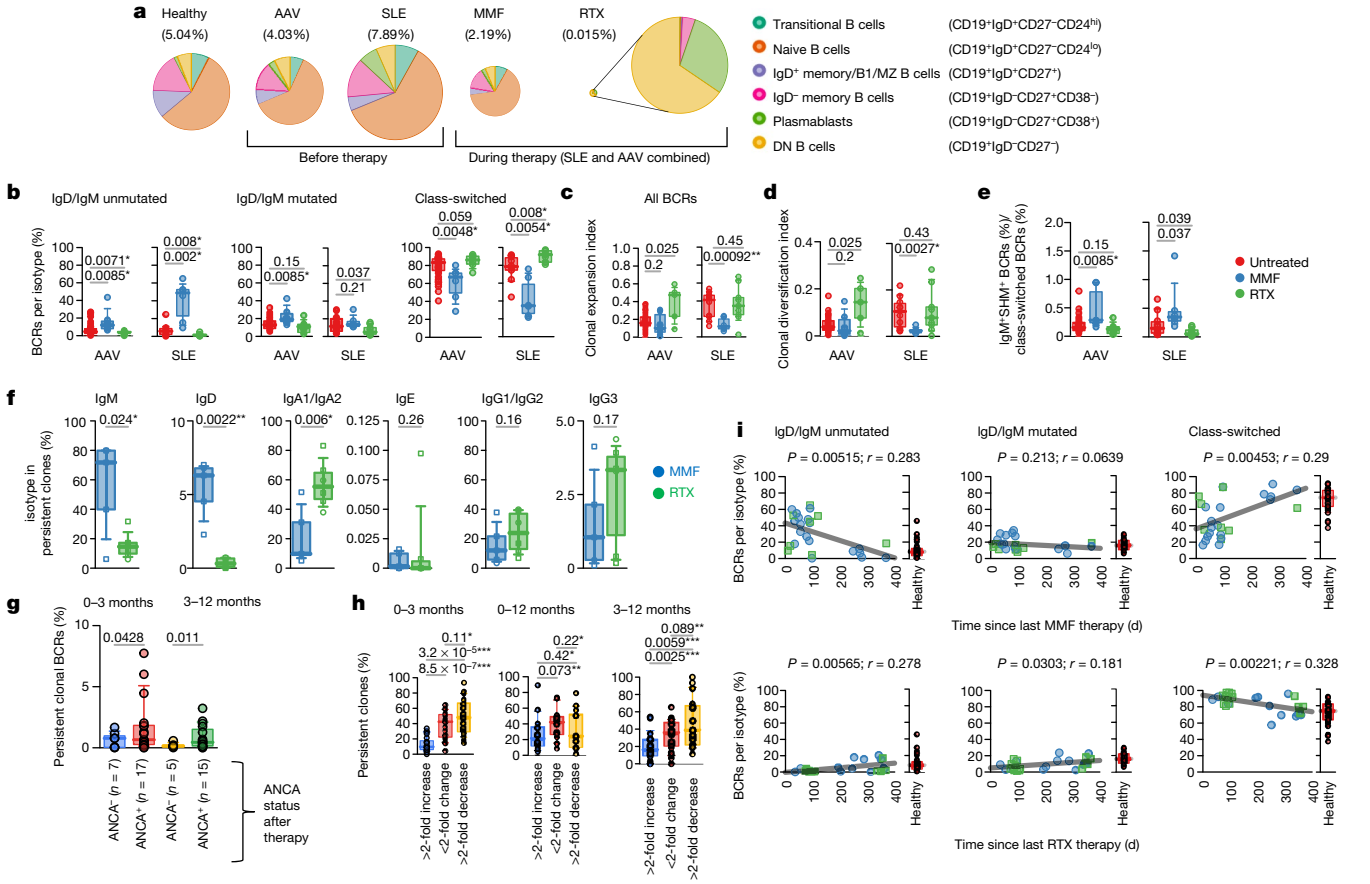


Fig. 3 | Effects of therapy on the BCR repertoire. a, Mean proportion and phenotype distribution of B cells within PBMCs in healthy controls and patients with AAV or SLE before and after therapy (MMF or RTX). The size of the pie chart corresponds to the proportion of B cells within PBMCs (the percentage of B cells is given in brackets), and pie segments represent the mean proportions of the indicated cell types. Data from patients with AAV and SLE are combined after therapy. DN, double negative; MZ, marginal zone. **b**–**d**, Percentage of unmutated IgD or IgM, mutated IgD or IgM, and antigen-experienced switched BCRs (IgA1, IgA2, IgG1, IgG2, IgG3 and IgE) (**b**), clonal expansion index (**c**), clonal diversification index (**d**) and ratio of the percentage of IgM^+SHM^+ BCRs over class-switched BCRs (**e**) of samples that were taken from patients with AAV or SLE at diagnosis (red, untreated) and 3 months after treatment with MMF (blue) or RTX (green). For AAV: untreated, $n = 42$; MMF, $n = 5$; RTX, $n = 5$; and for SLE: untreated, $n = 11$; MMF, $n = 6$; RTX, $n = 9$. **f**, Isotype percentage of persistent clones at diagnosis and after induction therapy in patients with AAV (MMF, $n = 5$; RTX, $n = 6$). **g**, Percentage of persistent BCRs shared between samples that were taken from patients with AAV at diagnosis and at 3 months, or between samples that were taken 3 months and 12 months after induction therapy, respectively, split between patients who became negative for serum ANCA after induction therapy versus those who remained serum-ANCA positive. **h**, Percentage of persistent clones that expanded by more than twofold, changed by less than twofold or decreased by more than twofold between the time of diagnosis and after induction therapy in patients with AAV ($n = 12$, $n = 20$ and $n = 19$ for 0–3 months, 0–12 months and 3–12 months, respectively). **i**, Correlation between proportions of BCR types with time since last treatment with MMF (top; $n = 27$) and RTX (bottom; $n = 26$) for AAV (blue) and SLE (green). Pearson's correlation coefficients and *P* values are indicated. BCR frequencies from healthy individuals are shown in red. For **b**–**h**, *P* values were calculated by two-sided ANOVA; *FDR < 0.05, **FDR < 0.005, ***FDR < 0.0005 (determined by the Šidák method). For all box plots, box lines show the 25th, 50th and 75th percentiles; whiskers show the upper and lower quartiles.

with differences in isotype use (Fig. 2d). All switching was reduced in AAV and Behçet's disease. In SLE and Crohn's disease, the representation of IgA increased, and switching between IgA and IgE was seen. The increase in isotype switching in Crohn's disease appeared to be independent of isotype frequency. In EGPA, there was a striking increase in switching to IgE from all isotypes (Fig. 2d), particularly IgG3—perhaps secondary to the increased frequency of IgG3 (Extended Data Fig. 9a, b). This first—to our knowledge—systematic analysis of isotype switching in IMD reveals disease-specific increases that contribute to isotype profiles. Some of these, such as the prominence of IgG3 and IgE in EGPA, and the reduced switching in AAV and Behçet's disease, were unexpected and may be relevant to disease pathogenesis.

Our analysis of the BCR repertoire supports suggestions in the literature[17,18] that human IgE clones usually arise from clonally diversified memory cells of precursor isotypes, as in mice[19]. Consistent with this, IgE^+ peripheral blood B cells are commonly plasmablasts (Extended Data Fig. 2), have fewer closest clonal relatives that express the IgE isotype (Fig. 2e, f, Extended Data Fig. 9c, d, Supplementary Discussion), and probably share a clonal origin with cells that do not express IgE. IgE also commonly arises from multiple independent switch events in large clones (Fig. 2g).

Different immunosuppressive regimens have different effects on the B cell compartment, and these can correlate with clinical efficacy[20]. We investigated the effect of treatment on the BCR repertoire in SLE and AAV, taking repeat samples at 3 or 12 months after diagnosis. Most patients were treated with rituximab (RTX; a B cell-depleting anti-CD20 monoclonal antibody) or mycophenolate mofetil (MMF; a precursor of an inosine 5′-monophosphate dehydrogenase inhibitor that predominantly affects proliferating cells[21]). These regimens were standardized based on international guidelines, but not administered according a formal protocol (reflecting real-world clinical practice) and were accompanied by similar steroid and subsequent maintenance therapy (Supplementary Discussion, Methods), allowing their effects on the BCR repertoire to be compared.

MMF and RTX had markedly different effects on the repertoire (Fig. 3a–e, Extended Data Fig. 10a–c). MMF therapy resulted in an

increased proportion of IgM^+ and IgD^+ B cells and concomitantly led to a reduction in the number of isotype-switched B cells and a reduction in clonality, with relative preservation of both IgM clones that underwent somatic hypermutation (SHM^+) and those that did not (SHM^-) compared to switched clones. This could be consistent with a shorter half-life for switched but not IgM memory B cells in humans (as seen in mice[22]), adding to the ongoing debate on this topic[23,24]. Conversely, after RTX, the numbers of circulating B cells were low[25] but persisting cells were largely isotype-switched and clonally expanded—predominantly IgA in AAV and IgG1 or IgG2 in SLE. Larger studies are required to determine whether these changes in the repertoire associate with disease subsets, pathogenic clonal persistence and/or treatment efficacy. Nonetheless, this suggests that the effect of a treatment on the BCR repertoire might inform the design of therapeutic strategies (for example, the ability of MMF to reduce class-switched clones might suggest that MMF could be effective in preventing relapse after RTX therapy).

Clonal persistence 3 months after therapy was observed in over 90% of patients, and the isotype of persistent clones differed between therapies (Fig. 3f, Extended Data Fig. 10d, e, Methods). In AAV, reduced persistence of isotype-switched clones was associated with reduced ANCA titre (Fig. 3g). Persistent clones could expand, undergo SHM and isotype switch despite continuing therapy (Fig. 3h). By considering the time between the last dose of MMF or RTX and the collection of the sample, we could analyse the 'recovery' of the repertoire. After treatment with MMF, the isotype-switched population reached healthy levels after approximately one year (Fig. 3i). By contrast, the slow reconstitution of IgD^+ and/or IgM^+ unmutated cells after RTX is consistent with the known kinetics of B cell recovery after such depletion[26].

This study reveals profound variation in many aspects of the BCR repertoire across IMDs, both at diagnosis and after therapy. Many of the disease-associated changes have not been previously described, in particular those relating to isotype use. The changes in the BCR repertoire in these diseases illustrate deficiencies in our understanding of disease pathogenesis (Supplementary Discussion). SLE, Crohn's disease and EGPA exhibited abnormal isotype-specific clonal expansion or diversity, as well as abnormal use of the IGHV genes; such a broad dysregulation of the repertoire is consistent with their associations with multiple antibodies. Increased IgA isotype usage was expected in an intestinal disease like Crohn's disease, but not in SLE, in which IgG is implicated in pathogenesis and intestinal inflammation is not prominent[27]. These observations suggest an unanticipated commonality in the pathogenesis of SLE, Crohn's disease and EGPA, suggesting that they might share unknown drivers—perhaps within the mucosal microbiome, given known IGHV affinities for microbial antigens[11–13,28]. EGPA also displayed an expansion of IgG3 and disproportionate switching to IgE. The IgE association was expected[29], but whether an increased IgG3 isotype usage is important in EGPA pathogenesis remains uncertain. IgAV was associated with increased IgA isotype usage and mucosal involvement, but showed no evidence of IgA clonal expansion or abnormal use of the IGHV genes—consistent with the pathogenesis of IgAV being distinct from that of Crohn's disease. It is also possible to have severe active autoimmune disease, such as AAV, without detectable changes in the BCR repertoire; the pathogenic anti-MPO or anti-PR3 clones are presumably too infrequent to skew repertoire analysis at the PBMC level. Finally, Behçet's disease showed a marked increase in the expression of *IGHV1-46*, *IGHV1-69* and *IGHV1-3*. The proteins that are encoded by these genes bind to both microbial antigens and autoantigens (Supplementary Table 4), enhancing previous speculation that infection might drive disease[30]. Future studies of the BCR repertoire that expand on this work—for example, by comparing the repertoire to the microbiome or determining the antigenic specificity of expanded clones—would be illuminating. Altogether, this comprehensive analysis of the BCR repertoire across diseases reveals a complex architecture, which may provide a platform for better understanding pathological mechanisms and designing therapeutic strategies.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1595-3.

Received: 20 December 2018; Accepted: 21 August 2019;
Published online 25 September 2019.

1. Nossal, G. J. V. & Lederberg, J. Antibody production by single cells. *Nature* **181**, 1419–1420 (1958).
2. Lydyard, P. M., Whelan, A. & Fanger, M. W. *Instant Notes in Immunology* (Bios Scientific, Oxford, 2000).
3. Nemazee, D. Mechanisms of central tolerance for B cells. *Nat. Rev. Immunol.* **17**, 281–294 (2017).
4. Wardemann, H. et al. Predominant autoantibody production by early human B cell precursors. *Science* **301**, 1374–1377 (2003).
5. Stavnezer, J. & Schrader, C. E. IgH chain class switch recombination: mechanism and regulation. *J. Immunol.* **193**, 5370–5378 (2014).
6. Stavnezer, J., Guikema, J. E. & Schrader, C. E. Mechanism and regulation of class switch recombination. *Annu. Rev. Immunol.* **26**, 261–292 (2008).
7. De Silva, N. S. & Klein, U. Dynamics of B cells in germinal centres. *Nat. Rev. Immunol.* **15**, 137–148 (2015).
8. Giltiay, N. V., Chappell, C. P. & Clark, E. A. B-cell selection and the development of autoantibodies. *Arthritis Res. Ther.* **14**, S1 (2012).
9. Petrova, V. N. et al. Combined influence of B-cell receptor rearrangement and somatic hypermutation on B-cell class-switch fate in health and in chronic lymphocytic leukemia. *Front. Immunol.* **9**, 1784 (2018).
10. Matsuda, F. et al. The complete nucleotide sequence of the human immunoglobulin heavy chain variable region locus. *J. Exp. Med.* **188**, 2151–2162 (1998).
11. Pascual, V. et al. Nucleotide sequence analysis of the V regions of two IgM cold agglutinins. Evidence that the VH4-21 gene segment is responsible for the major cross-reactive idiotype. *J. Immunol.* **146**, 4385–4391 (1991).
12. Schickel, J. N. et al. Self-reactive VH4-34-expressing IgG B cells recognize commensal bacteria. *J. Exp. Med.* **214**, 1991–2003 (2017).
13. 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).
14. 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).
15. Bashford-Rogers, R. J. M. et al. Network properties derived from deep sequencing of human B-cell receptor repertoires delineate B-cell populations. *Genome Res.* **23**, 1874–1884 (2013).
16. Horns, F. et al. Lineage tracing of human B cells reveals the *in vivo* landscape of human antibody class switching. *eLife* **5**, e16578 (2016).
17. Saunders, S. P., Ma, E. G. M., Aranda, C. J. & Curotto de Lafaille, M. A. Non-classical B cell memory of allergic IgE Responses. *Front. Immunol.* **10**, 715 (2019).
18. Croote, D., Darmanis, S., Nadeau, K. C. & Quake, S. R. High-affinity allergen-specific human antibodies cloned from single IgE B cell transcriptomes. *Science* **362**, 1306–1309 (2018).
19. He, J. S. et al. IgG1 memory B cells keep the memory of IgE responses. *Nat. Commun.* **8**, 641 (2017).
20. Jayne, D. R., Gaskin, G., Pusey, C. D. & Lockwood, C. M. ANCA and predicting relapse in systemic vasculitis. *QJM* **88**, 127–133 (1995).
21. Karnell, J. L. et al. Mycophenolic acid differentially impacts B cell function depending on the stage of differentiation. *J. Immunol.* **187**, 3603–3612 (2011).
22. Tarlinton, D. & Good-Jacobson, K. Diversity among memory B cells: origin, consequences, and utility. *Science* **341**, 1205–1211 (2013).
23. Seifert, M. & Küppers, R. Human memory B cells. *Leukemia* **30**, 2283–2292 (2016).
24. Macallan, D. C. et al. B-cell kinetics in humans: rapid turnover of peripheral blood memory cells. *Blood* **105**, 3633–3640 (2005).
25. Mei, H. E. et al. Steady-state generation of mucosal IgA^+ plasmablasts is not abrogated by B-cell depletion therapy with rituximab. *Blood* **116**, 5181–5190 (2010).
26. Anolik, J. H. et al. Delayed memory B cell recovery in peripheral blood and lymphoid tissue in systemic lupus erythematosus after B cell depletion therapy. *Arthritis Rheum.* **56**, 3044–3056 (2007).
27. Villalta, D. et al. Anti-dsDNA antibody isotypes in systemic lupus erythematosus: IgA in addition to IgG anti-dsDNA help to identify glomerulonephritis and active disease. *PLoS One* **8**, e71458 (2013).
28. Bende, R. J. et al. Identification of a novel stereotypic IGHV4-59/IGHJ5-encoded B-cell receptor subset expressed by various B-cell lymphomas with high affinity rheumatoid factor activity. *Haematologica* **101**, e200–e203 (2016).
29. Manger, B. J. et al. IgE-containing circulating immune complexes in Churg–Strauss vasculitis. *Scand. J. Immunol.* **21**, 369–373 (1985).
30. Galeone, M., Colucci, R., D'Erme, A. M., Moretti, S. & Lotti, T. Potential infectious etiology of Behçet's disease. *Patholog. Res. Int.* **2012**, 595380 (2012).

METHODS

Ethical approval. Ethical approval for this study was obtained from the Cambridge Local Research Ethics Committee (reference numbers 04/023, 08/H0306/21, 08/H0308/176) and Eastern NHS Multi Research Ethics Committee (07/MRE05/44). Informed consent was obtained from all subjects enrolled.

Samples from healthy participants. The inclusion criteria for healthy individuals were people aged between 20 and 77 years, with no serious co-morbidities, no direct family history of autoimmune disease, no use of immunosuppressants or steroids, and no hospitalization within the last 12 months. The healthy individual samples used for B cell sorting were recruited through the NIHR Cambridge BioResource.

Samples from patients with AAV. Patients with AAV who attended or were referred to the specialist vasculitis unit at Addenbrooke's Hospital, Cambridge between July 2004 and June 2016 were enrolled. Active disease at presentation was defined by at least one major or three minor criteria using the Birmingham Vasculitis Activity Score (BVAS)[31] and the clinical impression that induction immunosuppression would be required. Prospective disease monitoring was undertaken monthly with serial BVAS assessment[31] and serum ANCA status (Supplementary Discussion). Of a total of 54 patients, 41 were sampled at diagnosis and 13 at disease flare as defined above. A minority of patients (11/54) had received prior treatment with oral prednisolone, and 3 patients had received azathioprine within 6 months before sampling. Patients on low-dose steroids and azathioprine have been analysed separately, and their inclusion does not affect any of the findings described in this study.

Samples from patients with SLE. The SLE cohort comprised patients who attended or were referred to the Addenbrooke's Hospital specialist vasculitis unit between July 2004 and June 2016 who met at least four American College of Rheumatology SLE criteria[32] and presented with active disease. Active disease was defined as meeting all three of the following prospectively defined criteria: new British Isles Lupus Assessment Group (BILAG) score A or B in any system, clinical assessment of active disease by the reviewing physician and increase in immunosuppressive therapy as a result. After treatment with an immunosuppressant, patients were followed up monthly. Disease monitoring was undertaken with serial BILAG assessment and serum antinuclear antibody (ANA) status. Treatment of the patients was at the physician's discretion, not dictated by study participation and includes therapy used for induction of remission at enrolment ('induction'). Of a total of 10 patients, 8 were sampled at diagnosis and 2 at disease flare. A minority of patients (3/10) had received prior treatment with oral prednisolone and/or hydroxychloroquine.

Samples from patients with Crohn's disease. Patients with active Crohn's disease were recruited from a specialist inflammatory bowel disease (IBD) clinic at Addenbrooke's Hospital, before starting treatment. Of the 23 patients, 22 were recruited at the time of diagnosis. Diagnosis was made using standard endoscopic, histological and radiological criteria[33]. All patients had at least moderately active Crohn's disease at enrolment as evidenced by clinical symptoms in conjunction with some or all of elevated C-reactive protein, elevated faecal calprotectin, radiologically active disease or endoscopically active disease. All patients were treatment naive, with none receiving immunomodulators, corticosteroids or biological therapy.

Samples from patients with chronic lymphocytic leukaemia. Patients with chronic lymphocytic leukaemia (CLL) were recruited from the specialist leukaemia and lymphoma unit at Addenbrooke's Hospital between January 2011 and July 2014. CLL patient inclusion required the presence of at least 5×10^9 monoclonal B cells per litre of peripheral blood persisting for 3 months and a characteristic phenotype (typically CD5, CD19, CD20 and CD23).

Samples from patients with EGPA. Patients with EGPA who attended or were referred to the specialist vasculitis unit at Addenbrooke's Hospital between July 2004 and June 2016 were enrolled. Diagnosis of EGPA was based on the history or presence of both asthma and eosinophilia ($>1.0 \times 10^9\,l^{-1}$ and/or >10% of leukocytes) plus at least two additional features of EGPA as per the criteria used in the recent Phase III clinical trial 'Study to Investigate Mepolizumab in the Treatment of Eosinophilic Granulomatosis with Polyangiitis'[34]. Of 11 patients in total, 7 were sampled at diagnosis and 4 at disease flare. A minority of patients (4/11) had received prior treatment with oral steroids (methylprednisolone or prednisolone), 2/11 patients had been treated with azathioprine and 1/11 patients treated with cyclophosphamide within 6 months of sampling.

Samples from patients with IgAV and Behçet's disease. Patients with IgAV and Behçet's disease were recruited from the specialist vasculitis clinic at Addenbrooke's Hospital and enrolled into the present study between 2005 and 2015. The clinical data that were recorded for patients with Behçet's disease were as follows: (i) the basis for diagnosis, that is, orogenital mucosal ulceration, prior ocular inflammation and characteristic skin rash (erythema nodosum or pseudofolliculitis); (ii) major complications such as venous or arterial thrombosis, involvement of the central nervous system or of the pulmonary vascular system; and (iii) disease activity (expert physician global assessment). Of 11 patients, 5 had received prior treatment with oral steroids (prednisolone) and 3 had been treated with azathioprine within 6 months before sampling.

The diagnosis of IgAV was based on the American College of Rheumatology 1990 criteria for the classification of Henoch–Schönlein purpura[35] and the 2012 Revised International Chapel Hill Consensus Conference Nomenclature of Vasculitides[36]. All patients had to have a biopsy-proven diagnosis of IgAV. Patients were included on the basis of two criteria: i) severe involvement of at least one organ (including biopsy-proven IgAV-related nephritis class 3–4; gastrointestinal involvement with haemorrhage, ischaemia, perforation, and/or abdominal pain unresponsive to common analgesics and lasting for more than 24 h; pulmonary haemorrhage, episcleritis, cardiac and central nervous system involvement); and ii) exclusion of other systemic autoimmune or neoplastic diseases. Of 10 IgAV patients, 8 were sampled at diagnosis and 2 at disease flare; 4/10 patients had received prior treatment with oral prednisolone, 1/10 patients had been treated with azathioprine and 1/10 patients treated with cyclophosphamide within 6 months of sampling.

Cell separation, RNA extraction and antibody titres. For PBMCs and $CD19^+$ B cells: PBMCs were isolated from 110 ml of whole blood by centrifugation over Ficoll. $CD19^+$ B cells were isolated by positive selection using magnetic beads as previously described[37]. Total RNA was extracted from each sample using an RNeasy mini kit (Qiagen), quality was assessed using an Agilent BioAnalyser 2100 and RNA quantification was performed using a NanoDrop ND-1000 spectrophotometer.

For flow-sorted B cell samples[38]: flow sorting was performed using CD19-BV785, CD38-BV711, CD3-NC650, CD14-605NC, CD24-PerCP-Cy5.5, IgD-FITC. CD27-PE-Cy7 and Aqua (Invitrogen) (the flow cytometry protocol is outlined in Extended Data Fig. 1) into sorting buffer (10 mM Tris pH 8.0 and RiboLock RNase Inhibitor ($1\,U\,\mu l^{-1}$)) and frozen immediately.

Total IgA and IgE levels in patient serum were measured using a ProcartaPlex immunoassay kit (Thermo Fisher Scientific) using 25 μl of serum from each individual and run on a Luminex xMAP analyser. Raw data (MFI) were normalized to a concurrently measured 7-point standard curve according to the manufacturer's instructions to return an absolute quantification ($pg\,ml^{-1}$). All measured values were encompassed by the standard distribution.

Reverse transcription and amplification with barcoded primers. Reverse transcription was performed in a 23-μl reaction: 14 μl of reverse-transcription mix 1 (containing RNA template, 10 μM reverse primer mix, 1 μl dNTP (10 mM) and nuclease-free water) was incubated for 5 min at 70 °C. This mixture was immediately transferred to ice for 1 min, and the reverse-transcription mix 2 (4 μl 5× FS buffer, 1 μl DTT (0.1 M), 1 μl SuperScriptIII (Thermo Fisher Scientific)) was added and incubated at 50 °C for 60 min followed by 15 min inactivation at 70 °C. cDNA was cleaned with Agencourt AMPure XP beads and PCR-amplified with V-gene multiplex primer mix (10 μM each forward primer) and 3′ universal reverse primer (10 μM) using KAPA protocol and the following thermal cycling conditions: 1 cycle (95 °C, 5 min); 5 cycles (98 °C, 5 s; 72 °C, 2 min); 5 cycles (65 °C, 10 s; 72 °C, 2 min); 25 cycles (98 °C, 20 s; 60 °C, 1 min; 72 °C, 2 min); 1 step (72 °C, 10 min). Primers are provided in Supplementary Table 7.

Sequencing and barcode filtering. Sequencing libraries were prepared using Illumina protocols and sequenced using 300-bp paired-end sequencing on a MiSeq (Illumina). Raw reads were filtered for base quality (median Phred score of over 32) using QUASR (http://sourceforge.net/projects/quasr/)[39]. Forward and reverse reads were merged if they contained an identical overlapping region of more than 50 bp, or otherwise discarded. Universal barcoded regions were identified in reads and orientated to read from V-primer to constant-region primer. The barcoded region within each primer was identified and checked for conserved bases. Primers and constant regions were trimmed from each sequence, and sequences were retained only if there was over 80% per base sequence similarity between all sequences obtained with the same barcode; otherwise they were discarded. The constant-region allele with highest sequence similarity was identified by 10-mer matching to the reference constant-region genes from the IMGT database[40], and sequences were trimmed to give only the region of the sequence that corresponded to the variable (VDJ) regions. Isotype use information for each BCR was retained throughout the analysis hereafter. Sequences without complete reading frames and non-immunoglobulin sequences were removed and only reads with significant similarity to reference IGHV and J genes from the IMGT database using BLAST[41] were retained. Immunoglobulin gene use and sequence annotation were performed in IMGT V-QUEST, and repertoire differences were performed by custom scripts in Python.

Accounting for age in the BCR analysis. Age-related differences in the BCR repertoire have been previously described, and this could be important as immune-mediated diseases often have different ages of onset. We confirmed this in both healthy controls and disease, and then incorporated age as a covariate in repertoire analyses, as in previous studies[42–44].

As expected, correction for age usually made little difference (Extended Data Fig. 4a). In cases in which statistical discordance between uncorrected and corrected data did occur, the latter became not significant, indicating that this correction is appropriately conservative (that is, correction does not create spurious statistically significant positive associations). In these and most other cases (predominantly in diseases of later onset (AAV, EGPA)), age correction made *P* values less significant, which indicates that some observed repertoire differences are driven in part by age, and underlines the importance of correcting for it (Extended Data Fig. 3a–d). In some cases, results that were already significant became more so after correction; as expected, many of these were in cases of SLE or Crohn's disease, which are diseases with a younger age of onset (Extended Data Fig. 3c).

Isotype frequencies, somatic hypermutation, CDR3 lengths and IGHV gene use. To account for the greater numbers of BCR RNA molecules per plasmablast compared to other B cell subsets, we calculated two measures of isotype use: (i) the percentage of reads per isotype, which does not control for differential RNA per cell and thus reflects the effect that plasmablasts and plasma cells have on the repertoire; and (ii) the normalized isotype use, which is defined as the percentage of unique VDJ sequences per isotype, thus controlling for differential RNA per cell and reducing potential biases from differential RNA per cell. We did not control for ethnicity as the majority of patients (95%) in all disease groups were of northern European ancestry, with the exception of SLE in which four patients were Asian and five were Caucasian. We observed only two IGHV genes that exhibited differential frequencies between ethnicities with an FDR of less than 0.05 (Supplementary Table 6), and neither of these were differentially expressed between SLE and health.

Similarly, mean somatic hypermutation levels and CDR3 lengths were calculated per unique VDJ region sequence to reduce potential biases from differential RNA per cell. IGHV gene use was determined using IMGT[45], and proportions were calculated per unique VDJ region sequence. The representation of IGHV genes in the BCR repertoire reflects their presence in the germline, the naive repertoire and their expansion after antigenic exposure. We therefore compared the frequency of IGHV gene use in PBMC-derived BCRs identified by sequence as being enriched for naive B cells ($IgM^+IgD^+SHM^-$: more than 78% naive B cells by flow cytometry) and antigen-experienced B cells (including both unswitched ($IgM^+IgD^+SHM^+$) and class-switched memory ($IgA^+IgG^+IgE^+$) subsets).

Generation of the BCR repertoire and network analysis. The network generation algorithm and network properties were calculated as described previously[15]: each vertex represents a unique sequence, and the relative vertex size is proportional to the number of identical reads. Edges join vertices that differ by single nucleotide non-indel differences and clusters are collections of related, connected vertices. A clone (cluster) refers to a group of clonally related B cells, each containing BCRs with identical CDR3 regions and IGHV gene use, or differing by single point mutations, such as through SHM. Each cluster is assumed to arise from the same pre-B cell.

Repertoire parameters that were dependent on sequencing depth were generated by subsampling each sequencing sample to a specified depth. First, the clonal expansion index is a measure of the unevenness of the number of RNA molecules per unique VDJ region sequence by vertex Gini Index as defined previously[15]. This is calculated from the distribution of the number of unique RNA molecules per vertex within subsampled BCR repertoires at the specified depth defined below. The mean of 100 repeats of resulting clonal expansion indices was determined. Second, the clonal diversification index is a measure of the unevenness of unique VDJ region sequences per clone by cluster Renyi Index, as defined previously[15]. This is calculated from the distribution of the number of unique VDJ region sequences per clone within subsampled BCR repertoires at the specified depth defined below. The mean of 100 repeats of resulting clonal diversification indices was determined. Clone size distributions were also calculated from the same subsamples and the mean of 100 repeats was determined.

The number of sampled unique RNA molecules (for the clonal expansion index) and clones (for the clonal diversification index) per sample was: all isotypes, 3,500; IgD and IgM mutated, 600; IgD and IgM unmutated, 500; class-switched, 1,000; IgA1/2, 1,000; IgD, 75; IgE, 50; IgG1/2, 500; IgG3, 100; IgM, 750. These thresholds were chosen as a balance between including as many samples as possible per analysis but still remaining representative of the total BCR repertoire in each sample.

Sampling of the BCR network to preserve the overall clonal structure of visual representation. We developed network sampling methods to obtain a graphical representation of a network that preserves the overall clonal architecture. The rationale for this development, and for the selection of the clone-sampling method, is discussed in detail in the Supplementary Discussion. In brief, a fixed number of clones were subsampled, and a network was generated from all BCRs from these clones from a given sample. Subsampling was performed 1,000 times, and the sample that contained a maximum clone size closest to the median of all subsamples was chosen to generate a visual representation of the BCR repertoire.

Global measure of BCR repertoire. To define a global measure of the clonal normality or otherwise of the BCR repertoire, we combined three main BCR features (isotype frequency, clonal expansion index and clonal diversification index) using a multivariate ANOVA (MANOVA) comparison between disease groups with age as a covariate.

Analyses of class-switching events. The relative frequency of class-switching events was the frequency of unique VDJ regions expressed as two isotypes (that is, from more than one B cell, one of which has undergone CSR). This was determined as the proportion of unique BCRs present as both isotypes IgX and IgY within a random subsample of 8,000 BCRs, from which the mean of 1,000 repeats was generated (Extended Data Fig. 7e). This provides information on the frequency of BCRs observed to be associated with any two isotypes (class-switching events) and accounts for total read depth, but not for differences in the relative frequencies of BCRs per isotype.

The per-isotype normalized class-switch event frequency determines the frequency of unique VDJ regions expressed as two isotypes, normalizing for differences in isotype frequencies. To account for differences in isotype proportions, BCRs from each isotype were randomly subsampled to a fixed depth of 100 BCRs, and the proportion of unique VDJ sequences present between each pair of isotypes was counted (Extended Data Fig. 9a). The mean of 1,000 repeats was generated.

Clonal overlap between time points during therapy. The identification of persistent clones was performed using MRDARCY[46]. Clonal overlap frequencies between samples, including the quantification of persistent clones, were determined through subsampling each repertoire to a fixed depth of 2,000 unique BCRs and determining the proportion of overlapping clones. The mean of 1,000 repeats was generated.

Although quantitative conclusions are difficult as a blood draw samples such a small proportion of peripheral B cells, the clonal overlap estimate between time points is, as expected, significantly lower than that from technical BCR sequencing repeats from the same RNA samples and higher than the overlap between unrelated patient samples (Extended Data Fig. 10d).

Phylogenetic analysis. Phylogenetic trees from patients with AAV and EGPA were derived from all clusters containing at least one BCR sequence across multiple time points using the MRDARCY pipeline[46]. Alignments were performed using MAFFT[47] and maximum parsimony trees fitted using PAUP*[48]. The amino acid similarity tree for IGHV genes was generated using an alignment of reference IGHV genes from IMGT using MAFFT[47] and a maximum parsimony tree was fitted using PAUP*[48].

Statistical methods. Statistical differences between disease groups were performed using ANOVA or MANOVA with patient age as a covariate and correcting for multiple testing by Bonferroni correction. For cases in which patients were age-matched, Wilcoxon tests were performed.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Sequencing data are available from the EGA (accession numbers in Supplementary Table 3)

31. Stone, J. H. et al. A disease-specific activity index for Wegener's granulomatosis: modification of the Birmingham vasculitis activity score. *Arthritis Rheum.* **44**, 912–920 (2001).
32. Tan, E. M. et al. The 1982 revised criteria for the classification of systemic lupus erythematosus. *Arthritis Rheum.* **25**, 1271–1277 (1982).
33. Silverberg, M. S. et al. Toward an integrated clinical, molecular and serological classification of inflammatory bowel disease: report of a Working Party of the 2005 Montreal World Congress of Gastroenterology. *Can. J. Gastroenterol.* **19**, 5A–36A (2005).
34. Wechsler, W. E. et al. Mepolizumab or placebo for eosinophilic granulomatosis with polyangiitis. *N. Engl. J. Med.* **376**, 1921–1932 (2017).
35. Mills, J. A. et al. The American College of Rheumatology 1990 criteria for the classification of Henoch–Schönlein purpura. *Arthritis Rheum.* **33**, 1114–1121 (1990).
36. Jennette, J. C. et al. 2012 revised International Chapel Hill Consensus Conference Nomenclature of Vasculitides. *Arthritis Rheum.* **65**, 1–11 (2013).
37. Lyons, P. A. et al. Microarray analysis of human leucocyte subsets: the advantages of positive selection and rapid purification. *BMC Genomics* **8**, 64 (2007).
38. Espéli, M. et al. FcγRIIb differentially regulates pre-immune and germinal center B cell tolerance in mouse and human. *Nat. Commun.* **10**, 1970 (2019).
39. Watson, S. J. et al. Viral population analysis and minority-variant detection using short read next-generation sequencing. *Phil. Trans. R. Soc. Lond. B* **368**, 20120205 (2013).
40. Lefranc, M. P. IMGT unique numbering for the variable (V), constant (C), and groove (G) domains of IG, TR, MH, IgSF, and MhSF. *Cold Spring Harb. Protoc.* **2011**, 633–642 (2011).
41. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
42. Davydov, A. N. et al. Comparative analysis of B-cell receptor repertoires induced by live yellow fever vaccine in young and middle-age donors. *Front. Immunol.* **9**, 2309 (2018).
43. Marioni, R. E. et al. Genetic stratification to identify risk groups for Alzheimer's disease. *J. Alzheimers Dis.* **57**, 275–283 (2017).

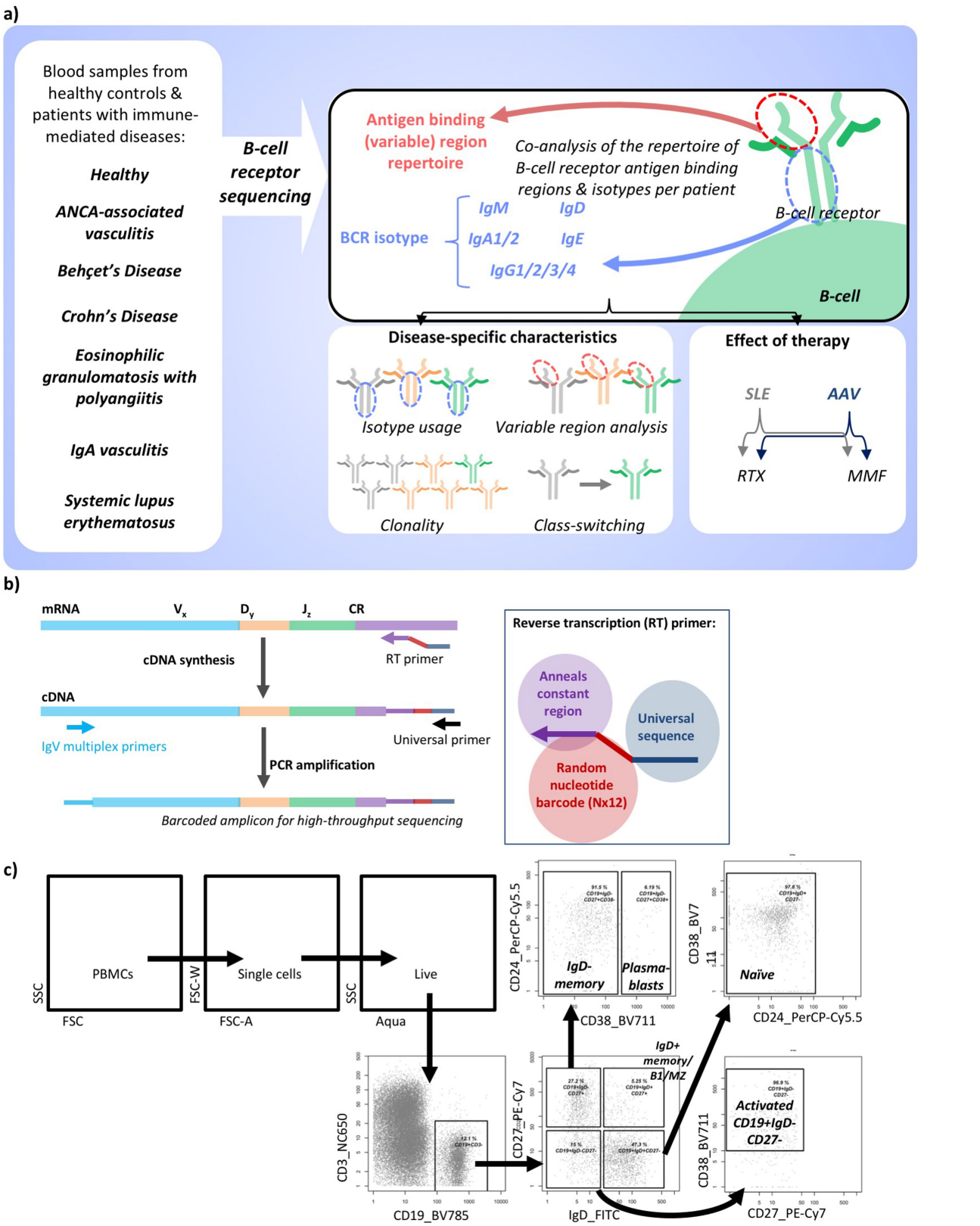
44. Ellis, J. A., Panagiotopoulos, S., Akdeniz, A., Jerums, G. & Harrap, S. B. Androgenic correlates of genetic variation in the gene encoding 5α-reductase type 1. *J. Hum. Genet.* **50**, 534–537 (2005).
45. Giudicelli, V., Chaume, D. & Lefranc, M. P. IMGT/V-QUEST, an integrated software program for immunoglobulin and T cell receptor V–J and V–D–J rearrangement analysis. *Nucleic Acids Res.* **32**, W435–W440 (2004).
46. Bashford-Rogers, R. J. et al. Eye on the B-ALL: B-cell receptor repertoires reveal persistence of numerous B-lymphoblastic leukemia subclones from diagnosis to relapse. *Leukemia* **30**, 2312–2321 (2016).
47. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
48. Wilgenbusch, J. C. & Swofford, D. Inferring evolutionary trees with PAUP*. *Current Protoc. Bioinformatics* **6**, 6.4.1–6.4.28 (2003).

Acknowledgements This work was supported by the Wellcome Trust (grants WT106068AIA and 083650/Z/07/Z), the EU H2020 project SYSCID (grant 733100), the UK Medical Research Council (program grant MR/L019027) and the UK National Institute of Health Research (NIHR) Cambridge Biomedical Research Centre. We thank the patients who participated in this study; V. Morrison; A. Reynolds; all NIHR Cambridge BioResource staff and volunteers; the Cambridge NIHR BRC Cell Phenotyping Hub (particularly A. Petrunkina Harrison, N. S.Yarkoni, E. Perez, S. McCallum and C. Bowman); F. Alberici, N. Noor and other members of the Addenbrooke's Vasculitis and Gastroenterology services; N. E. Urquijo for discussions about network subsampling, and P. Naydenova; and G. Manferrari. We are grateful to J. A. Todd and D. M. Tarlinton for reading the manuscript.

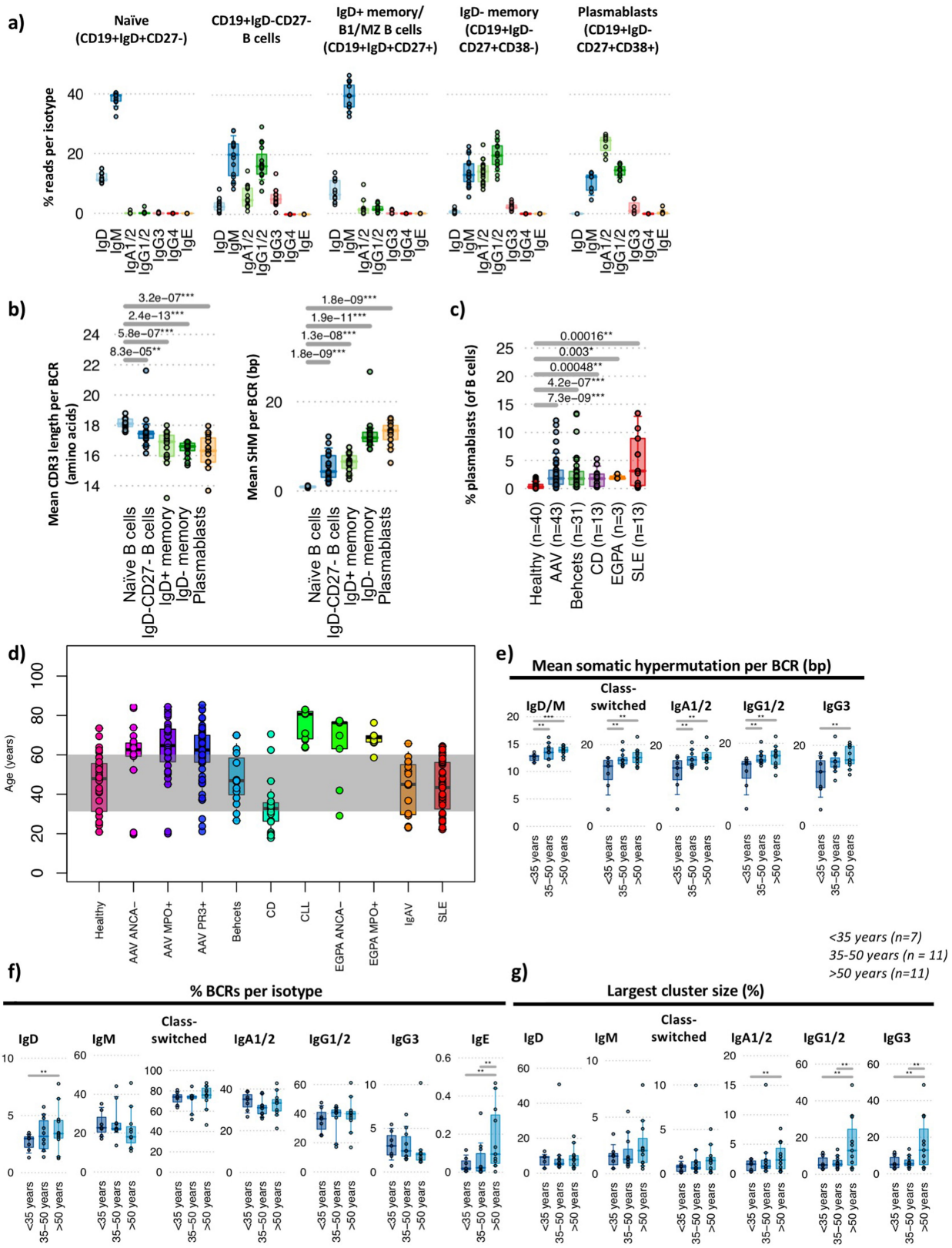
Author contributions R.J.M.B.-R. and K.G.C.S. planned the study. R.J.M.B.-R. performed BCR amplification and analysed sequencing data. F.M. analysed clinical data and L.B., D.C.P. and S.M.F. performed immunophenotyping. E.F.M., J.C.L., D.C.T., S.M.F., D.R.W.J. and P.A.L. contributed to sample collection and clinical data generation, and P.K. contributed to sample processing. R.J.M.B.-R., P.A.L., E.F.M., J.C.L., D.C.T. and K.G.C.S. provided intellectual contributions to analyses. R.J.M.B.-R. and K.G.C.S. wrote the manuscript. All authors edited the manuscript.

Competing interests R.J.M.B.-R., P.K. and K.G.C.S. are all named on a patent associated with the methodologies in this paper. S.M.F. is a current employee of GlaxoSmithKline, and holds shares in GlaxoSmithKline. R.J.M.B.-R. is a consultant for Imperial College London and VHSquared. P.K. is an employee and holder of shares in Kymab Ltd. D.R.W.J. is a recipient of a research grant from Roche and Genentech. K.G.C.S. is a co-founder of Rheos Medicines and K.G.C.S., P.A.L. and E.F.M. are co-founders of PredictImmune.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1595-3.
Correspondence and requests for materials should be addressed to R.J.M.B. or K.G.C.S.
Peer review information *Nature* thanks Felix Breden, George Georgiou and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.



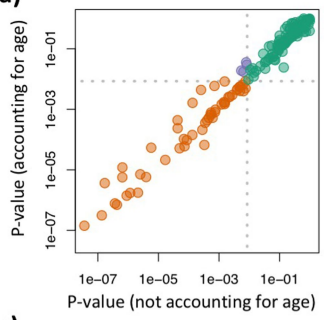
Extended Data Fig. 1 | Overview of BCR repertoire strategy. **a**, Schematic diagram of the strategy for BCR repertoire analysis. **b**, Schematic diagram of the BCR sequencing strategy. In the reverse transcription (RT) step, the primer anneals to the constant region of the BCR mRNA to generate cDNA with a random 12-nucleotide barcode. This barcode can be computationally used to reduce PCR amplification biases after sequencing. The product is cleaned and PCR-amplified using multiple primers that bind to the FR1 region of the IgH genes along with a universal sequence complementary to the end of the reverse-transcription primer. **c**, Gating strategy for sorting B cell subsets from healthy donor PBMCs by flow cytometry.



Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | Effects of B cell subset and age on the BCR repertoire. a, Frequencies of isotype use from BCR sequencing data from sorted naive B cells ($CD19^+IgD^+CD27^-$); $CD19^+CD27^-IgD^-$ B cells; IgD^+ memory, B1 and marginal-zone B cells ($CD19^+CD27^+IgD^+$); IgD^- memory B cells ($CD19^+CD27^+IgD^-CD38^-$); and plasmablasts ($CD19^+CD27^+IgD^-CD38^+$) from 19 healthy individuals. **b**, Mean CDR3 lengths (left) and mean SHM per BCR (right) from cell-sorted B cell populations from healthy individuals ($n = 19$). **c**, Plasmablast frequency per patient in peripheral blood at enrolment as a percentage of $CD19^+$ B cells. **d**, Distribution of patient ages within this study, grouped by disease. **e**–**g**, Correlations of the BCR repertoire in PBMCs with age in healthy individuals for the mean number of somatic hypermutations per BCR per bp (**e**), the percentage of BCRs per isotype (**f**) and percentage size of the largest cluster per sample (**g**). For **b**, **c**, *P* values were calculated by two-way ANOVA; *FDR < 0.05, **FDR < 0.005, ***FDR < 0.0005 (determined by the Šidák method). For **e**–**g**, *P* values were calculated by two-sided Wilcoxon test; *$P < 0.05$, **$P < 0.005$, ***$P < 0.0005$, all other values not significant. Raw *P* values are provided in Supplementary Table 4. For all box plots, box lines show the 25th, 50th and 75th percentiles; whiskers show the upper and lower quartiles.

a)



b)

Comparison	Isotype	Group 1	Group 2	Mean of group 1	Mean of group 2	P-value (not accounting for age)	P-value (accounting for age)	N in group 1	N in group 2
Clonal diversification index	Class-switched	Healthy	EGPA	0.0528	0.0881	0.007895	0.035615	32	11
Clonal diversification index	IgA1/2	Healthy	EGPA	0.0446	0.0792	0.008248	0.029550	31	11
Clonal expansion index	All BCRs	Healthy	EGPA	0.2035	0.3108	0.006401	0.017446	31	11
Clonal expansion index	IgD	Healthy	CD	0.0184	0.0501	0.005427	0.018978	32	24

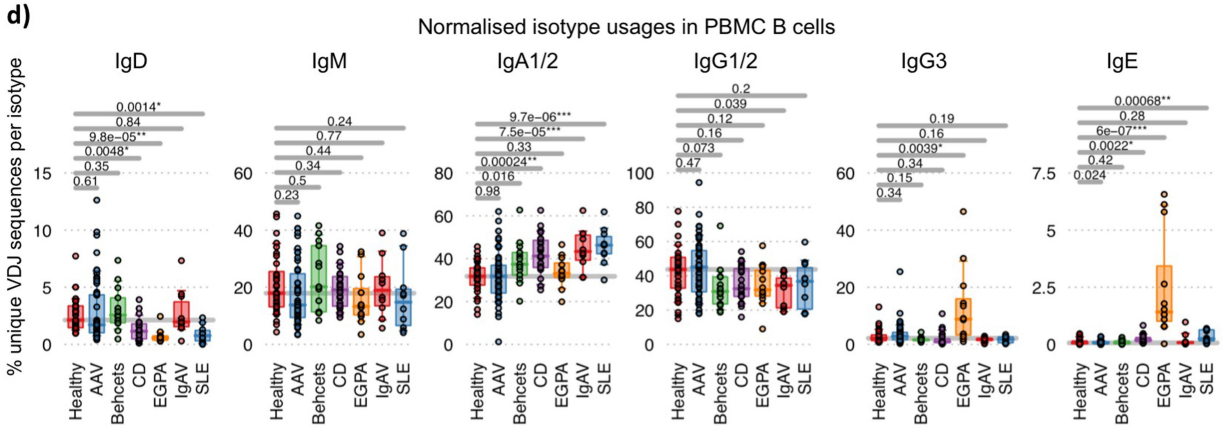
*significant threshold = 0.008512 after multiple testing correction

c)

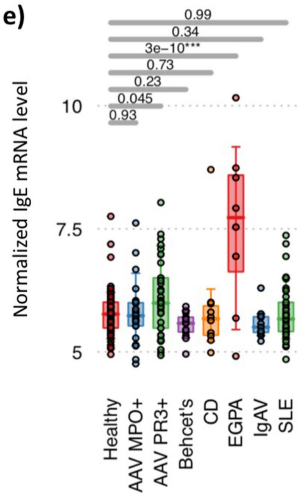
Comparison	Isotype	Group 1	Group 2	Mean of group 1	Mean of group 2	P-value (not accounting for age)	P-value (accounting for age)	N in group 1	N in group 2
Clonal diversification index	IgD/M unmutated	Healthy	SLE	0.05332	0.18753	2.11E-06	1.76E-06	31	9
Clonal diversification index	IgE	Healthy	EGPA	0.07334	0.17618	6.43E-03	3.97E-03	7	11
Clonal diversification index	IgM	Healthy	Behcets	0.10271	0.03552	1.54E-03	1.49E-03	32	13
Clonal expansion index	Class switched	Healthy	SLE	0.05282	0.12648	6.17E-04	5.23E-04	32	10
Clonal expansion index	IgD/M unmutated	Healthy	SLE	0.00938	0.03805	7.26E-05	6.00E-05	31	9
Clonal expansion index	IgG1/2	Healthy	SLE	0.06068	0.12238	4.92E-03	4.06E-03	32	10
Isotype usage (%)	IgD	Healthy	EGPA	3.37707	1.01856	2.81E-04	2.10E-04	32	11
Isotype usage (%)	IgE	Healthy	CD	0.13917	0.32650	3.28E-04	6.66E-05	32	24

*significant threshold = 0.008512 after multiple testing correction

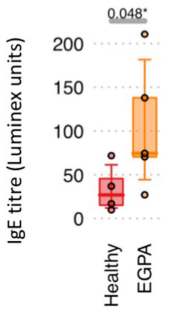
d)



e)

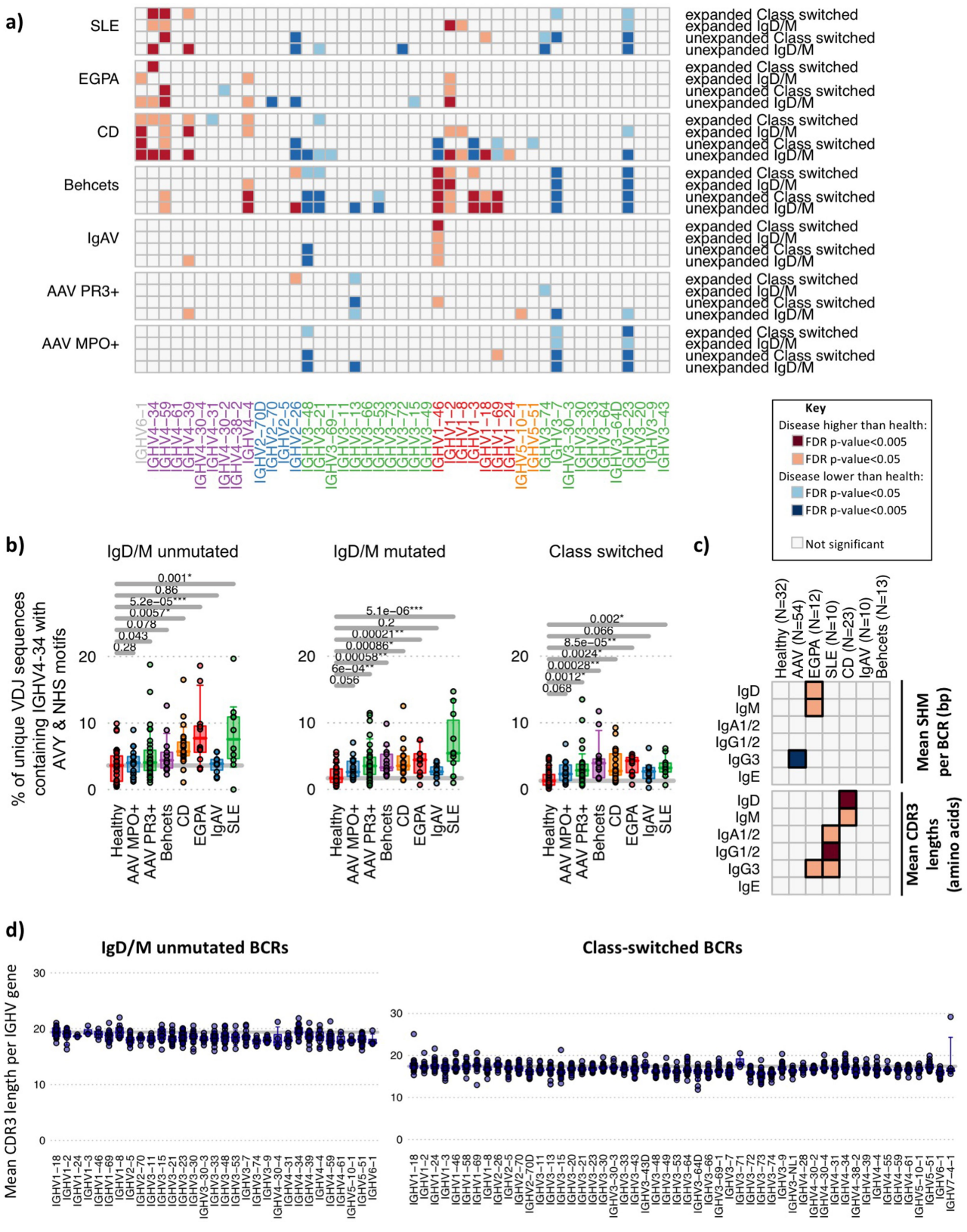


f)



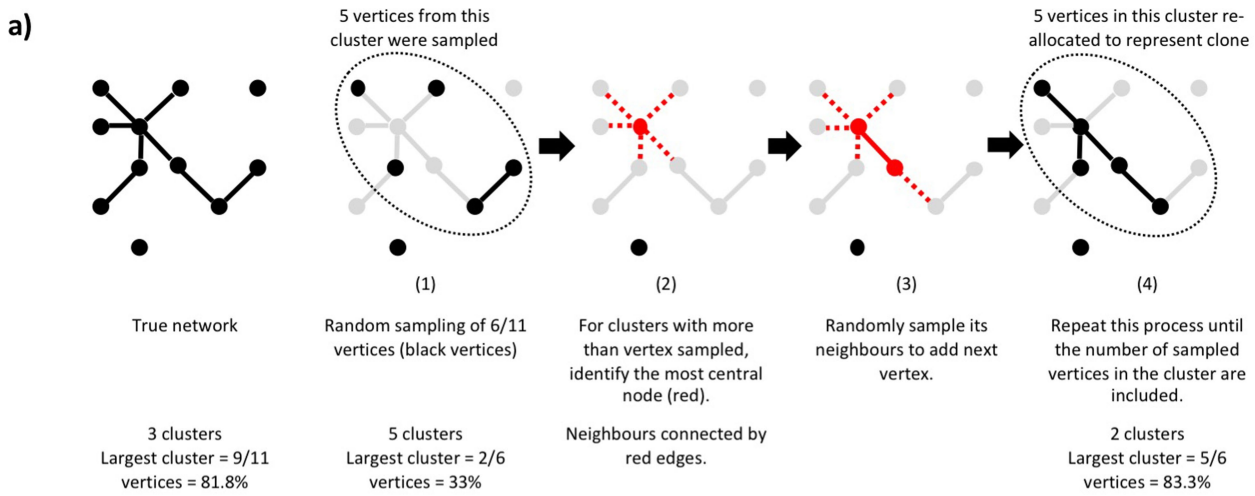
Extended Data Fig. 3 | See next page for caption.

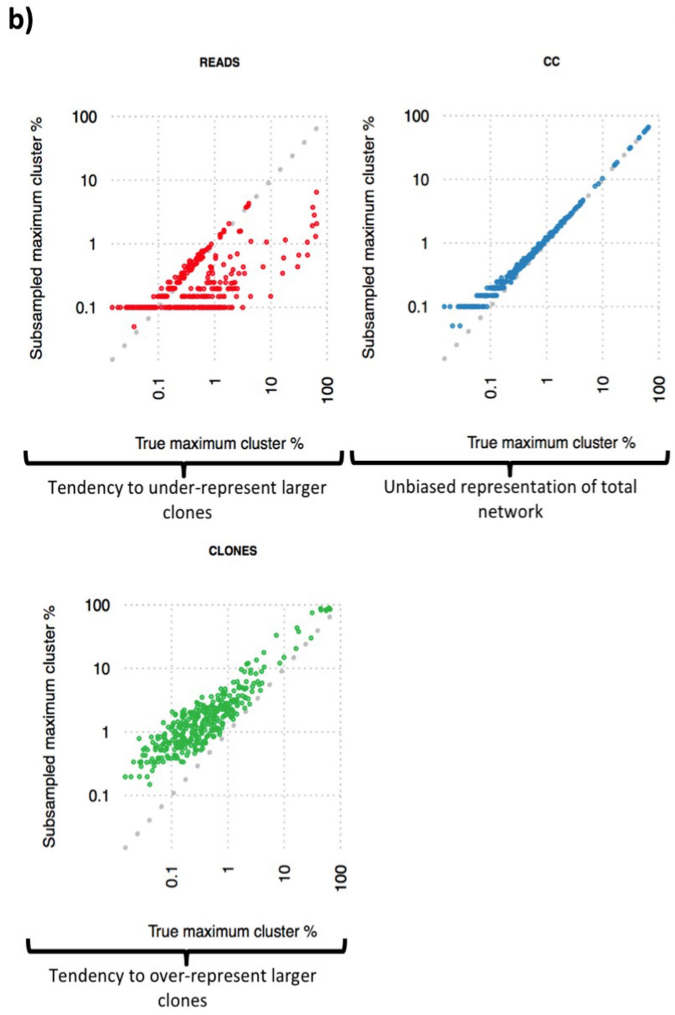
Extended Data Fig. 3 | Changes in the BCR repertoire with age and changes in isotype use with disease. a, Correlation of *P* values obtained using age as a covariate versus those obtained when age was excluded from the analysis across 178 BCR features (calculated by two-way ANOVA). Grey dotted lines indicate the threshold of significance after accounting for multiple testing (FDR < 0.05, determined by the Šidák method). **b**, Analyses in which statistical significance was discordant (that is, below the threshold for significance without accounting for age and above the threshold when age was used as a covariate, or vice versa; purple points in **a**). **c**, Analyses in which statistically significant *P* values were decreased by more than 1.5-fold after using age as a covariate. **d**, Percentage of normalized isotype use (unique VDJ sequences per isotype) for BCR repertoires in PBMCs per disease. **e**, Normalized transcript levels of the IgE immunoglobulin constant region between disease groups, from transcriptomic data. $n = 58$ for healthy individuals and $n = 23$, $n = 33$, $n = 13$, $n = 10$, $n = 8$, $n = 11$ and $n = 37$ for patients with AAV (MPO^+), AAV ($PR3^+$), Behçet's disease, Crohn's disease, EGPA, IgAV and SLE, respectively. **f**, IgE titre between healthy individuals ($n = 4$) and patients with EGPA ($n = 5$). For **d**, **e**, *P* values were calculated by two-way ANOVA; *FDR < 0.05, **FDR < 0.005, ***FDR < 0.0005 (determined by the Šidák method). For all box plots, box lines show the 25th, 50th and 75th percentiles; whiskers show the upper and lower quartiles.

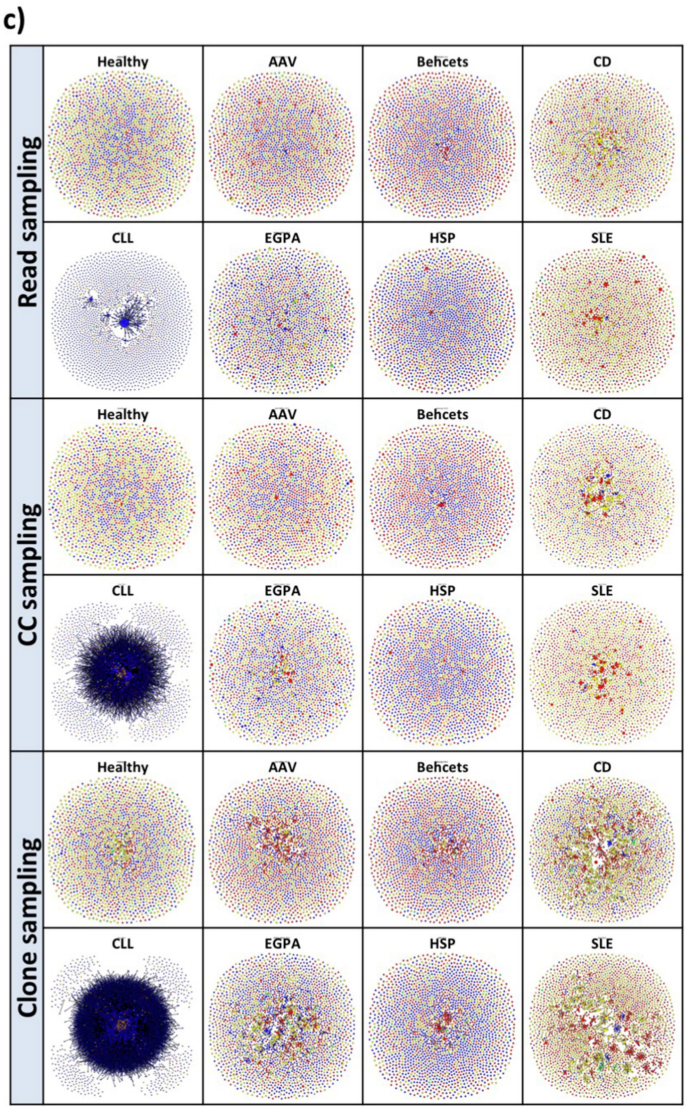


Extended Data Fig. 4 | See next page for caption.

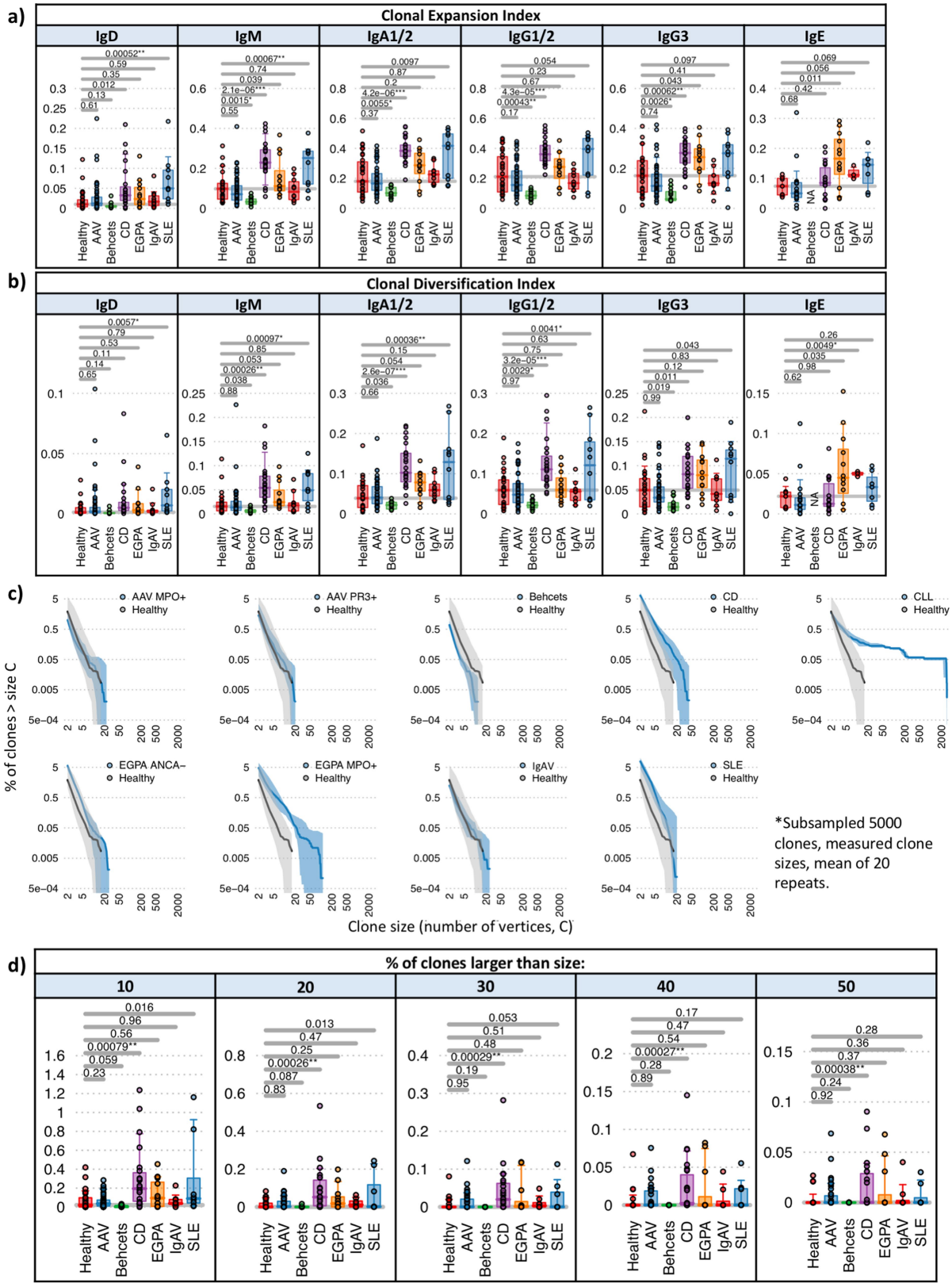
Extended Data Fig. 4 | Changes in IGHV gene use with disease. **a**, Changes in IGHV gene use between unexpanded and expanded clones. Heat map of the difference in the frequency of each IGHV gene between health and disease within BCRs from IgM^+IgD^+ or isotype-switched (IgA1, IgA2, IgG1, IgG2, IgG3, IgG4 or IgE) unexpanded clones (containing fewer than three unique BCRs) or expanded clones (containing three or more unique BCRs per clone). Only genes that occurred at a higher frequency than 0.1% are shown. IGHV genes are ordered according to amino acid similarity, as in Fig. 2. **b**, Frequencies of *IGHV4-34* BCRs with autoreactive AVY and NHS motifs compared between healthy individuals and disease groups, separated by BCR type: $IgM^+IgD^+SHM^-$, $IgM^+IgD^+SHM^+$ and IgM^-IgD^- BCR sequences (defined in **a**). **c**, Heat maps showing the mean SHM per BCR (top) and relative mean CDR3 lengths (bottom) per isotype per disease from total PBMC B cells. **d**, Distribution of the mean CDR3 lengths per IGHV gene in healthy individuals ($n = 32$). Each point represents the mean CDR3 length for an individual for unmutated IgD or IgM BCRs (left) and class-switched BCRs (right). Instances in which IGHV genes were represented by fewer than 10 BCRs in an individual are excluded. For **a**–**d**, $n = 32$ for healthy individuals and $n = 20$, $n = 34$, $n = 12$, $n = 10$, $n = 23$, $n = 10$ and $n = 13$ for patients with AAV (MPO^+), AAV ($PR3^+$), EGPA, SLE, Crohn's disease, IgAV and Behçet's disease, respectively. *P* values were calculated by two-way ANOVA. Orange squares indicate significantly higher, and blue squares significantly lower, corresponding gene frequency between healthy individuals and disease. FDR was determined by the Šidák method.





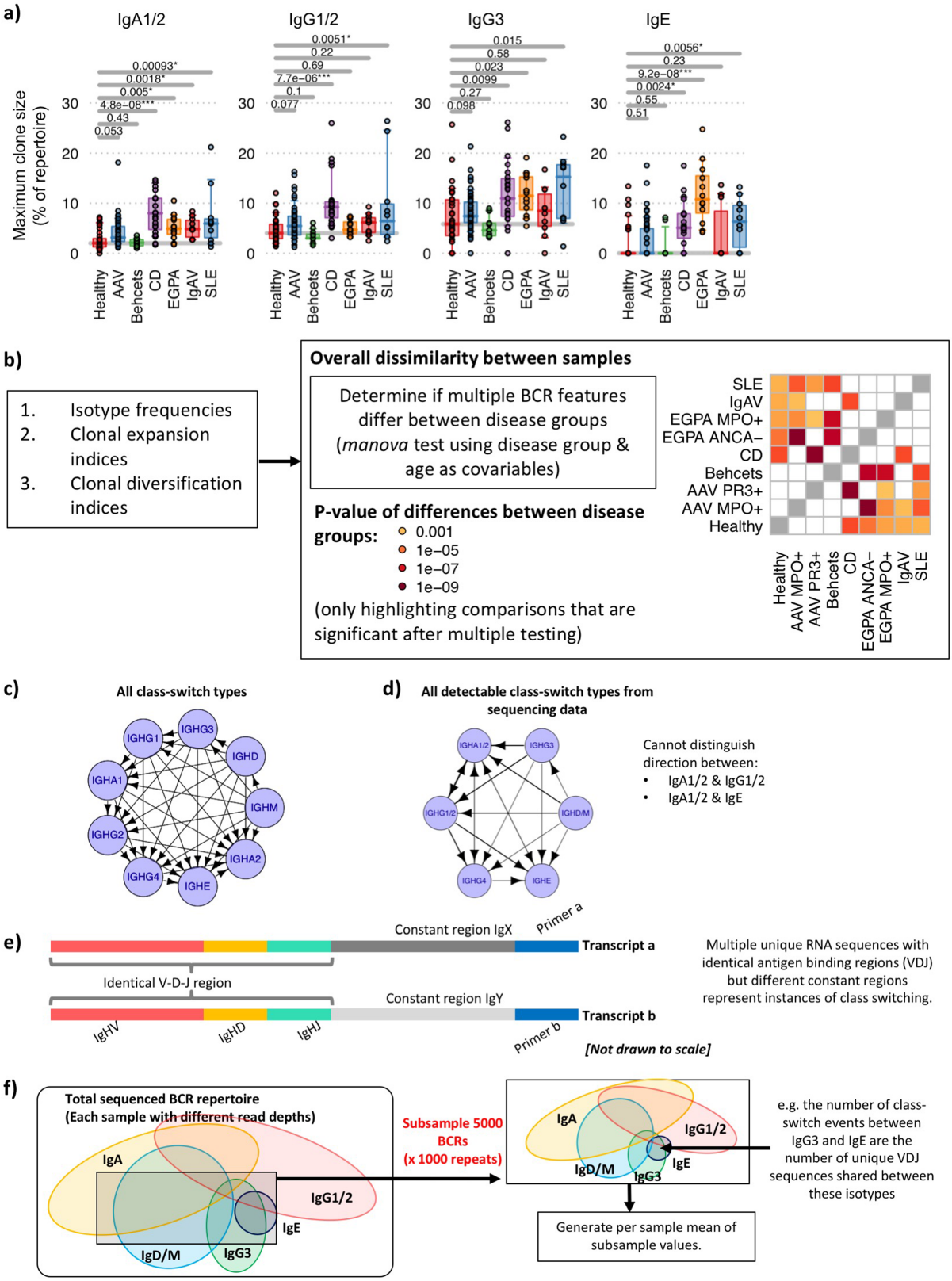


Extended Data Fig. 5 | Network subsampling methods for preserving repertoire structure. a, Schematic diagram of the cluster-vertex migration in the CC algorithm. **b**, Maximum cluster sizes between true (unsampled) networks and subsampled networks of 2,000 clones by the tree subsampling methods. **c**, Comparison of representative networks from each group of patients at diagnosis. The patient samples are represented across the three sampling methods. Each vertex represents a unique sequence, and the relative vertex size is proportional to the number of identical reads. Edges join vertices that differ by single nucleotide non-indel differences and clusters are collections of related, connected vertices. Networks comprise a subsample of 2,000 clones, using the corresponding subsampling method. Each vertex is represented by a pie chart that indicates the percentage of each isotype, in which blue represents IgD or IgM, red IgA1 or IgA2, yellow IgG1 or IgG2, green IgG3 and grey IgE.



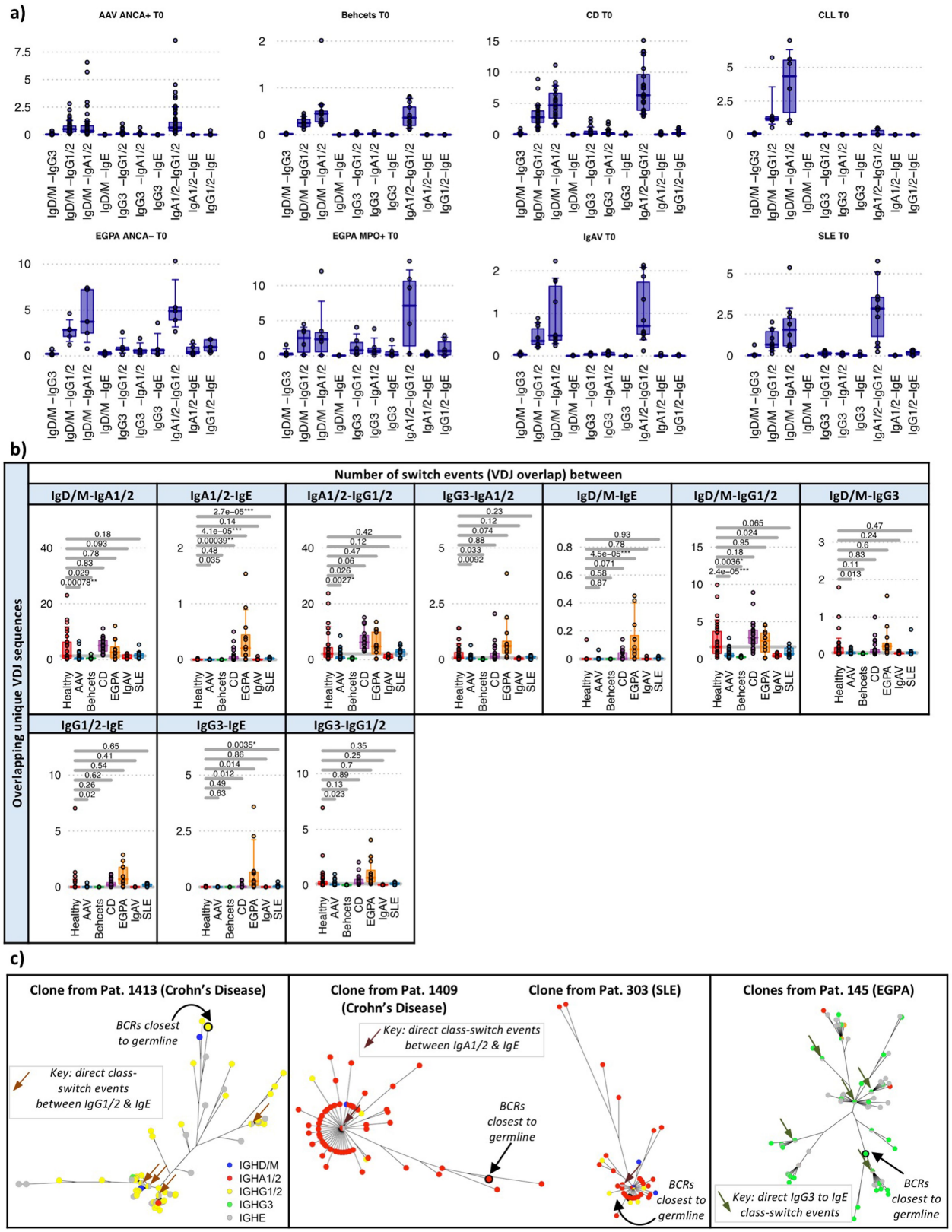
Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | BCR repertoire clonality between diseases. **a**, **b**, Box plots of the clonal expansion index (**a**) and the clonal diversification index (**b**) for BCR repertoires in PBMCs per disease. **c**, Plots of the percentage of clones per sample per disease that are greater than clone size (C). Clone size is defined as the number of unique VDJ sequences that are clonally related. For each disease, the mean percentage is indicated by the dark blue line, and the upper and lower interquartile ranges by the light blue areas. Overlaid in grey is the equivalent for healthy individuals. Differences in read depth were accounted for by subsampling 5,000 clones from each repertoire and determining the mean of 20 repeats. As a disease comparison, we show the distribution for CLL. **d**, Box plots of the percentage of clones that have more than 10, 20, 30, 40 or 50 unique VDJs per disease. Differences in read depth were accounted for by subsampling 5,000 clones and determining the mean of 20 repeats. For **a**, **b**, **d**, *P* values were calculated by two-way ANOVA; *FDR < 0.05, **FDR < 0.005, ***FDR < 0.0005 (determined by the Šidák method). $n = 32$ for healthy individuals and $n = 20$, $n = 34$, $n = 12$, $n = 10$, $n = 23$, $n = 10$ and $n = 13$ for patients with AAV (MPO^+), AAV ($PR3^+$), EGPA, SLE, Crohn's disease, IgAV and Behçet's disease, respectively. For all box plots, box lines show the 25th, 50th and 75th percentiles; whiskers show the upper and lower quartiles.

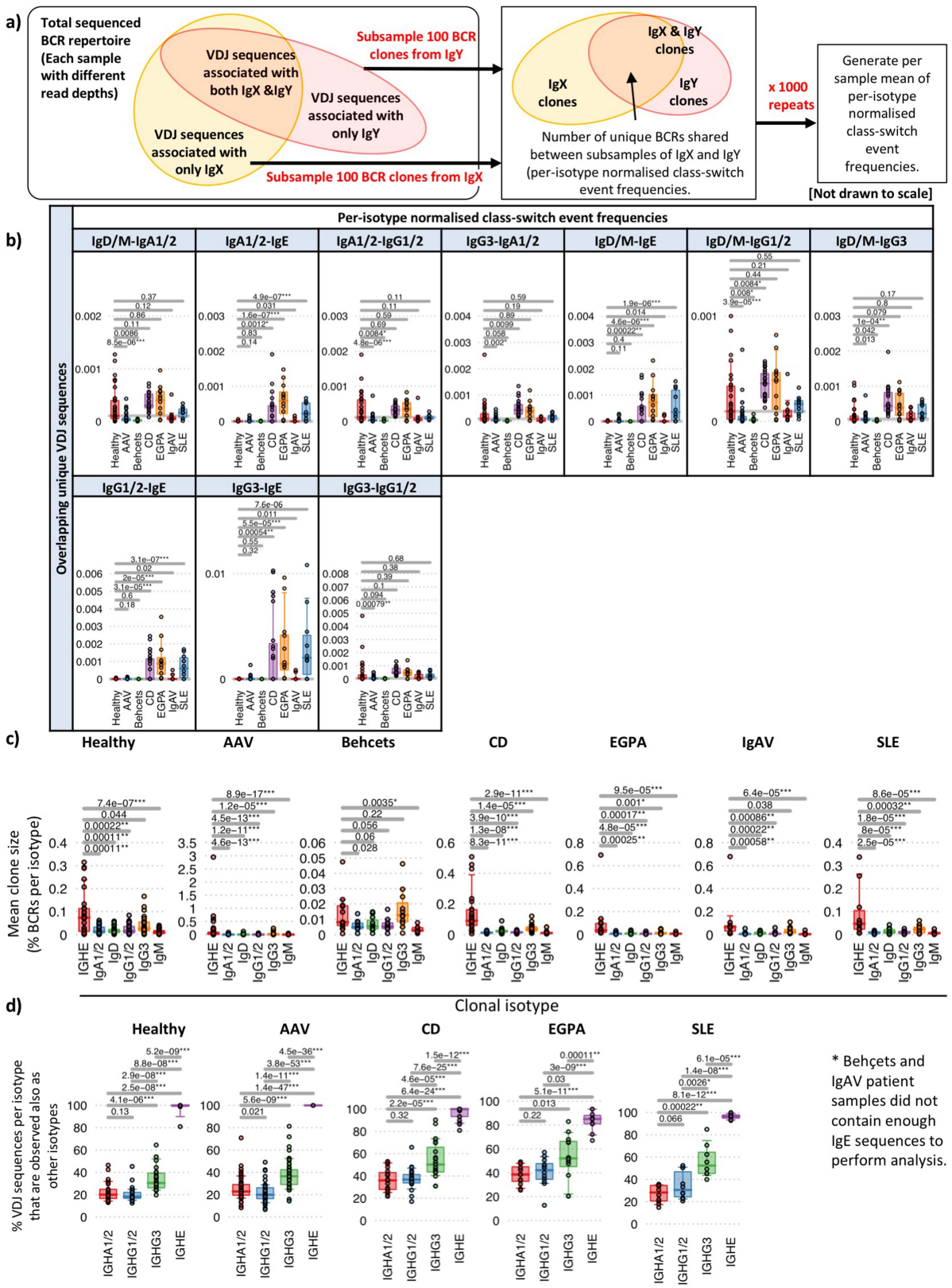


Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | BCR repertoire similarity between diseases and estimation of CSR. a, The maximum clone sizes (as a percentage of unique VDJ sequences of a given isotype in the largest clone divided by the total number of unique BCRs of that isotype) for BCR repertoires from PBMCs per disease across isotypes. For box plots, box lines show the 25th, 50th and 75th percentiles; whiskers show the upper and lower quartiles. **b**, Global repertoire dissimilarity measures between disease groups. Heat map showing the global repertoire dissimilarity measures between disease groups on the basis of a combination of three main BCR features (isotype frequency, clonal expansion index and clonal diversification index) and determining joint differences between groups (MANOVA test using disease group and age as covariables). The light and dark orange squares indicate significant differences between corresponding disease groups (FDR < 0.05 and FDR < 0.005, respectively, determined by Šidák method). $n = 32$ for healthy individuals and $n = 20$, $n = 34$, $n = 12$, $n = 10$, $n = 23$, $n = 10$ and $n = 13$ for patients with AAV (MPO^+), AAV ($PR3^+$), EGPA, SLE, Crohn's disease, IgAV and Behçet's disease, respectively. **c**, The sequence of B cell isotype expression is defined by the order of constant regions on the chromosome. Possible class-switching events are depicted by the arrows between constant regions. **d**, Schematic diagram of class-switching types that are detectable from the sequencing data. The differences from **c** are a result of the ambiguity of isotype between IgA1 and IgA2, and IgG1 and IgG2, in the isotype-specific sequencing, and splicing of IgD from IgM-containing transcripts. Possible class-switching events are represented by the arrows between constant regions. **e**, Several unique RNA sequences with identical antigen-binding regions (V–D–J) but different constant regions represent instances of class switching. **f**, Schematic diagram of the subsampling of BCR repertoires to generate the relative class-switch event frequency. This is the frequency of unique VDJ regions that are expressed as two isotypes (that is, from more than one B cell, one of which has undergone CSR), and determined as the proportion of unique BCRs that are present as both isotypes (IgX and IgY) within a random subsample of 8,000 BCRs, from which the mean of 1,000 repeats was generated. This provides information on the frequency of BCRs that are observed to be associated with any two isotypes (class-switching events), and accounts for total read depth, but not for differences in the relative frequencies of BCRs per isotype. For **a**, $n = 32$ for healthy individuals and $n = 54$, $n = 12$, $n = 10$, $n = 23$, $n = 10$ and $n = 13$ for patients with AAV, EGPA, SLE, Crohn's disease, IgAV and Behçet's disease, respectively. *P* values were calculated by two-way ANOVA, *FDR < 0.05, **FDR < 0.005, ***FDR < 0.0005 (determined by the Šidák method).

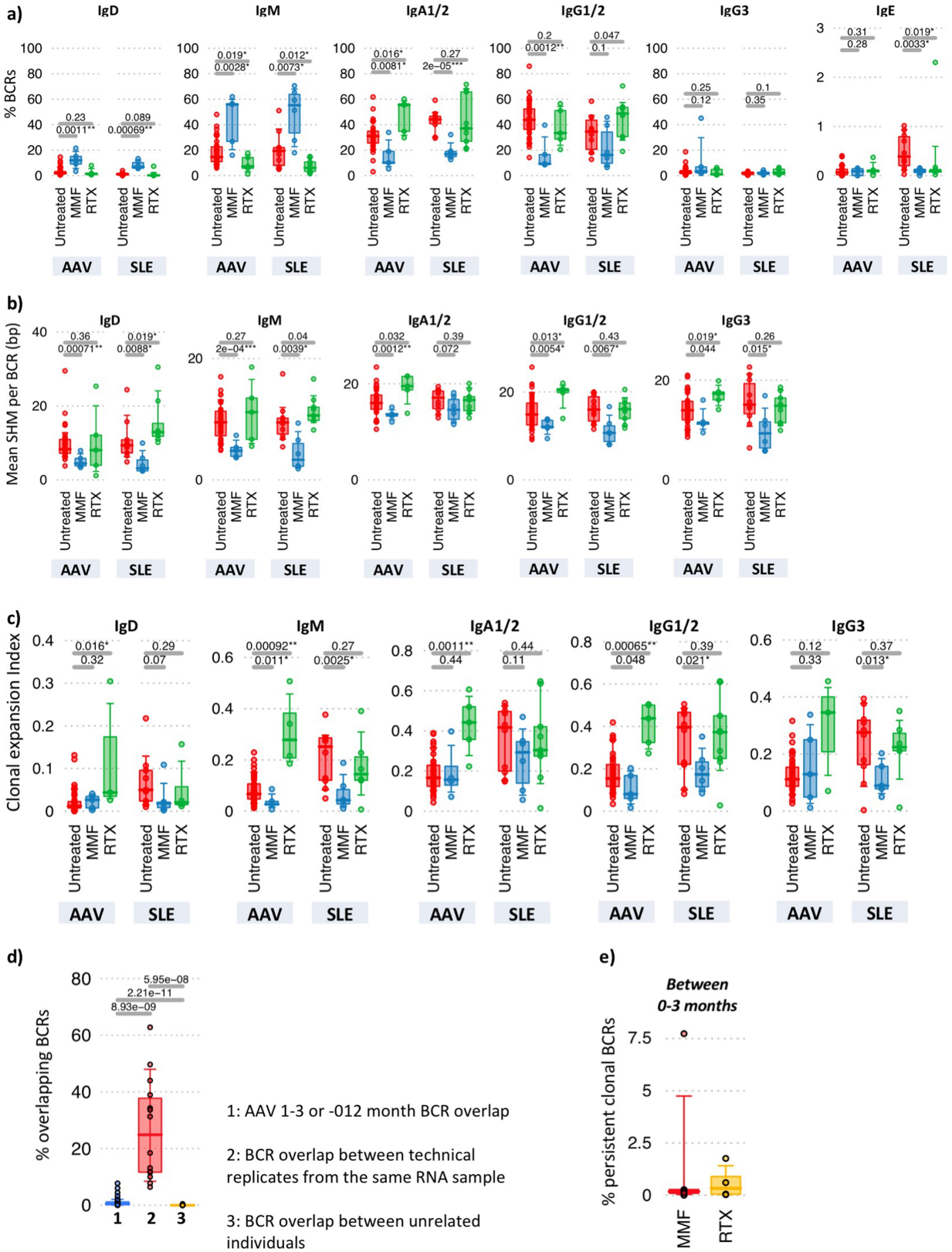


Extended Data Fig. 8 | Differences in CSR estimation between diseases. **a**, Box plots of the proportion of class-switching events between isotypes for each autoimmune disease. **b**, Box plots of the proportion of class-switching events between autoimmune diseases across isotypes for BCR repertoires in PBMCs by subsampling the total repertoire. *P* values were calculated by two-way ANOVA; *FDR < 0.05, **FDR < 0.005, ***FDR < 0.0005 (determined by the Šidák method). $n = 32$ for healthy individuals and $n = 54$, $n = 12$, $n = 10$, $n = 23$, $n = 10$ and $n = 13$ for patients with AAV, EGPA, SLE, Crohn's disease, IgAV and Behçet's disease, respectively. For all box plots, box lines show the 25th, 50th and 75th percentiles; whiskers show the upper and lower quartiles. **c**, Phylogenetic trees of representative clonal expansions of B cells from patients, demonstrating CSR events. Each vertex is represented by a pie chart that indicates the percentage of each isotype, in which blue represents IgD or IgM, red IgA1 or IgA2, yellow IgG1 or IgG2, green IgG3 and grey IgE. Branch lengths are estimated by maximum parsimony, and the BCRs with the lowest number of somatic hypermutations are indicated (denoted 'BCRs closest to germline').



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Normalized differences in CSR estimation between diseases and IgE clonal features. a, Schematic diagram of subsampling of BCR repertoires to generate the per-isotype normalized class-switch event frequencies (defined as the frequency of unique VDJ regions expressed as two isotypes, normalizing for differences in isotype frequencies). To account for differences in the proportions of the isotypes, BCRs from each isotype were randomly subsampled to a fixed depth of 200 reads, and the proportion of unique VDJ sequences present between each pair of isotypes was counted. The mean of 1,000 repeats was generated. **b**, Box plots of the proportion of the per-isotype normalized class-switch event frequencies between isotypes for each autoimmune disease. *P* values were calculated by two-way ANOVA; *FDR < 0.05, **FDR < 0.005, ***FDR < 0.0005 (determined by the Šidák method). $n = 32$ for healthy individuals and $n = 54$, $n = 12$, $n = 10$, $n = 23$, $n = 10$ and $n = 13$ for patients with AAV, EGPA, SLE, Crohn's disease, IgAV and Behçet's disease, respectively. **c**, Box plots of the mean cluster sizes per patient per isotype as a percentage of BCRs per isotype, comparing IgE-associated clones with non-IgE-associated clones for each disease. **d**, The proportion of VDJ sequences per isotype that are observed also as other isotypes for each disease. *P* values were calculated by two-sided Wilcoxon test; *FDR < 0.05, **FDR < 0.005, ***FDR < 0.0005 (determined by the Šidák method). $n = 32$ for healthy individuals and $n = 54$, $n = 12$, $n = 10$, $n = 23$, $n = 10$ and $n = 13$ for patients with AAV (MPO^+), AAV ($PR3^+$), EGPA, SLE, Crohn's disease, IgAV and Behçet's disease, respectively. For all box plots, box lines show the 25th, 50th and 75th percentiles; whiskers show the upper and lower quartiles.



Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | Effects of therapy on the BCR repertoire. **a**–**c**, Percentage of BCRs per isotype (**a**), mean SHM per BCR per isotype (**b**) and clonal expansion indices (**c**) of samples that were taken from patients with AAV and SLE at diagnosis (red, untreated), and after 3 months of induction therapy with MMF (blue) or RTX (green). For AAV, the patients per group were: untreated, $n = 42$; MMF, $n = 5$; RTX, $n = 5$; and for SLE: untreated, $n = 11$; MMF, $n = 6$; RTX, $n = 9$. **d**, Percentage of BCRs shared between samples that were taken from patients with AAV at diagnosis and 3 or 12 months after induction therapy (blue); BCRs shared between repertoires from the same RNA tube (red); and BCRs shared between unrelated patient samples. Zero overlap was found between unrelated samples, whereas there was a significantly higher overlap between BCRs shared between repertoires from the same RNA tube compared to BCRs shared between AAV samples taken at diagnosis and 3 or 12 months after induction therapy. This suggests that the overlap measurements yield realistic and normalized values at this sampling depth. **e**, The percentages of persistent BCRs shared between diagnosis and 3 months after induction therapy, split between patients who received different therapies. *P* values were calculated by two-sided Wilcoxon test. For all box plots, box lines show the 25th, 50th and 75th percentiles; whiskers show the upper and lower quartiles.

nature research

Corresponding author(s): Bashford-Rogers

Last updated by author(s): 19/07/19

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	NA
Data analysis	QUASR (Watson, Welkers et al. 2013) http://sourceforge.net/projects/quasr/ BLAST (Altschul, Gish et al. 1990) https://blast.ncbi.nlm.nih.gov/Blast.cgi IMGT V-QUEST (Lefranc 2011) http://www.imgt.org/HighV-QUEST/ Immune_receptor_NETWORK-GENERATION (Bashford-Rogers, R. J. et al 2013) https://github.com/rbr1/Immune_receptor_NETWORK-GENERATION R R version 3.3.3 (2017-03-06) https://www.r-project.org R package: ape version 5.0 CRAN https://cran.r-project.org R package: stringr version 1.2.0 CRAN https://cran.r-project.org R package: igraph version 1.1.2 CRAN https://cran.r-project.org R package: phangorn version 2.3.1 CRAN https://cran.r-project.org

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Sequencing data available from the EGA (accession numbers in Table S3).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample sizes were determined such each disease group contained >=8 patients, which we have previously shown to be sufficient to distinguish during active immune responses, such as in early HIV infection (Hoehn et al 2015).
Data exclusions	Samples with <5000 unique BCRs from the sequencing were excluded.
Replication	We have previously shown that technical replicates using these methods are highly correlated (Petrova et al 2018) and therefore replicates were not performed on an individual patient basis.
Randomization	NA. Patients selected on the basis of no/limited prior treatment.
Blinding	During sample processing and sequencing.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☐	☒ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	c BioLegend Cat#302240 Anti-human CD38 BV711 BioLegend Cat#303528 Anti-Human CD3 eVolve™ 655 eBiosciences (Thermo Fisher Scientific) Cat#86-0037-42 Anti-Human CD14 eVolve™ 605 eBiosciences (Thermo Fisher Scientific) Cat#83-0149-42 PerCP-Cy™5.5 Mouse Anti-Human CD24 BD Biosciences Cat#561647 FITC Mouse Anti-human IgD BD Pharmigen Cat#555778 CD27 (PE-Cy7) - 100 tests eBiosciences (Thermo Fisher Scientific) Cat#25-0279-42
Validation	See manufacturers' notes.

Human research participants

Policy information about studies involving human research participants

Population characteristics	Age, current diagnosis and (where applicable) treatment.
Recruitment	Healthy participants Inclusion criteria for healthy individuals were people aged between 20-77 years, with no serious co-morbidities, no direct family history of autoimmune disease, no use of immunosuppressants or steroids, and no hospitalization within the last 12 months. The healthy individual samples used for B cell sorting were collected through the NIHR Cambridge BioResource. Patients with AAV AAV patients attending or referred to the specialist vasculitis unit at Addenbrooke's Hospital, Cambridge, UK, between July 2004

and June 2016 were enrolled into the present study. Active disease at presentation was defined by at least 1 major or 3 minor Birmingham Vasculitis Activity Score (BVAS) criteria and the clinical impression that induction immunosuppression would be required. Prospective disease monitoring was undertaken monthly with serial BVAS assessment 47 and serum ANCA status (Supplemental Item 1). 41/54 patients were sampled at diagnosis and 13/54 patients at disease flare as defined above. A minority of patients (11/54) had received prior treatment with oral prednisolone, and 3 patients had received Azathioprine within 6 months prior to sampling. Patients on low dose steroids and azathioprine have been analysed separately, and their inclusion does not impact upon any of the findings described in this study.

Patients with SLE
The SLE cohort comprised patients attending or referred to the Addenbrooke's Hospital specialist vasculitis unit between July 2004 and June 2016 who met at least four American College of Rheumatology SLE criteria, presenting with active disease. Active disease was defined as meeting all three of the following prospectively defined criteria: new British Isles Lupus Assessment Group (BILAG) score A or B in any system, clinical assessment of active disease by the reviewing physician and increase in immunosuppressive therapy as a result. After treatment with an immunosuppressant, patients were followed up monthly. Disease monitoring was undertaken with serial BILAG assessment and serum ANA status. Patients' treatment was at the physician's discretion, not dictated by study participation and includes therapy used for induction of remission at enrolment ('induction'). 8/10 patients were sampled at diagnosis and 2/10 patients at disease flare. A minority of patients (3/10) had received prior treatment with oral prednisolone and/or hydroxychloroquine.

Patients with CD
Patients with active Crohn's disease were recruited from a specialist IBD clinic at Addenbrooke's Hospital, before starting treatment. 22/23 patients were recruited at the time of diagnosis. Diagnosis was made using standard endoscopic, histological and radiological criteria. All patients had at least moderately active Crohn's disease at enrolment as evidenced by clinical symptoms in conjunction with some or all of elevated C-reactive protein, elevated fecal calprotectin, radiologically active disease or endoscopically active disease. All patients who were treatment naïve, with none receiving immunomodulators, corticosteroids or biological therapy.

Patients with CLL
Patients with CLL were recruited from the specialist leukemia/lymphoma unit at Addenbrooke's Hospital unit between January 2011 and July 2014. CLL patient inclusion required the presence of at least 5×109 B cells/L circulating clonal B cells persisting for 3 months and a characteristic phenotype (typically CD5, CD19, CD20, and CD23).

Patiens with EGPA
EGPA patients attending or referred to the specialist vasculitis unit at Addenbrooke's Hospital, Cambridge, UK, between July 2004 and June 2016 were enrolled into the present study. EGPA diagnosis was based on the history or presence of both asthma and eosinophilia (>1.0 x 109/L and/or > 10% of leukocytes) plus at least two additional features of EGPA, critria used in the recent Phase III clinical trial "Study to Investigate Mepolizumab in the Treatment of Eosinophilic Granulomatosis With Polyangiitis". 7/11 patients were sampled at diagnosis and 4/11 patients at disease flare. A minority of patients (4/11) had received prior treatment with oral steroids (methylprednisolone or prednisolone), 2/11 patients treated with azathioprine and 1/11 patients treated with cyclophosphamide within 6 months of sampling.

Patients with IgAV and Behçet's Disease
IgAV patients and Behçet's disease patients were recruited from the specialist vasculitis clinic at Addenbrooke's Hospital were enrolled between into the present study between 2005 and 2015. Clinical data recorded for Behçets disease patients comprised: (i) Basis for diagnosis i.e. orogenital mucosal ulceration, prior ocular inflammation, and characteristic skin rash (erythema nodosum or pseudofolliculitis); (2) Major complications such as venous or arterial thrombosis, central nervous system involvement or involvement of the pulmonary vascular system; and (iii) disease activity (expert physician global assessment). 5/11 of patients had received prior treatment with oral steroids (prednisolone) and 3/11 patients had been treated with azathioprine within 6 months prior to sampling.

The diagnosis of IgAV was based on the American College of Rheumatology 1990 criteria for the classification of Henoch-Schönlein purpura and the 2012 Revised International Chapel Hill Consensus Conference Nomenclature of Vasculitides. All patients had to have a biopsy-proven diagnosis of IgAV. Patient inclusion was based on if they had i) severe involvement of at least 1 organ (including biopsy-proven IgAV-related nephritis class 3–4; gastrointestinal involvement with haemorrhage, ischemia, perforation, and/or abdominal pain unresponsive to common analgesics and lasting for >24 hours; pulmonary haemorrhage, episcleritis, cardiac and central nervous system involvement); and ii) other systemic autoimmune or neoplastic diseases were excluded. 8/10 patients were sampled at diagnosis and 8/10 patients at disease flare. 4/10 of patients had received prior treatment with oral prednisolone, 1/10 patients treated with azathioprine and 1/10 patients treated with cyclophosphamide within 6 months of sampling.

Ethics oversight

Ethical approval for this study was obtained from the Cambridge Local Research Ethics Committee (reference numbers 04/023, 08/H0306/21, 08/H0308/176) and Eastern NHS Multi Research Ethics Committee (07/MRE05/44), with informed consent obtained from all subjects enrolled.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☐ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	For PBMCs and CD19+ B cells: PBMCs were isolated from 110 ml of whole blood by centrifugation over Ficoll. CD19+ B cells were isolated by positive selection using magnetic beads as previously described. Total RNA was extracted from each sample using an RNeasy mini kit (Qiagen) with quality assessed using an Agilent BioAnalyser 2100 and RNA quantification performed using a NanoDrop ND-1000 spectrophotometer. For flow-sorted B cell samples: Flow sorting was performed using CD19-BV785, CD38-BV711, CD3-NC650, CD14-605NC, CD24-PerCP-Cy5.5, IgD-FITC. CD27-PE-Cy7 and Aqua (Invitrogen), where flow protocol is outlined in Figure S1b, into sorting buffer (10mM Tris pH 8.0 and RiboLock RNase Inhibitor (1U/µL)) and frozen immediately.
Instrument	BD Influx lasers/colours 4/16 jet-in-air
Software	FlowJo and R
Cell population abundance	Post-sort purity check (>95%), as well as through assessment of isotype usages within the sorted populations from the sequencing data.
Gating strategy	In Supplemental figure 1

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.