SeQuoIA: a single-cell BCR sequencing analysis pipeline for tracking selection mechanisms in germinal centres

Eglantine Hector[1,‡] , Pierre Milpied[1,‡]

[1] Aix Marseille University, CNRS, Inserm, Centre d'Immunologie de Marseille-Luminy (CIML), Marseille, France
[‡] Corresponding authors: Eglantine Hector (hector@ciml.univ-mrs.fr), Pierre Milpied (milpied@ciml.univ-mrs.fr)

Abstract

Germinal centres (GCs) are specialized structures where B cells undergo iterative steps of B-cell receptor (BCR) somatic hypermutation and selection of best antigen binders in a darwinian-like fashion. The accelerated evolutionary process leads to the production of high-affinity antibodies that are crucial for robust and long-term humoral immunity. Within this frame, single-cell BCR sequencing analysis is a method of choice to track GC B cell dynamics as somatic mutations can be utilised as an *in vivo* molecular tracer. Herein, we present SeQuoIA, a start-to-finish pipeline for the analysis of BCR repertoire sequencing data at the single-cell level, including improved clonotype assignment and phylogeny reconstruction. Most importantly, we introduce a new method for the inference of BCR-driven selection pressure based on somatic mutation patterns, that was validated with biological data. With this pipeline, we explored public datasets and proposed new selection mechanisms in GCs.

Significance: Our pipeline should contribute to a better understanding of the basic biology of GC dynamics, and potentially help in laboratory animal usage reduction. Clinical applications could include assessment of vaccine efficacy, monitoring of B cell anti-tumoral responses, and identification of BCR-mediated processes in B cell lymphomas.

Key words:
BCR, AIRR, Germinal centre, Phylogeny, Modeling, Plasma cell, scRNA-seq

Short running title:
Inference and exploration of BCR-based selection in germinal centres

Introduction

B cells are major actors of the adaptive immune response against pathogens. These cells are armed with immunoglobulins (Ig), in the form of receptor (BCR) or secreted antibodies, which can bind foreign antigens with high affinity and specificity. These features are acquired through successive diversification and selection steps in the bone marrow (BM) and secondary lymphoid organs (SLO). VDJ recombination during early B cell development generates combinatorial diversity of BCR repertoires. Upon infection or immunisation, circulating mature B cells can undergo additional development steps in specialised micro-anatomic niches called germinal centres (GCs). GC entry is determined by B cell ability to bind the antigen (Ag) and receive T cell signals. This bottleneck provides a first clonal selection step. Within GCs, B cells migrate between two compartments in the cyclic process of affinity maturation (AM). They diversify their BCR through somatic hypermutation (SHM) in the dark zone (DZ). Point mutations are introduced by activation cytidine deaminase (AID) enzyme. The binding properties of newly generated receptors are tested in the light zone (LZ), where B cells compete for antigen access and those harbouring adequate BCR affinity are selected in a darwinian-like process. After several rounds of diversification and selection, B cells may eventually exit GCs and differentiate into memory B cells (MBCs) or plasma cells (PCs).

How Ag binding properties translate into increased fitness is not fully understood. B cell selection has been mostly scrutinised through the lens of antigen presentation to follicular helper T cells (T_{FH}), which provide cues (CD40L, IL4, IL21) involved in B cell proliferation and differentiation. Within this framework, GC B cells failing to receive T_{FH} cues are assumed to die by neglect[1]. The role of other signals is yet to be fully explored. Fate decisions (differentiation, death and recycling) and the signals that govern them is still an area of active research. In most experimental model systems, it has been shown that antibody-producing PC differentiated preferentially from GC B cells expressing BCR with high affinity for the antigen, although in polyclonal responses to complex antigens broad ranges of BCR affinities may end up in the GC-derived PC compartment[2,3]. Within this context, the extent to which PC differentiation relies on BCR affinity has not been settled.

Single cell datasets including gene expression and Ig sequence data constitute an insightful approach for the exploration of GC reaction (GCR) selection mechanisms, as they link clonal and mutation information to a transcriptional state characteristic of cell subset or recent signal transduction. Many tools have already been developed for processing Ig sequencing (IgSeq) datasets, identifying clonal families and tracing Ig sequence evolution through phylogenetic tree reconstruction. For an exhaustive list of such tools as well as typical repertoire analyses, the reader is referred to several reviews[4,5,6]. Despite the large amount of softwares dedicated to IgSeq analysis, few attempts have been made to infer BCR-driven selection pressure from SHM patterns[7]. Tools integrating gene expression, Ig heavy and light chain information are lacking, although high-throughput integrative single cell sequencing made strides over the past few years.

We developed a bioinformatic pipeline for comprehensive BCR repertoire analysis at the single cell level: **Se**lection **Qu**antification in **I**ntegrative **A**IRR data (SeQuoIA). SeQuoIA achieved improved clonotype assignment and phylogenetic reconstruction accuracy over state-of-the-art existing tools. Most importantly, *SeQuoIA* uniquely leverages BCR clonal tree knowledge to quantify selection based on Ig mutation patterns. Finally, our pipeline links BCR features to transcriptional states. In this article, we applied SeQuoIA to publicly available datasets and we show that the light chain is a valuable source of information, both for phylogeny reconstruction and tracking selection. We also demonstrate that SeQuoIA is a powerful predictive tool to generate new hypotheses on BCR-driven selection mechanisms, in the LZ but also subsequently in the DZ cells or after commitment to PC differentiation.

Material and Methods

SeQuoIA workflow

The complete pipeline is summarised in Fig.1 and pipeline code description can be found in Supp.Table 1.

Sequencing data pre-processing and annotation

Sequencing error correction, read assembly and read quality filtering are expected to be done by the user prior to SeQuoIA. The SeQuoIA pipeline takes as input assembled contigs in fasta format and an annotation file linking Ig contig information to cell barcodes. These inputs are typically generated with CellRanger V(D)J but could easily be obtained with other BCR sequencing data preprocessing methods. Additional but highly recommended input includes cell information from GEX data analysis (e.g. metadata from Seurat object). Input formatting requirements are detailed in Supp.Table 2.

Contigs are aligned onto a reference genome (mouse or human) and annotated with the *Immcantation* suite[8], including V, D, J gene segment usage and functional annotation (framework regions FRW and complement determining regions CDR). Contig filtering is based on alignment quality (non-missing VJ portions) and cellular data (contig with higher UMI counts per chain is kept). All intermediate outputs are in Adaptive Immune Receptor Repertoire (AIRR) format[9].

Clonotype assignation

Clonally related sequences descending from the same B cell are identified with classical similarity approaches. Similarity based grouping on sequences sharing the same V and J allele is performed for each chain with a hierarchical clustering on CDR3 aa sequences hamming distance matrix, and a 80% similarity cutoff is applied. The added value of SeQuoIA as compared to most clonal assignment tools is the pairing of heavy and light chain information through single-cell ID (10x barcode or well position). Heavy and light chain clones are combined into single consistent clones. Cells missing an Ig chain can be incorporated into clonal groups defined with both chain information if the corresponding chain meets the above-mentioned criteria. Sequences with indels are usually filtered out from clonal families at this stage as frameshifts lead to dramatic changes in CDR3 aa sequence. Clonotype assignment can be performed at the dataset level or for each donor independently, if donor information is provided by the user. The donor-restricted approach allows computation time to be reduced and increases the likelihood that clonal cells are indeed related, even in the case of “public” clones or murine genetically engineered B cells harbouring the same V(D)J combination.

Ancestral sequence reconstruction

For each clonotype, a clone ancestor is inferred. This ancestor corresponds either to the germline (*e.g.* naive B cell derived clones) or the Nearest Common Ancestor (NCA), a virtual or sampled sequence from which all sampled sequences should derive. The NCA is relevant in memory B cell derived clones or clonal families that are advanced in the affinity maturation process.

At this stage, germline sequences lack junction information, which cannot be inferred from genome alignment (*cf.* pre-processing section). Junction sequences are thus filled with a 50% consensus of least mutated contigs (first two ranks).

NCA reconstruction is based on a strict consensus of least mutated sequences for each nucleotide. Germline nucleotides are taken in positions lacking consensus. Ancestor sequence productivity is ensured by SeQuoIA.

Phylogeny reconstruction

Ancestral relationships within clonal families are inferred with a maximum parsimony based phylogenetic reconstruction of heavy and light chains. Trees are rooted either to the NCA or the germline (if the distance to the germline in non-junction sequence is inferior to 3). Identical Ig sequences are made non redundant with *--deduplicate* utility from *GcTree*[10] (v4.0.0), which also summarises the information of unique Ig sequences abundance. Phylip *dnapars* is called to generate a set of most parsimonious trees, henceforth termed the "Parsimony Forest". Non parsimonious trees are derived from the parsimonious ones through a prune and regraft (SPR) approach where nodes can be moved upstream of the trees. One pair of heavy and light chain trees is selected within the extended forest (featuring parsimonious and SPR-derived trees). Tree selection is informed by several criteria based on current knowledge of affinity maturation biology which are, by order of priority:

- Mutation load and reversions: reverse mutations are unexpected except in pathological generation of bystander and autoreactive cells as evidenced in BCR sequencing at different timepoints in WT and ALPS mouse models[11]. More generally, trees with least mutation 'aberrations' (reversion, decreased mutation load along tree progression axis) are pre-selected with SeQuoIA.
- Parent productivity: DZ B cells were shown to have a quick BCR turnover and to undergo apoptosis before reaching the LZ if a deleterious mutation led to a destabilisation of the BCR complex[12]. As a result, cells expressing non-productive sequences should not be able to have a descendance. SeQuoIA takes into account the need for tonic BCR signalling to persist in GCs by filtering out trees with non-productive (*i.e.* with a stop codon or a frameshift) internal nodes, whether they are observed or inferred.
- CSR parsimony: this criterion takes into consideration the isotypes of observed nodes and minimises the number of independent CSR events along the tree. CSR parsimony is already implemented in the *GcTree* package (*isotype_parsimony* function).
- Chain concordance: this criterion is based on paired heavy and light chain data recovered in scRNAseq. Heavy and light chain sequences expressed within the same cell should have the same lineage in both heavy and light chain trees. However, checking the concordance of lineages is not trivial due to an unbalanced evolution speed between both chains. As a result, heavy chains tend to have many divergent sequences matched to the same light chain sequence. SeQuoIA traces back each node with paired data in heavy and light chain trees and checks whether:
    - nodes have shared parents in both trees. If not, unshared parent could correspond to sequences derived from the same grandparent where heavy and light chains mutations were non concomitant. SeQuoIA thus extends the search to grandparents in heavy and light chain trees. Otherwise, the chain pairs are considered as not directly related in heavy and light chain trees. The tree pair combination is excluded from the exploration space for subsequent filtering criteria.
    - Paired sequences are in the same order from ancestor to terminal leaves. If the order is not conserved, the tree pair is excluded from the exploration space.

  SeQuoIA then ranks the remaining trees so as to choose the tree pairs that maximise the overlap in chain combination lineages. The degree of overlap is quantified with a modified jaccard index normalised by the total number of tree pairs:

$$\frac{|\,heavy\ parents\ \cap\ light\ parents\,| \times |heavy\ parents| \times |light\ parents|}{|total\ combinations|}$$

  The chain concordance-based selection process is illustrated in Supp.Fig.1.
- CSR order: CSR is an irreversible genomic rearrangement involving the excision of exons encoding alternative Ig classes. CSR proceeds from upstream to downstream exons along the IGH constant region locus in the following order: IgD, IgM, IgG3, IgG1, IgG2b, IgG2c, IgE, IgA in mouse, and IgD, IgM, IgG3, IgG1, IgA1, IgG2, IgG4, IgE, IgA2 in human. The number of breaches from this order is

counted by SeQuoIA and used to rank trees that passed previous filters. Since most CSR are thought to occur prior to GC entry [13], these CSR events should not occur during affinity maturation. However, they may occur during precursor activation and possibly during differentiation into effector B cells.

- GcTree likelihood: the *GcTree* tool incorporates node abundance (*i.e.* identical Ig sequence frequencies) to rank trees in terms of likelihood within the parsimony forest. This criterion is based on the premise that the more an Ig sequence is represented, the more likely it is to have a descendance. Most expanded nodes should thus be located upstream of the tree axis.
- Mutability Parsimony: Equally parsimonious trees could also be ranked according to the likelihood of SHM mutability patterns, as implemented in SAMM[14] and later incorporated in the *GcTree* pipeline. However, clonotypes subject to selection should deviate from the basal AID mutability patterns. This criterion is thus used as the last filter.

Mutation Simulation

Parent sequences (i.e. tree internal nodes) are used to initialise several independent models recapitulating point mutation accumulation in the absence of selection. Inferred internal nodes can be taken if the selected tree has a non-ambiguous topology. Mutations are introduced in parent sequences until the mutation level of its daughters is reached, and taking into account SHM intrinsic bias which has been characterised for hundreds of 5-mer nucleotide patterns[15]. In practice, these mutations are applied with the shmulateSeq function of the *SHazaM* package, with either previously published human targeting model[15] or our own mouse targeting model depending on the species configuration parameter. SeQuoIA applies mutations sequentially to update mutability patterns and ensures sequence productivity at each step. Newly *in silico* generated sequences are translated to check whether the mutation is synonymous or not. The mutation steps are repeated in 2000 replicates and mutation indices are gathered to estimate conditional probabilities of a site being mutated:

$$Pn = P(\ n\ is\ mutated \mid N, parent) = \frac{\#sequences\ where\ n\ is\ mutatated\ with\ N\ mutations}{\#replicates}$$

where n is the nucleotide index, $parent$ is the parental sequence used for initialization, and *N* the total (synonymous and non-synonymous) number of mutations accumulated from this parent. *N* is low in trees where intermediate states could be captured. In the default mode, the pipeline does not assume that only non-synonymous (NS) mutations affect BCR fitness and all mutations are considered. SeQuoIA generates figures as QCs to compare *Pn* in well described SHM hotspot (WRC) and coldspot (SYC) (Supp.Fig.2).

Selection Quantification

Experimental sequences mutation analysis includes:

- Mutation load: since parent and ancestor
- Mutation localization: nucleotide or codon index in RNA or protein sequences respectively
- Mutation effect: synonymous or non-synonymous

Theoretical mutation probabilities estimated in the previous step are used to compute selection scores. A codon selection score summarises the degree of expectancy of mutation or non-mutation in mutated and conserved codons respectively:

$$Score_{codon} = \frac{\delta_i}{1 + \delta_i \times \frac{Pi - Pbasal}{Pbasal}}$$

where

- δ_i = 1 if codon i is affected by a NS mutation, δ_i = 0 otherwise
- P_i the probability that at least one of the nucleotides in codon i is non-synonymously mutated. $P_i = \sum_{n=1}^{3} Pn$
- *Pbasal* is the mean over all Pi within the sequence. This normalisation factor allows to visualise the sites of appreciable positive and negative selection along a BCR sequence in a single plot.

Intuitively, the presence of a mutation although it was unlikely to occur according to SHM model (P_i> Pbasal) may be an indicator of positive selection ($Score_{codon}$ > 1). Conversely, unmutated sites in mutation hotspot may be an indicator of negative selection (($Score_{codon}$ < -1).

We use Kulback-Leibler (KL) divergence between observed mutations and the theoretical mutabilities to estimate positive selection imprint on the whole Ig sequence. The divergence of observed mutation from the model is normalised by the entropy to compare sequences with different ancestries and mutation loads. Sequence scores account either for instantaneous or cumulated selection imprints thanks to the phylogeny knowledge.

$$Score_{instantaneous} = \frac{\sum_{i=1}^{Sequence\ length} \lim_{Di \to \delta_i} D_i \times log(\frac{P_i}{D_i})}{\sum_{i=1}^{Sequence\ length} P_i \times log(P_i)}$$

$$Score_{cumulative} = \sum_{t=1}^{T} \frac{Score_{instantaneous}(S)}{T}$$

with *T* the number of tree layers until node S.

By default, unless stated otherwise, selection scores will refer to the instantaneous selection score in this manuscript.

Result aggregation and analysis

Experimental and transcriptomic data is incorporated into the analysis to compare selection between groups of cells or along continuous variables (tree pseudo-time, distance to germline and clone size are analysed by default). Details on generated outputs and script details can be found in Supp.Table 1. This report only provides basic analyses. The user is strongly encouraged to check tree quality (as in Supp.Fig.3) and perform gene expression UMI-based filtering prior to custom analysis.

Interface with IGOr

SeQuoIA can incorporate IGOr[16] to estimate VDJ combination probabilities of each clonotype. Examples of applications are provided in Supp.Fig.4. Generation probabilities (Pgen) can serve as a proxy of clonal selection (as in pre-GC steps), on the basis that low probability BCRs could be enriched through selection steps in GC and the follicle. This approach was validated in Supp.Fig.4B, where Pgens were strongly predictive of experimental affinity measurements.

Resources, environment and reproducibility

The SeQuoIA pipeline is based on the *snakemake*[17] workflow management system. *Snakemake* executes interdependent tasks (rules) and ensures reproducibility of the workflow by deploying dependencies stored in a compiled format. *Docker* images for each *snakemake* rule were thus converted to the *Singularity* 'sif' format. The SeQuoIA pipeline can be run locally or on a cluster. CPU resources to be used are customizable. Job parallelization is automated.

256 **Construction of AID mutability model**
257
258 *Sequencing datasets*
259 The human model for SHM 5-mers mutability and substitution probabilities was already available through
260 the Shazam package. Mouse model accounting for SHM targeting in the heavy chain was still lacking. To
261 compute mutation probabilities with minimal selection interference, we chose an IgSeq dataset from Peyer's
262 Patches spontaneous GCs from germ-free mice (Supp.Fig.5A). Indeed, in the absence of gut microbiota
263 antigens, selection for GC entry was shown to be severely dampened[18].
264
265 *Mutability and substitution matrices generation*
266 The model for murine Ig sequences was built according to a standard workflow from the Immcantation suite.
267 One sequence per clone was selected to avoid biases due to clonal expansions. Substitution rates were
268 inferred with the *createSubstitutionMatrix*. 5-mer mutabilities were inferred with the *createMutabilityMatrix*
269 function, considering all mutations (synonymous and non-synonymous). Both outputs were combined into
270 a single model using the *createTargetingModel* function. 5-mer mutabilities are displayed in Supp.Fig.5B.
271
272 **scRNAseq processing and annotation**
273 Public integrative datasets used in this study are detailed in Supp.Table 3. If provided in the form of a Seurat
274 object, additional cell filterings were applied (at least 300 features, mitochondrial content < 5% of genes
275 expressed). A standard Seurat (v5.0.3) workflow was applied, including cell cycle regression, dimensionality
276 reduction and clustering. Clustering resolution was determined with an automated script based on cell cluster
277 assignment stability with the *Clustree* package. Supervised cell annotation was performed with cluster
278 marker genes as depicted in Supp.Fig.6, Supp.Fig.7 and Supp.Fig.8.
279
280 Specificities of human Covid dataset:
281 Same QC filters were applied as described above. However, the donor effect was still noticeable in the original
282 dataset (Supp.Fig.8B). Donor-based integration was performed with the semi-supervised method STACAS[19].
283 Original annotations (PCs, GC, Naive, Memory B cells) were used to refine anchor usage. GC B cells were pre-
284 annotated into global LZ and DZ categories and cells with intermediate phenotypes were labelled as
285 "unknown" and not taken into account for inconsistent anchor rejection. Histone, ribosomal,
286 immunoglobulin and cell cycle coding genes were excluded from variable features prior to integration.
287 Additional difficulties arose from this dataset, due to concomitant and intricate GC reaction and massive
288 extrafollicular (EF) response. Clustering was performed with the same variable features as for the integration.
289 Clusters were annotated based on marker genes and 3D UMAP visualization, where 3 "paths" to terminally
290 differentiated PC clusters stood out (Supp.Fig.8C). These transitional clusters were all made of *XBP1*$^{+}$ cells
291 and were annotated as follow:
292 - prePC EF: emerging from the naive B cell cluster and CD38^{neg}, suggesting that they are not GC-
293 experienced. Not BLIMP1^{+} like other prePC clusters.
294 - prePC: emerging from the LZ
295 - prePC DZ: very small cluster of proliferating cells emerging from DZ recycling cells. Whether they are
296 related to prePC remained unclear to us.
297 Likewise, an intermediary CD38^{-} cluster bridged naive and GC "meta" clusters in the 3D umap and was hence
298 annotated as "pre-GC". Despite our efforts to refine original annotations, some clusters remained ambiguous
299 and displayed marked heterogeneity. Among them, the MYC+ intermediate cluster between LZ and DZ B cells
300 was dominated by CD38^{-} and was the entry point of pre-GC cells (Supp.Fig.8D). This cluster thus likely
301 corresponded to both rare selected LZ-to-DZ B cells and newly recruited GC entrants starting their first round
302 of mutation within GCs. For some analysis, *bona fide* LZ-to-DZ were enriched by selecting cells that were not

303 directly connected to the ancestor and thus should have completed at least one GC cycle. We also
304 endeavoured to segregate EF and GC-derived PCs. However, in the absence of GC-origin lineage tracing
305 system like the one used in the murine datasets, PC origin was not obvious. Since PC cluster sub-clustering
306 did not seem to segregate two relevant PC categories, putative EF PCs were identified based on clonotype
307 overlap with prePC EF cluster. Remaining PCs were labelled “PC other” and should be enriched in cells of GC
308 origin. All analyses performed with this re-annotated dataset are gathered in Supp.Fig.11.
309
310

Results

SeQuoIA improves Ig phylogeny accuracy with biologically-based optimization and by reducing the occurrence of dubious clonotypes

Many tree reconstruction methods have been developed to recover Igs evolutionary dynamics. Albeit based on pioneering work in ecological sciences, most of the tools accommodate BCR biology. Indeed, as GC B cells proliferate and mutate in the short evolutive time span of GCR, both child and parent may coexist at the time of sampling, hence the incorporation of internal nodes. In particular, GcTree[10] achieved good rankings in simulation-based benchmarking studies[20,21] and thus served as a basis for the implementation of the SeQuoIA pipeline. This tool consists in an informed maximum parsimony method, where the best tree within the parsimony forest is picked based on node size distribution and the compliance to mutation models. As most accurate parent to daughter relationships are needed in the SeQuoIA mutation model (see next section), we added additional optimization steps to guide tree selection. Due to the presence of mutation hotspots and the high mutation rate in Ig sequences, the number of mutations may not be minimal. As a result, tree search was not restricted to the parsimony forest, but also trees derived from the parsimonious ones with a SPR approach. Heavy and light chains were kept separate for several reasons: (1) the mutation model differs between heavy and light chain[22], (2) only one chain could be recovered for a substantial proportion of single cells (Supp.Fig.9A). SeQuoIA thus selects a tree pair with most consistent lineages between both chains, and penalises trees breaching current knowledge of BCR biology based on mutation load, productivity and CSR order (see Methods and Fig.2A for details). For each clonotype in an NP-KLH immunised mouse dataset[23], we compared biology-based metrics between trees selected either with SeQuoIA or the original GcTree algorithm. GcTree could select trees with reverted mutations, while this criteria was systematically corrected with SeQuoIA (Fig.2B, Supp.Fig.9C,D). Aberrant parent to child relationships affected a larger number of cells in large clones (Fig.2B). Likewise, SeQuoIA's ability to optimise tree pairs concordance was compared against tree pairs selected independently with GcTree. SeQuoIA could erase or decrease the number of aberrations in heavy/light chain lineages for most clonotypes (Fig.2C). One example of SeQuoIA correction is shown in Fig.2E. Among clonotypes with non-aberrant chain lineages, SeQuoIA further optimised lineage overlap in tree pairs (Fig.2D).

Despite exploring a larger space, SeQuoIA sometimes failed to find concordant tree pairs. We hypothesised these residual aberrations could stem from upstream steps in the pipeline and a propagation of errors in subsequent stages. Notably, the large size of some clones for which SeQuoIA did not converge to a solution raised the question as to whether they corresponded to a unique lineage descending from the same precursor. Clonotype accuracy was all the more questionable given that NP hapten immunisation is well known to elicit canonical variable segments (IGHV1-72/IGLλ, IGHV1–53). We thus checked whether such clonotypes corresponded to a single donor or were artefactual, by combining BCR data and donor data from the gene expression library. This analysis revealed that a substantial proportion of clonotypes were of mixed donor origin (Supp.Fig.10A). The share of each donor was balanced within artefactual clonotypes, thus ruling out technical noise from donor hashtag demultiplexing (Supp.Fig.10B). Largest clones (193 and 79 cells) did not correspond to single lineages, although small clones could also be inaccurate. Artefactual clones were not specific to the low diversity hapten response. Similar observations were made in polyclonal mouse response to influenza infection[3] (Supp.Fig.10C-D). Conversely, clonotypes in the human tonsil dataset were not affected (data not shown). This divergence was further investigated with IGOr in Supp.Fig.4, which predicted higher recombination probabilities in murine clones. We sought to improve clonotype accuracy by incorporating donor origin in the assignment process. This option in the SeQuoIA pipeline enabled to decrease the number of aberrant lineages in phylogenetic trees (Fig.2F), suggesting that part of SeQuoIA's

356 failures to reconstruct concordant trees is due to the presence of unrelated cells within a clone when donor
357 origin is not specified.
358
359 **SeQuoIA selection inference and validation in public datasets**
360
361 Inferred selection increases during LZ transit and correlates with Ig affinity
362 The substrate of selection is the accumulation of different somatic mutations in Ig encoding regions. SeQuoIA
363 thus analyses mutation patterns to infer selection pressure based on their occurrence likelihood. Briefly,
364 parental sequences (i.e. sampled internal nodes in the trees) seed a model of neutral BCR evolution.
365 Mutations are applied according to the appropriate AID mutation model depending on species and Ig chain.
366 Experimental and *in silico* generated sequences derived from the same parent are then compared and the
367 divergence from non-selected virtual sequences is quantified with selection scores either at the codon or
368 sequence level (Fig.3A). We chose to include Ig synonymous mutations in our analysis, although they are
369 often presumed to be neutral, as they were shown to affect RNA stability and splicing and translation rates
370 in other contexts[24,25], and might affect BCR turnover in the DZ[12].
371 SeQuoIA selection inference was validated against current knowledge of B selection during immune
372 response. We identified cell subsets at different stages of B cell maturation in public integrative single-cell
373 datasets[23,26], through cluster-based reannotation (Supp.Fig.6-8), and applied SeQuoIA. Our pipeline
374 measured an increase in selection scores between mutations accumulated prior to GC entry and within GCs,
375 where affinity maturation takes place (Fig.3C, Supp.Fig.11A). In the extra-GC compartment, antigen-
376 independent IgD^+ Memory B cells which are thought to be activated through their TLRs[27,28] displayed the
377 lowest score of all human LN mutated B cell subsets (Supp.Fig.11A).
378 It was shown that GC selection primarily takes place in the LZ and that cognate T_{FH} signals may instruct B cell
379 recycling to the DZ[29]. Selection imprint is thus expected to be stronger in cells transitioning from the LZ to
380 the DZ. An increase in selection imprint was detected from the LZ entry to the LZ exit, in GCs from immunised
381 mouse and human tonsil specimens or FNA biopsies (Fig.3B-C, Supp.Fig.11C). PCs were associated with higher
382 selection scores, consistent with enhanced antigen binding affinity in immunisation response reported in
383 previous publications[2,3]. Similar patterns in BCR selection imprint were observed for both heavy and light
384 chains (Supp.Fig.12A, Supp.Fig.11D).
385 Another validation was performed at the gene level. Current monitoring of selected cells in the LZ rely on c-
386 Myc expression[30], which is controlled by synergistic BCR signalling and T_{FH} feedback[31]. Selection imprint was
387 not only higher in *Myc*-positive cells as compared to *Myc*-negative LZ B cells (Supp.Fig.12D), but selection
388 scores scaled with *Myc* expression levels in *Myc*-positive cells in mouse response to NP (Fig.3D) and human
389 response to COVID vaccine (Supp.Fig.11D). The BCR signaling marker *NR4A1* (encoding for Nur77), also scaled
390 with selection scores in human LNs. Therefore, SeQuoIA selection scores are relevant in the identification of
391 B cells harbouring BCRs with enhanced signalling and/or Ag internalisation capacity.
392 Finally, we sought to determine whether SeQuoIA selection scores were a good indicator of Ig binding
393 efficiency. BCRs generated in GCs induced by influenza infection were cloned *in vitro* and their equilibrium
394 dissociation constants (Kd) for hemagglutinin (HA) binding were determined experimentally in a recent
395 publication[3]. Heavy chain selection scores strongly correlated with measured affinity (1/Kd), regardless of the
396 phenotype (Fig.3E). Another feature of SeQuoIA is the inference of selection pressure at the codon scale.
397 Examples of this application can be found in Supp.Fig.2. One particular mutation is well described for its
398 antigen binding enhancing effect after immunisation with 4-hydroxy-3-nitrophenyl acetyl (NP) hapten: W→L
399 transition on the 33rd codon (W33L) in the IGHV1-72 variable segment. This mutation was associated with a
400 positive selection score. Because its maintenance was not associated with a significant negative selection
401 score (Supp.Fig.2C), W33L monitoring at late GCR stages may not be informative on BCR selection.
402

403
404
405 ### Synonymous vs non synonymous mutations
406 The NS-only option of SeQuoIA was tested likewise and resulted in similar selection score distribution across
407 cell subsets (Supp.Fig.12B). However, this analysis led to a loss in statistical power due to fewer observations.
408 Differences between DZ-to-LZ and LZ selection scores distribution were also less marked, suggesting a role
409 for synonymous mutations in DZ to LZ transition. We next examined synonymous to NS ratios and respective
410 scores across GC phenotypes. This analysis uncovered an enrichment in synonymous mutations after DZ exit
411 (Supp.Fig.12C), while most mutations in the neutral evolution model are NS (Supp.Fig.13A). Synonymous
412 mutations had intermediate to high selection scores, while a cluster of NS mutations displayed low selection
413 scores. Altogether, this suggests that most mutations are deleterious even though Ig sequence is still
414 productive and are purged upon LZ entry. Synonymous mutations, albeit unlikely, could be more prone to
415 pass this putative DZ-to-LZ checkpoint, as they would not lead to decreased Ig fitness. Conversely, NS
416 mutations were enriched in the LZ-to-DZ compartment, suggesting a more stringent selection step.
417
418 ### Comparison with BaseLine
419 Previous attempts of selection quantification based on somatic mutation patterns have been published[7] and
420 integrated within the *Immcantation* suite as the BaseLine tool. BaseLine compares replacement to silent
421 mutation ratio from observed sequences and theoretical values from a bayesian estimator based on a
422 binomial likelihood function and β prior. With this assumed distribution, the greater the mutation load, the
423 smaller the confidence interval of the selection strength estimate. SeQuoIA, on the other hand, does not
424 necessarily assume that silent mutations are neutral, and minimises the number of mutations for observed
425 and theoretical sequence comparison. The phylogenetic reconstruction upstream of selection quantification
426 allows to narrow down the modelling time scale to a few mutations (Supp.Fig.3F), thus limiting uncertainty
427 on re-mutation and mutation hotspots/coldspots reshuffles. Our quantification method does not rely on
428 replacement to silent ratio, which allows us to quantify selection at the codon level or to individual sequences
429 without aggregation[7]. Furthermore, replacement probability is equal or close to 1 in most of modelled
430 mutations (Supp.Fig.13A), which could lead to decreased ability to detect enrichment in NS mutation as
431 compared to the model. Accordingly, when we applied *BaseLine* on the same dataset as Fig.3B, no significant
432 difference was found between B cells entering or exiting the LZ (Supp.Fig.13B).
433
434 **SeQuoIA usage to investigate selection mechanisms in the GC LZ**
435 Having validated our selection quantification approach, we next attempted to identify new LZ B cell selection
436 mechanisms with our tool. To identify transcriptional programs elicited in cells expressing selected BCRs, we
437 set up a gene screening approach in LZ B cells, where normalised gene expression was correlated to Ig
438 selection score at the single-cell level. Candidate genes (spearman correlation coefficient > 0.2, $p < 0.05$) were
439 submitted to Gene Ontology (GO) analysis to interrogate their functional relevance. Biological processes (BP)
440 linked to Ig selection included expected cellular responses to B cell stimulation, such as *NF*$_{\kappa}$B signal
441 transduction which was shown to be triggered by CD40 stimulation in a subset of LZ B cell[31], Ag receptor
442 signalling, and Ag presentation (Fig.4A). At the gene level, *Irf4* expression, which was shown to be induced in
443 response to a combination of high affinity antigen and T_{FH}-derived cues in GC B cells[32], displayed a strong
444 correlation to the selection score (Supp.Fig.14). BCR signalling may also synergize or sensitise to interferons
445 in the LZ, as *Stat2* levels were correlated to selection scores (Supp.Fig.14). Overall, SeQuoIA was able to
446 identify activated LZ B cells involved in recent Ag binding events and interactions with T cells. Another
447 relevant BP was regulation of cell cycle. Altering T_{FH} feedback to LZ GC B cells in *in vivo* mouse response
448 altered GC size and cell division in the DZ, suggesting that BCR-mediated selection in the LZ regulates the
449 proliferative capacity of GC B cells[33]. In line with these observations, the expression of regulators of cell cycle

450 scaled with SeQuoIA selection scores (Fig.4A, Supp.Fig.14), supporting BCR-based selection in the LZ as a
451 prerequisite for cell cycle progression.
452 Actin cytoskeleton organisation is also one of the BP worth mentioning ; *in vitro* assays revealed that LZ B
453 cells were able to discriminate high and low affinity antigens independently of cell to cell competition thanks
454 to pulling forces exerted by their contractile cytoskeleton[34]. This discrimination resulted in affinity-dependent
455 Ag internalisation[32]. Actin dynamics can also modulate BCR signalling capacity through BCR microcluster
456 formation[35]. SeQuoIA disclosed BCR-related upregulation of actin cytoskeleton remodelling or vesicle-
457 forming proteins (Supp.Fig.14), which supports the existence of a positive feedback at the transcriptional
458 level where higher affinity BCR would benefit from stronger signalling transduction cascade in subsequent
459 interactions and increased B-T synapse formation capacity.
460 Finally, selected BCR features seem to confer a competitive advantage to LZ B cells via enhancing resistance
461 to apoptosis. BCR-driven selection was predicted to occur partly through the inhibition of pro-apoptotic
462 factors from the BCL2 family (Supp.Fig.14). Interestingly, protection from the extrinsic pathway was also
463 predicted to be fostered by selected mutations (Fig.4A), suggesting that LZ B cells with unfit BCRs could be
464 actively eliminated instead of dying by neglect. Mechanistically, this protection would be conferred by the
465 recruitment of caspase 8 inhibitors (e.g. FAIM, encoded by *Faim1*) in the Fas associated death domain FADD
466 (Supp.Fig.14).
467
468 **BCR isotypes exhibit different selection strengths**
469 In addition to Ig variable part diversifiable through recombination and SHM, BCRs have different cytoplasmic
470 domains depending on their isotypes. These structural features could lead to different signalling capacities,
471 mechanical force sensing or co-receptor interaction, thus conferring additional competitive advantage on top
472 of affinity for the Ag. Notably, IgM BCRs were shown to have a higher activation threshold as compared to
473 IgG1 BCRs in a biophysical *in vitro* assay[36], and were outcompeted in *in vivo* GCRs[11]. In keeping with these
474 observations, SeQuoIA inferred decreased selection pressure in IgG1 GC B cells variable domain, which could
475 be possible through a "native" competitive advantage (Fig.4B). Another interesting observation was the
476 increased selection imprint in IgG3 BCRs from both immunised mouse LN and human tonsil (Fig.4B-C). IgG3
477 may thus have attenuated signalling capacity, which would put an additional constraint on this population.
478
479 **Selection impact after LZ exit**
480 *DZ B cells*
481 Unlike c-Myc expression reporters, which can only detect selected cells among LZ GC B cells, Ig mutation
482 monitoring with SeQuoIA allows insight into selection strength impact on B cell phenotype beyond LZ exit,
483 either as recycling DZ B cell or effector B cell. To investigate whether and how LZ selection determines GC B
484 cell fate once they enter the DZ, we set up a new correlation analysis relying on phylogenetic tree knowledge.
485 Briefly, genes expressed in the DZ were screened for correlation between gene expression levels and parental
486 LZ selection score, if sampled (Fig.5A). LZ parent and DZ progeny pairs were retained if their BCR differed by
487 less than 2 mutations, to restrict the analysis to direct parent-progeny relationships. A GO analysis was run
488 next as described previously (Fig.5B). Among the BPs linked to LZ selection, SeQuoIA found DNA biosynthesis
489 and cell cycle regulation, in accordance with previous study concluding in accelerated S phase progression
490 for DZ B cells after enhancing T_{FH} feedback in the LZ[37]. Additionally, response to DNA damage and DNA repair
491 emerged from this analysis, which could be consistent with increased SHM mutation in selected cells. On the
492 other hand, there was no correlation between LZ selection score and AID or error prone Polη expression in
493 DZ descendance (data not shown). CXCR4 expression scaled with LZ parent selection score, suggesting that
494 fit GC B cells could benefit from a DZ stromal cell niche, or stay longer in the DZ, as shown in photoactivation
495 assays of DZ B cells[38]. In turn, extended DZ residence time could lead to an increased number of divisions and
496 somatic mutations. GO analysis also revealed the possible involvement of a transcriptional shift in metabolic

497 pathways (Fig.5B), including the upregulation of fatty acid transport (*FABP5*), and mitochondrial ATP synthase
498 and oxidoreductases (*ATP5F1A/SCD1*). The upregulation of genes involved in oxidative phosphorylation in
499 the DZ seems to emanate from BCR-related expression of regulators of metabolism (amide and carbohydrate)
500 and the generation of precursors in the LZ (Fig.4A). The regulation by LZ cues of subsequent DZ metabolism
501 supports previous observation of increased OXPHOS related gene expression in B cells carrying affinity
502 enhancing mutation W33L in the response to NP antigen[39]. Finally, inositol phosphatase *INPP5B* was strongly
503 correlated to LZ selection score (Fig.5C). This phosphatase is involved in inositol triphosphate inactivation,
504 and could thus terminate calcium mobilisation downstream of BCR signalling. This branch of BCR signalling
505 was shown to be involved in a negative feedback, with notably the upregulation of inhibitory coreceptor
506 *CD72*[40].
507
508 *Proliferation rates*
509 As discussed in the previous section, reduced cell death in the LZ seems to contribute greatly to the
510 competitive advantage of selected cells and BCR-based selection may be a prerequisite to cell cycle entry.
511 Whether increased cell division from fit cells also contributes to their maintenance over the GCR has already
512 been hypothesised[41] but not demonstrated in untouched polyclonal responses to complex antigens to our
513 knowledge. On top of GC "bursts", recent GC-derived PCs were shown to proliferate upon GC exit[23,3]. The
514 decoupling between DZ and PC proliferation raises the question as to whether one of them is driven by BCR-
515 mediated selection. We thus reanalyzed Sprumont et al dataset and compared PC and GC tree node sizes
516 correlations to their selection scores. Of note, nodes were split if a same node was associated with multiple
517 phenotypes. PC selection scores, unlike those of GC B cells, were mildly correlated to the node size (Fig.5D),
518 suggesting that BCR features may determine, at least partly, the extent of post-GC proliferation. This pattern
519 was also observed in human response to SARS-Cov2 vaccination (Supp.Fig.15A). However, it could be argued
520 that somatic mutations are a confounding factor as they often result in selection score decrease and co-occur
521 with DZ proliferation. Hence, another correlation was drawn between SeQuoIA selection scores and GC node
522 size and direct descendance, which was not significant (Fig.5E). Altogether, these results suggest some
523 stochasticity in GC proliferation, whereas PC expansion (or survival) would be somehow linked to Ig features.
524
525 *PC differentiation*
526 PC differentiation is another fate after transiting through the LZ, which has been a focus of numerous
527 publications[2,3,42] since PCs are at the frontline in the fight against pathogens. This compartment was reported
528 to be enriched in Ag binding cells as compared to total GC compartment in mice immunised with NP[43] or HEL[2]
529 in past and recent literature. On the other hand, low affinities were measured in BCRs harboured by GC-
530 derived MBCs[44]. SeQuoIA selection scores were measured in GC and output cells in murine (Fig.3B,
531 Supp.Fig.12A) and human tonsillar or LN aspirate datasets (Fig.3C, Supp.Fig.15C). In all these datasets, in light
532 and heavy chains (if available), selection imprint was lower in MBCs as compared to LZ B cells, whereas PCs
533 displayed higher selection scores, consistent with enhanced affinity reported previously[2,43].
534 Nevertheless, recent publications suggest that PC differentiation could be more permissive to low-affinity
535 cells than anticipated. In polyclonal GC responses to plasmodium and influenza infections, PC precursors[45]
536 and PCs[3] displayed similar affinities as GC B cells. This recent literature may seem at odds with stringent
537 selection inferred with SeQuoIA in four independent datasets. However, this selection imprint could be
538 affinity-independent, including autoreactive prePC/PC elimination[46]. Since the nature of selection cannot be
539 deconvoluted with SeQuoIA, we tried different strategies to further investigate PC selection in complex
540 human and above-mentioned murine responses datasets. In agreement with Sprumont et al.[3] affinity
541 measurements, unique BCRs from PC compartments had similar selection scores as GC B cells (Supp.Fig.15A).
542 However, global PC population displayed an increased selection imprint thanks to expanded selected cells

543 (Supp.Fig.15B), suggesting that selection operates in the PC compartment through differential proliferation,
544 possibly in an affinity regulated manner[23].
545 Conversely, a shift in selection scores between GC B cells and PCs was observed on unique sequences from
546 the human covid dataset (Supp.Fig.11D). We had similar findings in human tonsil dataset (Fig.3C), comprising
547 small clones without expanded nodes. Selection into or within PC compartment may thus not be restricted
548 to proliferative effect. Divergence between results obtained on human datasets and polyclonal murine model
549 could be explained by (1) exclusion of non-binding cells through cell sorting (2) FTY720 treatment preventing
550 PCs from egressing LNs, possibly confounding past and present PC generation. Within this line, selection
551 imprint increased in both GC and PCs within a few days (Supp.Fig.15C), suggesting that selection imprint
552 distribution in influenza dataset might have been partly brought down by earlier-generated PCs.
553 We thus investigated in untouched GC responses whether PC differentiation could be subject to selective
554 pressure. Interestingly, cells having an intermediate phenotype between GC and PCs, henceforth referred to
555 as "prePC", displayed lower selection scores than PCs and putative late DZ PC precursors (Fig.3B-C,
556 Supp.Fig.11A, Supp.Fig.11C). Most often, they had equivalent selection scores to LZ B cells. Altogether,
557 increasing selection score distribution over PC differentiation trajectory suggests that a selection step could
558 occur during and/or after the prePC to PC transition. Unique PC Ig sequences displayed higher selection
559 scores as compared to prePC (as well as LZ B cells) regardless of post GC expansion (Fig.3C, Supp.Fig11C),
560 suggesting that prePC to PC transition is not only subject to positive selection through proliferative
561 advantage. Negative selection could clear unfit prePC or positive selection could licence fit prePC into
562 becoming fully fledged PCs.
563 We therefore analyzed prePC and PC gene expression correlations with SeQuoIA selection scores to identify
564 putative competitive advantages conferred by Igs during PC differentiation. This analysis was carried out in a
565 published dataset[23] where PCs and prePCs were enriched prior to scRNAseq (Fig.5D). We found significant
566 correlation between Ig selection score and *Tnfrsf13b* (Fig.5E), suggesting that late stages of PC differentiation
567 are linked to BCR-driven selection. APRIL signaling through the TACI receptor, encoded by *Tnfrsf13b*, has been
568 shown to support PC differentiation upon GC exit at the GC/medullary interface (GMI)[47]. We also identified
569 other relevant genes such as *Cd37* (Supp.Fig.15D), a marker of long-lived plasma cells (LLPCs)[48]. Anti-
570 apoptotic factor *Caap1* also turned out to be correlated to Ig selection (Supp.Fig.15D), indicating that a
571 competitive advantage in terms of survival may be conferred to prePCs and PCs during or after
572 differentiation. Finally, cell growth regulator *Mtor* scaled with SeQuoIA selection scores (Supp.Fig.15D),
573 supporting results from the previous section where PC numbers correlated with their Igs selection imprint.
574 All in all, Ig features seem to determine PC fate during and after their differentiation, in terms of survival,
575 proliferation and possibly BM homing.
576
577 **Selection pressure evolution over time**
578 GCs are highly dynamic structures with variable lifetime. Despite the fact that individual GCs are
579 asynchronous[49], several mechanisms may modify the selection pressure over the course of GCRs, including
580 the emergence of regulatory GC T cell population[50], a switch from MBC- to PC-dominant exportation
581 regime[42], or increasing serum antibody affinity resulting in tighter antigen retention at the FDC surface[51]. To
582 assess whether selection pressure evolved over time in GCRs, we took advantage of a published dataset[23]
583 consisting in mouse LN sampling at different timepoints after immunisation (Fig.6A). SeQuoIA detected
584 increased selection imprint in GC B cell heavy chains over time post-immunisation (Fig.6B). Temporal changes
585 in selection score were also correlated to the mutation load in GC B cell light chains, suggesting that selection
586 strength relies on time spent in GCs (Fig.6C). Finally, SeQuoIA was applied to a clinical cohort of SARS-CoV2
587 vaccination to assess time required for immune protection (Fig.6D). Selection scores were still low at day 60
588 after primary vaccine injection, but increased between day 60 and day 201 post vaccination, suggesting that

589 effective and long-lasting affinity maturation was elicited by the vaccine after a phase of relative latency in
590 selection (Fig.6E).

591 Discussion

592 We herein introduced a start-to-finish pipeline for single-cell Ig repertoire analysis in clonally expanded B
593 cells, with improvements of current clonotype assignment and Ig tree reconstruction methods. Most
594 importantly, we developed a modelling approach for BCR-driven selection inference.
595 To date, BCR-based selection in GCs has been overwhelmingly studied through a few mouse immunisation
596 models with small antigens (NP, HEL), which are characterised by stereotypical BCRs and short-lived GCs, and
597 the monitoring of a handful of recurrent mutations. Although those models have proved valuable to explore
598 basic concepts of GC function in controlled settings, our analyses with SeQuoIA highlighted some of the
599 limitations of such models: low clonotype assignment accuracy, low pre-GC selection, irrelevance of W33L
600 mutation monitoring at late stages. SeQuoIA, on the other hand, opens up new avenues for studying
601 competitive advantage conferred by selected mutations, including in complex responses and human samples.
602 Taking into account the light chain not only increased phylogeny reconstruction accuracy, as assessed by
603 others in trees of concatenated heavy and light sequences[52], but was also useful for detecting selection
604 patterns. The increased availability and low mutation load since the last parent proved to be valuable,
605 especially when screening gene expression for correlations to the selection scores. These selection scores
606 were validated against affinity measurements and against the expression of known activation markers (NFκB,
607 MYC, IRF4).
608 Having validated our method, new mechanisms for BCR-driven selection were put forward. We propose BCR
609 signalling capacity as being an integral part of the LZ selection process, that could be a first checkpoint
610 allowing the acquisition of LZ phenotype (Supp.Fig.16). A shift in selection score distribution was indeed
611 detected in productive cells from DZ-to-LZ to LZ phenotype. NS mutations seemed to be purged in late DZ
612 stages, before DZ-to-LZ transition. Altogether, these observations suggest that failure to reach the LZ or
613 acquire LZ phenotype is not solely due to an absence of BCR or tonic signalling. One mechanism for this first
614 checkpoint can be found in the Luo et al. study which demonstrated that BCR stimulation is sufficient for
615 Foxo1 translocation from nucleus to cytoplasm, thus allowing the DZ to LZ switch[31].
616 Once in the LZ, the screening approach identified resistance to apoptosis (intrinsic and extrinsic) as a
617 transcriptional pattern associated with BCR-driven selection. Selection in the LZ is thus likely not to be
618 reduced to competition for limiting factors and death by neglect of outcompeted B cells, as stated in previous
619 work[1], especially as FAS-mediated cell death was shown to be involved in autoreactive and bystander GC B
620 cell elimination[46]. Changes in FADD composition in connection with BCR properties provides one mechanism
621 explaining the protective effect conferred by BCR signalling regarding FAS-mediated cell death which has
622 been reported in early experiments[53], but remained unelucidated so far. Besides the competitive advantage
623 in terms of survival, the expression levels of membrane proteins involved in B/T synapse were correlated to
624 SeQuoIA selection scores. BCR signalling may thus influence T_{FH} help beyond antigen internalization, through
625 transcriptional regulation of surface receptors. SeQuoIA also identified regulation of metabolic programs
626 which then translated into differential expression of enzymes/transporters connected with OXPHOS in the
627 DZ progeny. This metabolic shift may support DZ B cells survival and proliferation and be directly or indirectly
628 regulated by BCR binding properties.
629 SeQuoIA did not find a correlation between BCR fitness and GC B cells abundance, but genes involved in the
630 regulation of cell cycle seemed to licence fit cells to undergo cell divisions. The link between BCR properties
631 and proliferation may be blurred by concomitant accumulation of somatic mutations (most of which are
632 deleterious), or a stochastic component. Stochasticity in proliferation rates may lie in cognate T_{FH} encounter,
633 which may have not been captured in current mouse studies resorting to anti-DEC-205-mediated antigen

634 delivery and antigen-specific transgenic T cell transfer[33]. If differential proliferation did not seem to be elicited
635 by BCR selective value, cell death in the LZ appears as a major driver of BCR-driven selection.
636 Finally, selection imprint was found to evolve along the GCR and immunisation/vaccine time. Several putative
637 mechanisms for increased selection pressure over time were also provided in this study, including the
638 aforementioned calcium-dependent BCR negative feedback in unfit cells and enhanced BCR-signalling and Ag
639 presentation capacity through cytoskeleton remodelling in fit cells. BCR-dependent phosphorylation of actin
640 nucleators was already demonstrated[35,54], thus altering cytoskeleton through post translational regulation,
641 but SeQuoIA application to *in vivo* datasets revealed the possible involvement of a long term bidirectional
642 feedback loop with modulation of cytoskeleton composition through BCR-dependent transcriptional
643 regulation.
644 The last application of SeQuoIA provided insight into the selection events leading to PC generation. We
645 propose a model that could partly reconcile diverging literature on PC selection[3,45]. Our results are compatible
646 with Sutton et al study[45], where no enrichment in affinity enhancing mutations was found in prePCs (at least
647 when they still have a LZ phenotype). We identified a putative new BCR selection checkpoint during or after
648 prePC to PC transition upon GC exit, explaining increased affinity (or autoimmune cell clearance) in the PC
649 compartment (Supp.Fig.16). PrePC-to-PC selection checkpoint could control maturation, local PC
650 proliferation rates, and PC fate once they exit GCs and LNs, such as their longevity. The latter falls in line with
651 a publication[55] where LLPC numbers were not affected by the different amounts of newcomers in the BM
652 niche, suggesting that the turnover is determined, at least partly, by intrinsic properties. We identified anti-
653 apoptotic Caap1 as a possible determinant factor of LLPC lifetime, which expression would be settled upon
654 GC exit according to the Ig properties. Whether the Ig related competitive advantage conferred to some
655 prePCs and LLPCs is determined by upstream LZ cues or by residual surface BCR expression[56–58] warrants
656 further investigation (see Supp.Fig.16).
657 The extensive usage of SeQuoIA in this study to investigate GC selection mechanisms is a foretaste of its
658 potential applications. For instance, whether the different selection imprints in GC B cells harbouring
659 different isotypes is linked to intrinsic BCR properties or is dependent on the cytokinic context has not been
660 explored. We hope that this pipeline could have applications to study autoimmune or lymphoma BCR-related
661 pathogenesis, to assess B-cell responses in tumour tertiary lymphoid structures. Clinical applications could
662 include patient stratification in the aforementioned pathologies, monoclonal antibody design and vaccine
663 response assessment.

664 Limitations of the pipeline and the study

665 We would like to stress that the pipeline has no application in datasets restricted to non-clonally expanded
666 B cells (*e.g.* circulating B cells) since clonal phylogenies are required for inferring selection pressure. Despite
667 the robustness of SeQuoIA selection scores, some Ig repertoire analysis issues have not been resolved and
668 the application of this pipeline may be limited by the computation time and computational resources usage.
669 Applications in murine datasets may be less reliable than in human datasets, since clonotypes (as classically
670 defined) may not be fully accurate. Furthermore, the murine mutation model that we built is likely to lead to
671 an underestimation of selection imprint as it was derived from spontaneous GCs in PPs, where GC B cells
672 should nonetheless still be subject to self-tolerance. As our donor-based clonotype assignment did not fully
673 resolve the quality of phylogeny reconstruction, residual intra-donor chimeric clones or uncorrected
674 sequencing errors could explain the absence of solution following tree space exploration. Finally, mutation
675 gaps in trees and low overlap between GEX data and post-filtering Ig data lead to decreased statistical power.
676 Indeed, our results proved robust as they were convergent in four independent human and murine setups
677 but the sensitivity of our tests was limited by the low number of informative cells (cells with complete gene
678 expression and BCR sequencing data, no mutation gap). Within this line, one major limitation of this study is

679 that the most abundant clonotype of each dataset had to be discarded from the analysis due to tree
680 reconstruction errors. We are nevertheless confident that improvements in sequencing technologies and
681 BCRseq preprocessing would partly increase the overlap between informative BCR sequencing and gene
682 expression data. Experimental strategies could optimise the amount of analysable data, especially in human
683 datasets, such as small tissue wedge dissection, which allowed recovery of larger clonal portions as compared
684 to whole tissue processing in previous publication[59]. Finally, we were not always able to perform gene
685 screening on both chains. We hypothesise that the ability of SeQuoIA to correlate selection scores and gene
686 expression levels depends on the chain transcript availability, antigen type, and evolution speed as there
687 could be some latency for last mutation to translate into gene expression modulation.

688 Code & data availability

689 The SeQuoIA source code and dockers are available on Zenodo. All information and a tutorial can be found
690 on GitHub (https://github.com/MilpiedLab/SeQuoIA). All datasets used in this paper are summarised in
691 Supp.Table 3.

692 Acknowledgments

693 This work would not have come to fruition without the help of Romain Fenouil for docker environments
694 troubleshooting, including the containerization and adaptation of GcTree to the pipeline. Sabrina Baaklini's
695 help on the onset of the project was highly valuable with her pioneering work in heavy and light chain
696 clonotype assignment. We thank Thierry Mora for his support and the profitable discussions that led to the
697 current sequence selection score. We acknowledge the Computational Biology, Biostatistics and Modeling
698 (CB2M) hub at CIML for setting the computational working environment and the discussions on the pipeline
699 availability and reproducibility. We are also grateful to Chuang Dong for the datasets handover. We thank
700 past and current members of the "Integrative B cell immunology" lab at CIML for fruitful discussions on GC
701 selection mechanisms. Finally, we would like to thank Will Dumm, Erik Matsen, Adam Ye, Frederick Alt, Ali
702 Ellebedy, Adrien Sprumont and Oliver Bannard for their availability and for kindly providing code and data
703 information. This work was supported by thesis scholarships from the French Ministry of Research and Higher
704 Education and Foundation for Medical Research (FRM), by a grant from Agence Nationale pour la Recherche
705 (projet "GCselection", ANR-23-CE15-0025-01) to PM, and by institutional grants from INSERM, CNRS, and Aix-
706 Marseille University to the CIML.

707 # References

708 1. Mayer, C. T. *et al.* The microanatomic segregation of selection by apoptosis in the germinal
709 center. *Science* **358**, eaao2602 (2017).
710 2. Phan, T. G. *et al.* High affinity germinal center B cells are actively selected into the plasma cell
711 compartment. *J. Exp. Med.* **203**, 2419–2424 (2006).
712 3. Sprumont, A., Rodrigues, A., McGowan, S. J., Bannard, C. & Bannard, O. Germinal centers
713 output clonally diverse plasma cell populations expressing high- and low-affinity antibodies. *Cell*
714 **186**, 5486-5499.e13 (2023).
715 4. Victora, G. D. & Nussenzweig, M. C. Germinal Centers. *Annu. Rev. Immunol.* **40**, 413–442
716 (2022).
717 5. Chaudhary, N. & Wesemann, D. R. Analyzing Immunoglobulin Repertoires. *Front. Immunol.* **9**,
718 (2018).
719 6. Yaari, G. & Kleinstein, S. H. Practical guidelines for B-cell receptor repertoire sequencing
720 analysis. *Genome Med.* **7**, 121 (2015).
721 7. Yaari, G., Uduman, M. & Kleinstein, S. H. Quantifying selection in high-throughput
722 Immunoglobulin sequencing data sets. *Nucleic Acids Res.* **40**, e134 (2012).
723 8. Gupta, N. T. *et al.* Change-O: a toolkit for analyzing large-scale B cell immunoglobulin repertoire
724 sequencing data. *Bioinformatics* **31**, 3356–3358 (2015).
725 9. Marquez, S. *et al.* Adaptive Immune Receptor Repertoire (AIRR) Community Guide to
726 Repertoire Analysis. *Methods Mol. Biol. Clifton NJ* **2453**, 297–316 (2022).
727 10. DeWitt, W. S., III, Mesin, L., Victora, G. D., Minin, V. N. & Matsen, F. A., IV. Using
728 Genotype Abundance to Improve Phylogenetic Inference. *Mol. Biol. Evol.* **35**, 1253–1265
729 (2018).
730 11. Sundling, C. *et al.* Positive selection of IgG+ over IgM+ B cells in the germinal center
731 reaction. *Immunity* **54**, 988-1001.e5 (2021).
732 12. Stewart, I., Radtke, D., Phillips, B., McGowan, S. J. & Bannard, O. Germinal Center B Cells
733 Replace Their Antigen Receptors in Dark Zones and Fail Light Zone Entry when
734 Immunoglobulin Gene Mutations are Damaging. *Immunity* **49**, 477-489.e7 (2018).
735 13. Class-Switch Recombination Occurs Infrequently in Germinal Centers. *Immunity* **51**, 337-

736 350.e7 (2019).
737 14. Feng, J., Shaw, D. A., Minin, V. N., Simon, N. & Matsen, F. A. Survival analysis of DNA
738 mutation motifs with penalized proportional hazards. *Ann. Appl. Stat.* **13**, 1268–1294 (2019).
739 15. Yaari, G. *et al.* Models of Somatic Hypermutation Targeting and Substitution Based on
740 Synonymous Mutations from High-Throughput Immunoglobulin Sequencing Data. *Front.*
741 *Immunol.* **4**, 358 (2013).
742 16. Marcou, Q., Mora, T. & Walczak, A. M. High-throughput immune repertoire analysis with
743 IGoR. *Nat. Commun.* **9**, 561 (2018).
744 17. Köster, J. & Rahmann, S. Snakemake—a scalable bioinformatics workflow engine.
745 *Bioinformatics* **28**, 2520–2522 (2012).
746 18. Chen, H. *et al.* BCR Selection and Affinity Maturation in Peyer's Patches Germinal Centers.
747 *Nature* **582**, 421–425 (2020).
748 19. Andreatta, M. *et al.* Semi-supervised integration of single-cell transcriptomics data. *Nat.*
749 *Commun.* **15**, 872 (2024).
750 20. Davidsen, K. & Matsen, F. A. Benchmarking Tree and Ancestral Sequence Inference for B
751 Cell Receptor Sequences. *Front. Immunol.* **9**, 2451 (2018).
752 21. Zhang, C., Bzikadze, A. V., Safonova, Y. & Mirarab, S. A scalable model for simulating
753 multi-round antibody evolution and benchmarking of clonal tree reconstruction methods. *Front.*
754 *Immunol.* **13**, (2022).
755 22. Cui, A. *et al.* A Model of Somatic Hypermutation Targeting in Mice Based on High-
756 Throughput Ig Sequencing Data. *J. Immunol.* **197**, 3566–3574 (2016).
757 23. Binet, L. *et al.* Specific pre-plasma cell states and local proliferation at the dark zone –
758 medulla interface characterize germinal center-derived plasma cell differentiation in lymph node.
759 at https://doi.org/10.1101/2024.07.26.605240 (2024).
760 24. Sarkar, A., Panati, K. & Narala, V. R. Code inside the codon: The role of synonymous
761 mutations in regulating splicing machinery and its impact on disease. *Mutat. Res. Mutat. Res.*
762 **790**, 108444 (2022).
763 25. Bailey, S. F., Alonso Morales, L. A. & Kassen, R. Effects of Synonymous Mutations beyond
764 Codon Bias: The Evidence for Adaptive Synonymous Substitutions from Microbial Evolution

765 Experiments. *Genome Biol. Evol.* **13**, evab141 (2021).
766 26. King, H. W. *et al.* Single-cell analysis of human B cell maturation predicts how antibody
767 class switching shapes selection dynamics. *Sci. Immunol.* **6**, eabe6291 (2021).
768 27. Weller, S. *et al.* IgM+IgD+CD27+ B cells are markedly reduced in IRAK-4–, MyD88-, and
769 TIRAP- but not UNC-93B–deficient patients. *Blood* **120**, 4992–5001 (2012).
770 28. Weller, S. *et al.* CD40-CD40L independent Ig gene hypermutation suggests a second B cell
771 diversification pathway in humans. *Proc. Natl. Acad. Sci.* **98**, 1166–1170 (2001).
772 29. Victora, G. D. *et al.* Germinal Center Dynamics Revealed by Multiphoton Microscopy with a
773 Photoactivatable Fluorescent Reporter. *Cell* **143**, 592–605 (2010).
774 30. Dominguez-Sola, D. *et al.* c-MYC is required for germinal center selection and cyclic re-
775 entry. *Nat. Immunol.* **13**, 1083–1091 (2012).
776 31. Luo, W., Weisel, F. & Shlomchik, M. J. B Cell Receptor and CD40 Signaling Are Rewired
777 for Synergistic Induction of the c-Myc Transcription Factor in Germinal Center B Cells. *Immunity*
778 **48**, 313-326.e5 (2018).
779 32. Kwak, K. *et al.* Intrinsic properties of human germinal-center B cells set antigen-affinity
780 thresholds. *Sci. Immunol.* **3**, eaau6598 (2018).
781 33. Jing, Z., McCarron, M. J., Dustin, M. L. & Fooksman, D. R. Germinal center expansion but
782 not plasmablast differentiation is proportional to peptide-MHCII density via CD40-CD40L
783 signaling strength. *Cell Rep.* **39**, 110763 (2022).
784 34. Spillane, K. M. & Tolar, P. B cell antigen extraction is regulated by physical properties of
785 antigen-presenting cells. *J. Cell Biol.* **216**, 217–230 (2017).
786 35. Bhanja, A., Rey-Suarez, I., Song, W. & Upadhyaya, A. Bidirectional feedback between BCR
787 signaling and actin cytoskeletal dynamics. *FEBS J.* **289**, 4430–4446 (2022).
788 36. Wan, Z. *et al.* The activation of IgM- or isotype-switched IgG- and IgE-BCR exhibits distinct
789 mechanical force sensitivity and threshold. *eLife* **4**, e06925 (2015).
790 37. Gitlin, A. D. *et al.* T cell help controls the speed of the cell cycle in germinal center B cells.
791 *Science* **349**, 643–646 (2015).
792 38. Gitlin, A. D., Shulman, Z. & Nussenzweig, M. C. Clonal selection in the germinal centre by
793 regulated proliferation and hypermutation. *Nature* **509**, 637–640 (2014).

794 39. Chen, D. *et al.* Coupled analysis of transcriptome and BCR mutations reveals role of
795 OXPHOS in affinity maturation. *Nat. Immunol.* **22**, 904–913 (2021).
796 40. Verma-Gaur, J., Hauser, J. & Grundström, T. Negative Feedback Regulation of Antigen
797 Receptors through Calmodulin Inhibition of E2A. *J. Immunol.* **188**, 6175–6183 (2012).
798 41. Shlomchik, M. J., Luo, W. & Weisel, F. Linking signaling and selection in the germinal
799 center. *Immunol. Rev.* **288**, 49–63 (2019).
800 42. Weisel, F. J., Zuccarino-Catania, G., Chikina, M. & Shlomchik, M. J. A Temporal Switch in
801 the Germinal Center Determines Differential Output of Memory B and Plasma Cells. *Immunity*
802 **44**, 116–130 (2016).
803 43. ElTanbouly, M. A. *et al.* Role of affinity in plasma cell development in the germinal center
804 light zone. *J. Exp. Med.* **221**, e20231838 (2023).
805 44. Viant, C. *et al.* Antibody Affinity Shapes the Choice between Memory and Germinal Center
806 B Cell Fates. *Cell* **183**, 1298-1311.e11 (2020).
807 45. Sutton, H. J. *et al.* Lack of affinity signature for germinal center cells that have initiated
808 plasma cell differentiation. *Immunity* **57**, 245-255.e5 (2024).
809 46. Butt, D. *et al.* FAS Inactivation Releases Unconventional Germinal Center B Cells that
810 Escape Antigen Control and Drive IgE and Autoantibody Production. *Immunity* **42**, 890–902
811 (2015).
812 47. Zhang, Y. *et al.* Plasma cell output from germinal centers is regulated by signals from Tfh
813 and stromal cells. *J. Exp. Med.* **215**, 1227–1243 (2018).
814 48. van Spriel, A. B. *et al.* The Tetraspanin CD37 Orchestrates the α4β1 Integrin–Akt Signaling
815 Axis and Supports Long-Lived Plasma Cell Survival. *Sci. Signal.* **5**, ra82–ra82 (2012).
816 49. Arulraj, T., Binder, S. C., Robert, P. A. & Meyer-Hermann, M. Germinal Centre Shutdown.
817 *Front. Immunol.* **12**, (2021).
818 50. Wollenberg, I. *et al.* Regulation of the Germinal Center Reaction by Foxp3+ Follicular
819 Regulatory T Cells. *J. Immunol.* **187**, 4553–4560 (2011).
820 51. Takahashi, Y., Dutta, P. R., Cerasoli, D. M. & Kelsoe, G. In Situ Studies of the Primary
821 Immune Response to (4-Hydroxy-3-Nitrophenyl)Acetyl. V. Affinity Maturation Develops in Two
822 Stages of Clonal Selection. *J. Exp. Med.* **187**, 885–895 (1998).

823 52. Jensen, C. G., Sumner, J. A., Kleinstein, S. H. & Hoehn, K. B. Inferring B Cell Phylogenies
824 from Paired H and L Chain BCR Sequences with Dowser. *J. Immunol.* **212**, 1579–1588 (2024).
825 53. Rathmell, J. C., Townsend, S. E., Xu, J. C., Flavell, R. A. & Goodnow, C. C. Expansion or
826 Elimination of B Cells In Vivo: Dual Roles for CD40- and Fas (CD95)-Ligands Modulated by the
827 B Cell Antigen Receptor. *Cell* **87**, 319–329 (1996).
828 54. Miletic, A. V. *et al.* Vav Links the T Cell Antigen Receptor to the Actin Cytoskeleton and T
829 Cell Activation Independently of Intrinsic Guanine Nucleotide Exchange Activity. *PLoS ONE* **4**,
830 e6599 (2009).
831 55. Robinson, M. J. *et al.* Intrinsically determined turnover underlies broad heterogeneity in
832 plasma-cell lifespan. *Immunity* **56**, 1596-1612.e4 (2023).
833 56. Iseki, M., Hidano, S., Kudo, F. & Takaki, S. Control of germinal center B cell survival and
834 IgE production by an adaptor molecule containing PH and SH2 domains, Aps/Sh2b2. *Sci. Rep.*
835 **14**, 17767 (2024).
836 57. Pinto, D. *et al.* A functional BCR in human IgA and IgM plasma cells. *Blood* **121**, 4110–
837 4114 (2013).
838 58. Blanc, P. *et al.* Mature IgM-expressing plasma cells sense antigen and develop
839 competence for cytokine production upon antigenic challenge. *Nat. Commun.* **7**, 13600 (2016).
840 59. Le Coz, C. *et al.* Human T follicular helper clones seed the germinal center-resident
841 regulatory pool. *Sci. Immunol.* **8**, eade8162 (2023).
842

Figures and Legends

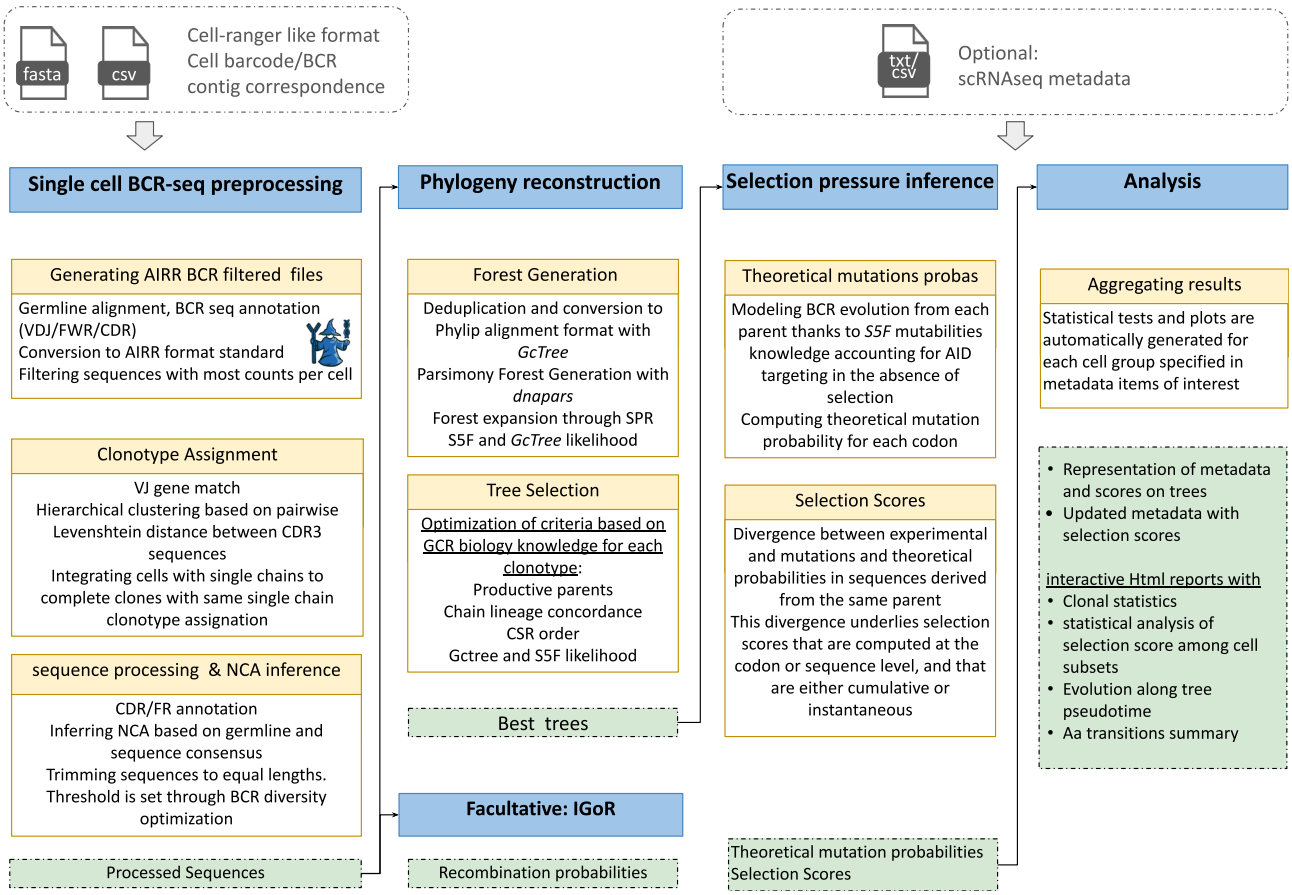


Figure 1 | SeQuoIA pipeline. Inputs and outputs are depicted in gray and green respectively. Each yellow box corresponds to scripts described in supplementary tables.

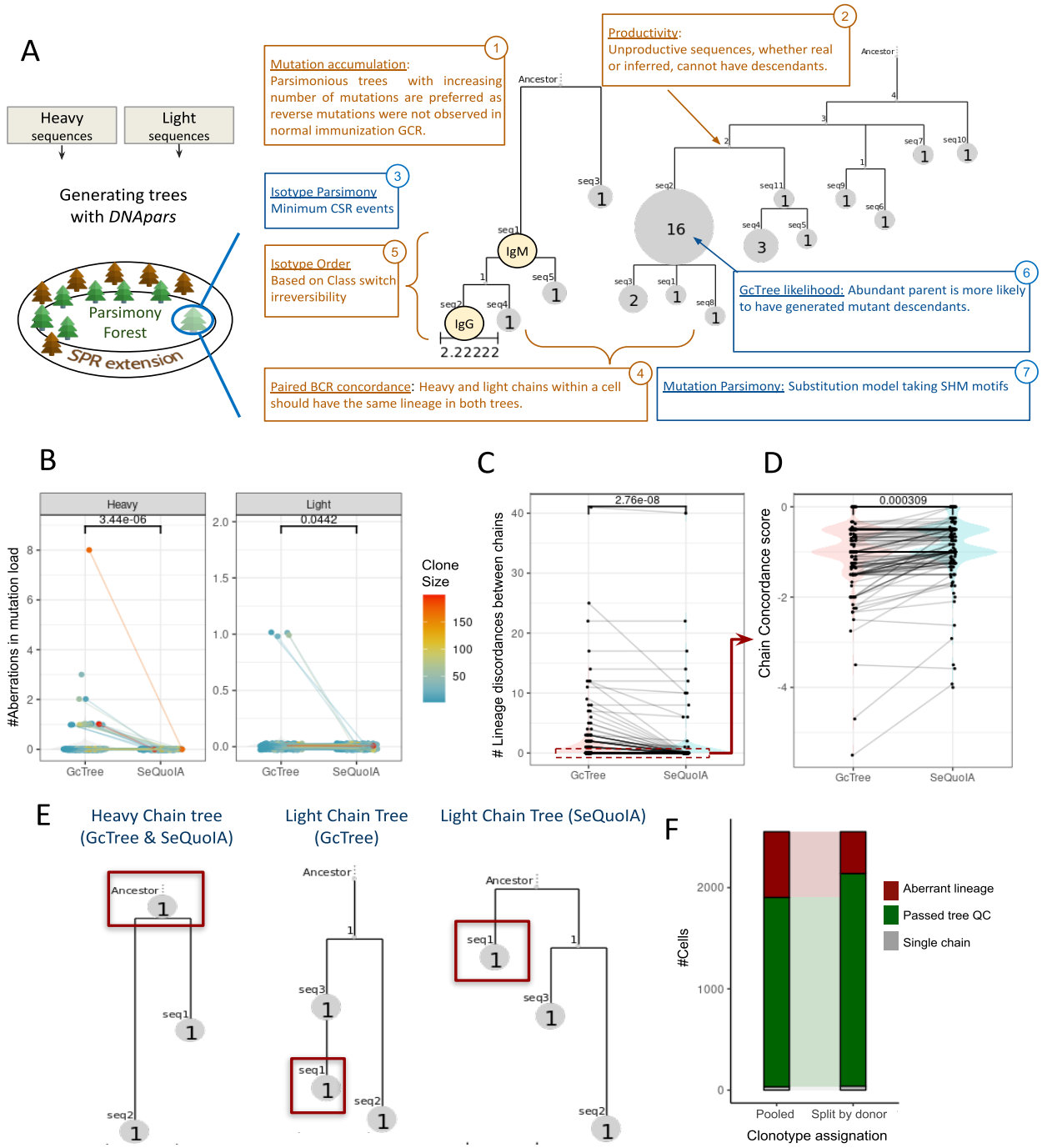


Figure 2 | BCR phylogeny reconstruction in SeQuoIA. **A)** Schematic overview of phylogeny reconstruction. Parsimony method is applied for heavy and light chain phylogeny reconstruction independently, generating two tree exploration spaces. This space is extended thanks to the SPR approach. Several criteria (boxes) are then used to select one pair of tree from the heavy and light chain spaces. Purple boxes correspond to criteria already implemented in the GcTree package, while orange criteria are custom in SeQuoIA. Numbers in the boxes corners indicate the order in which the criteria are applied. **B)** Quantification of unexpected mutation loads (criterion n°1). Each dot corresponds to a single chain tree. **C)** Dot plot comparison of the number of incompatible tree pairs (criterion n°4) between SeQuoIA and GcTree. Each dot corresponds to a clonotype with paired heavy and light sequences. The p-value from a paired wilcoxon rank sum test is indicated. **D)** Dot plot comparison of chain concordance scores between SeQuoIA and GcTree for trees with no lineage discordance across chains. The p-value from a wilcoxon rank sum test is indicated. **E)** Example of lineage aberrance correction by SeQuoIA on a small clone. Red boxes correspond to a single cell, with incompatible positions in heavy and light chain trees in the GcTree solution. **F)** Bar graph comparison of tree reconstruction performance after two clonotype assignation strategies: considering all cells from the dataset in one analysis, or splitting cells by donor sample. Only clones with more than 5 distinct Igs were considered here to discard trivial trees. All analyses were performed on mouse NP-KLH1 dataset and GcTree was applied to heavy and light chains separately.

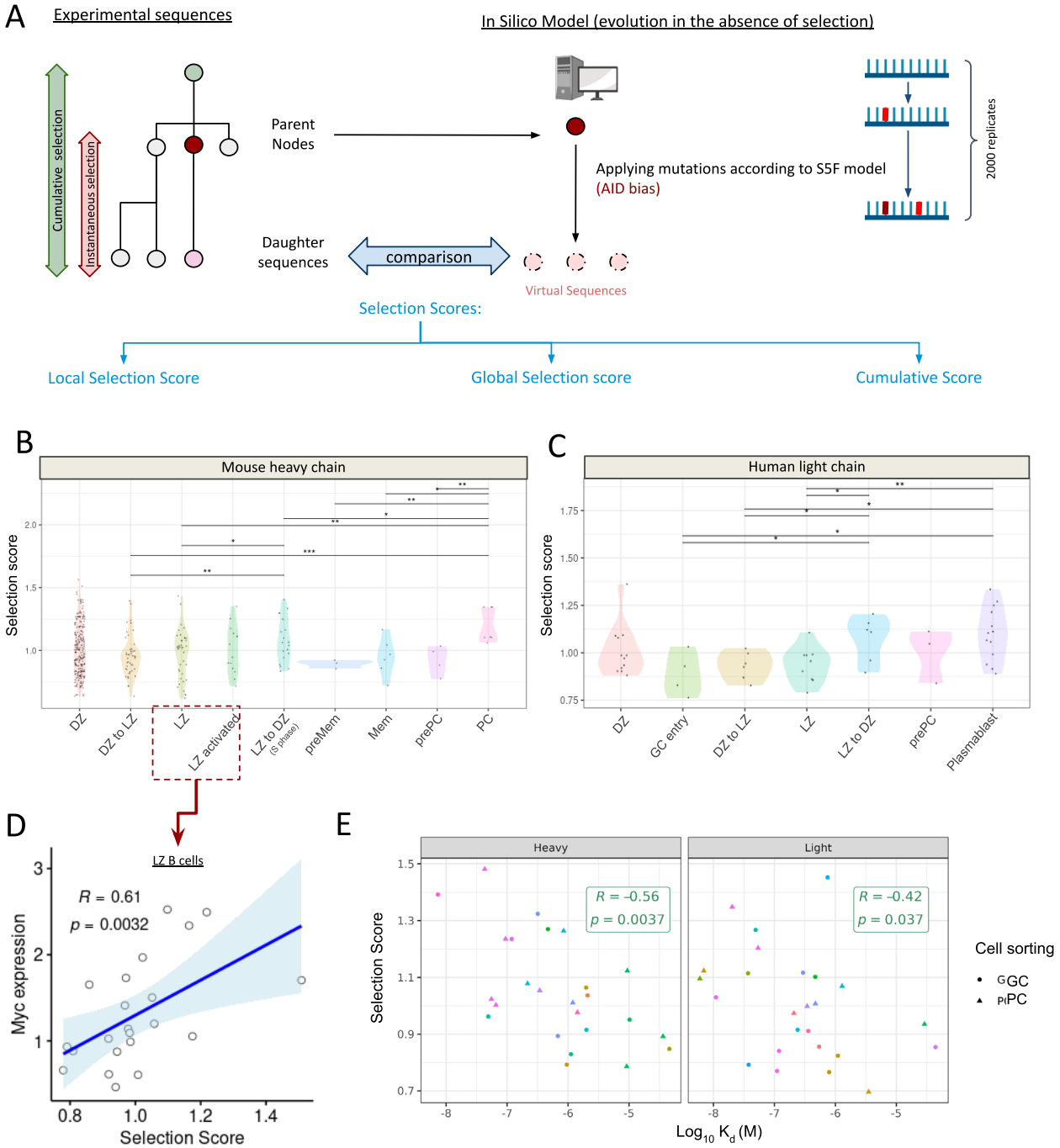


Figure 3 | Inference of selection pressure in SeQuoIA. **A)** Schematic overview of the selection quantification part in SeQuoIA. Each circle corresponds to a single chain BCR sequence. Their relationships are represented by solid lines. **B)** Dot plot distribution of global selection scores in heavy chains across B cell subsets identified through scRNAseq cluster annotation in mouse NP-KLH1 dataset. Each dot is a cell. * p<0.05, ** p<0.01, *** p<0.001, non-adjusted one-sided wilcoxon tests. **C)** same as B in human tonsil dataset **D)** Spearman correlation of light chain BCR selection score and *Myc* expression for total LZ B cells in mouse NP-KLH1 dataset. Each dot corresponds to a cell. **E)** Spearman correlation between in vitro measured affinity to HA and heavy chain (left) or light chain (right) BCR selection score in mouse influenza dataset. Colours correspond to clonotypes.

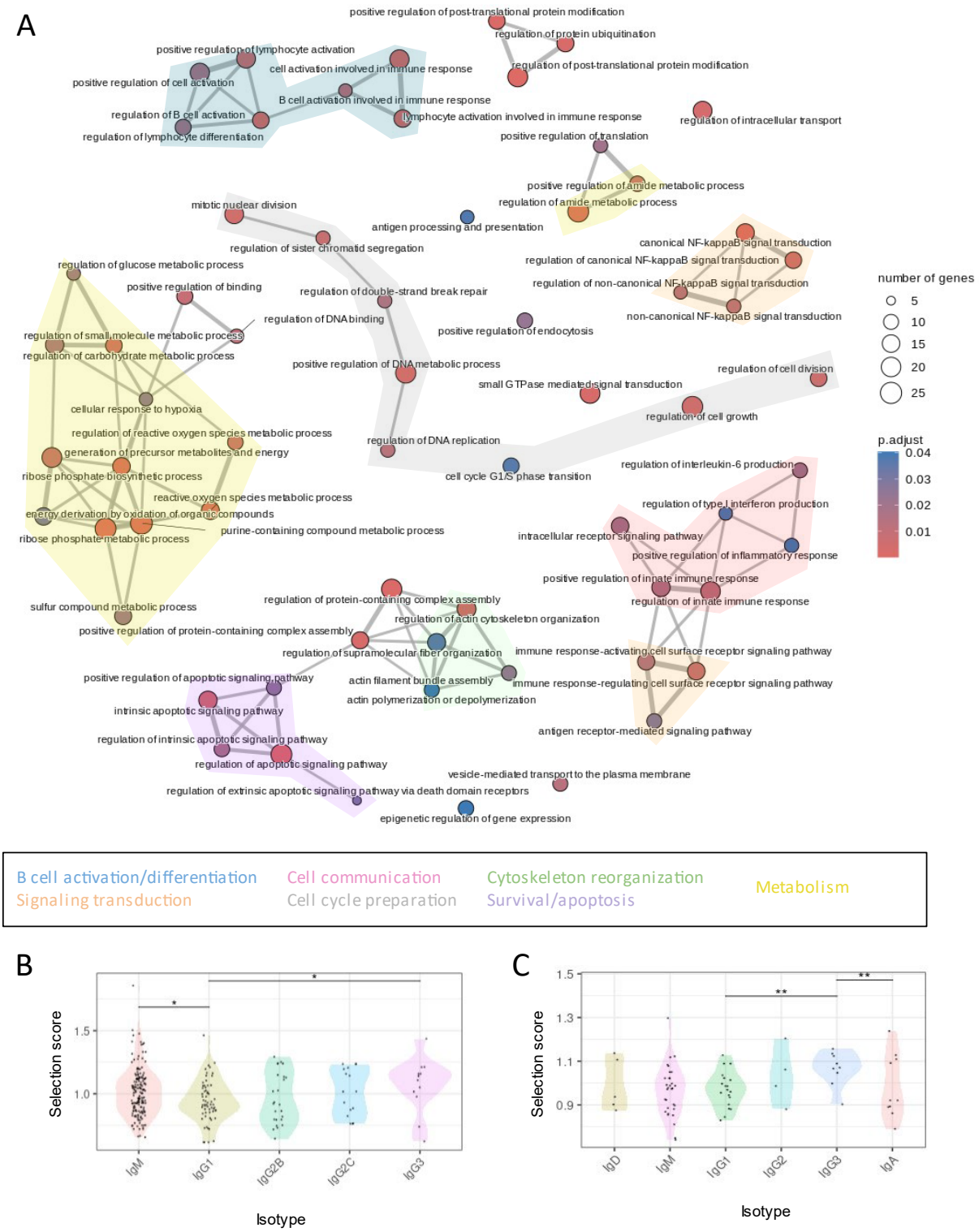


Figure 4 | Application of SeQuoIA to find biological processes involved in affinity maturation in the LZ. **A)** Graph representation of Gene Ontology terms identified for genes whose expression was positively correlated to the light chain selection score among LZ B cells from mouse NP-KLH1 dataset. Nodes correspond to biological processes from the ClusterProfiler database and edges connect overlapping gene sets. B) Dot plot distribution of light chain selection scores for GC B cells grouped by isotype in mouse NP-KLH1 dataset **C)** Dot plot distribution of light chain selection scores for GC B cells grouped by isotype (entrants and GC B cells from secondary responses are excluded) in human tonsil. P values denote two sided wilcoxon tests, * $p<0.05$, ** $p<0.01$

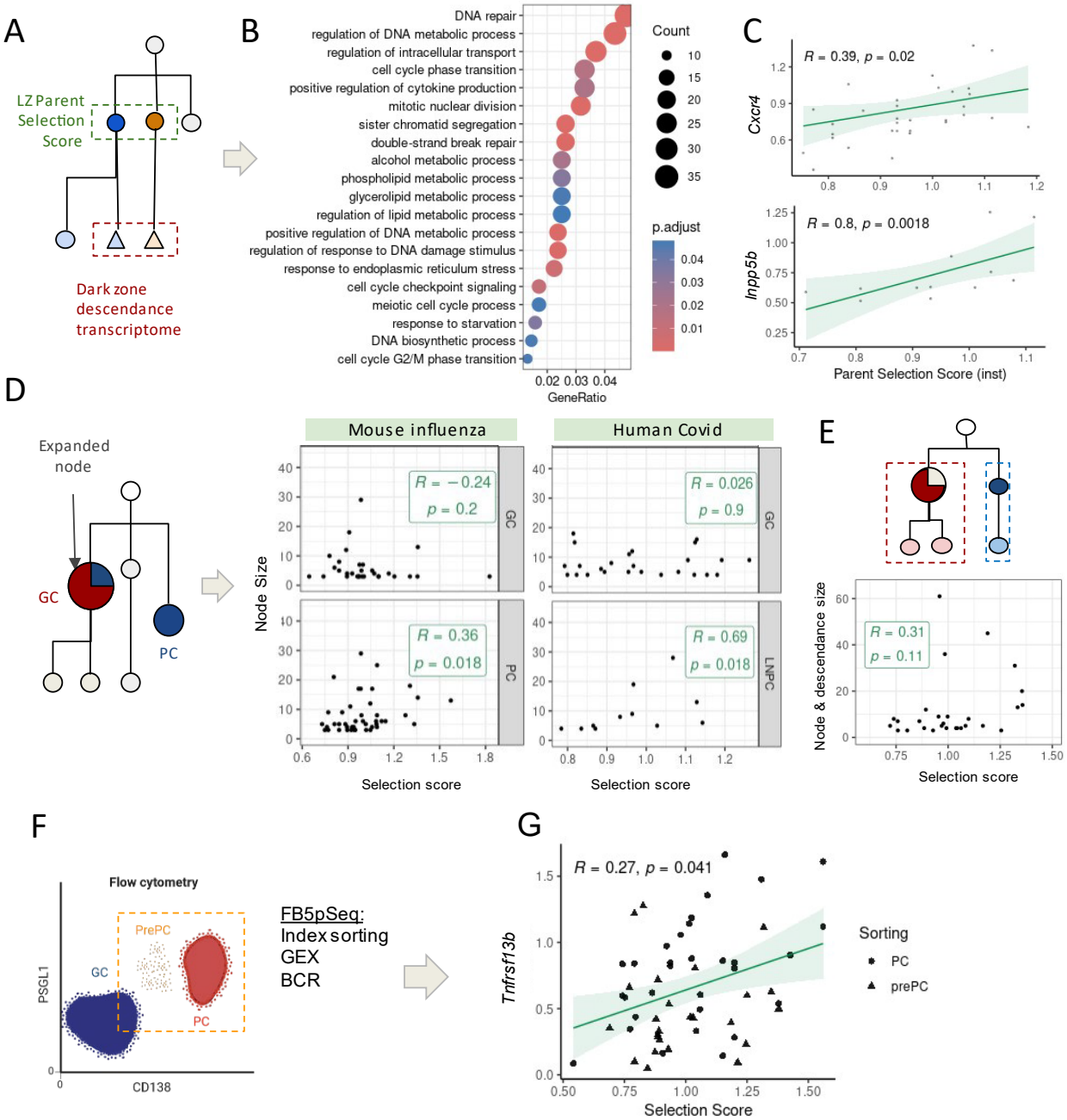


Figure 5 | Application of SeQuoIA to investigate the impact of selection after LZ exit **A)** Schematic overview of the analysis. DZ B cell gene expression was correlated to its LZ parent (if sampled) selection score. This analysis was performed on mouse NP-KLH1 dataset. **B)** Significant Gene Ontology terms from analyzing genes in DZ progeny with significant correlation (ρ > 0.2, p.value < 0.05) with LZ light chain selection score **C)** Scatter plot and Spearman correlation of the expression of selected genes in DZ B cells and light chain BCR selection score in their LZ parent. **D)** Schematic overview and analysis of Spearman correlation between node size and its associated heavy chain selection score, with nodes split by GC and PC cell types. Only expanded nodes (i.e. > 3 sequences) were considered. This analysis was performed in in mouse influenza and human covid datasets. **E)** Schematic overview and analysis of Spearman correlation between node plus progeny size and the node's light chain selection score, focusing on GC B cells in mouse influenza dataset. **F)** Schematic of mouse PC and prePC enrichment experiment from mouse NP-KLH2 dataset **G)** Spearman correlation of prePC and PC light chain selection scores with *Tnfrsf13b* gene expression.

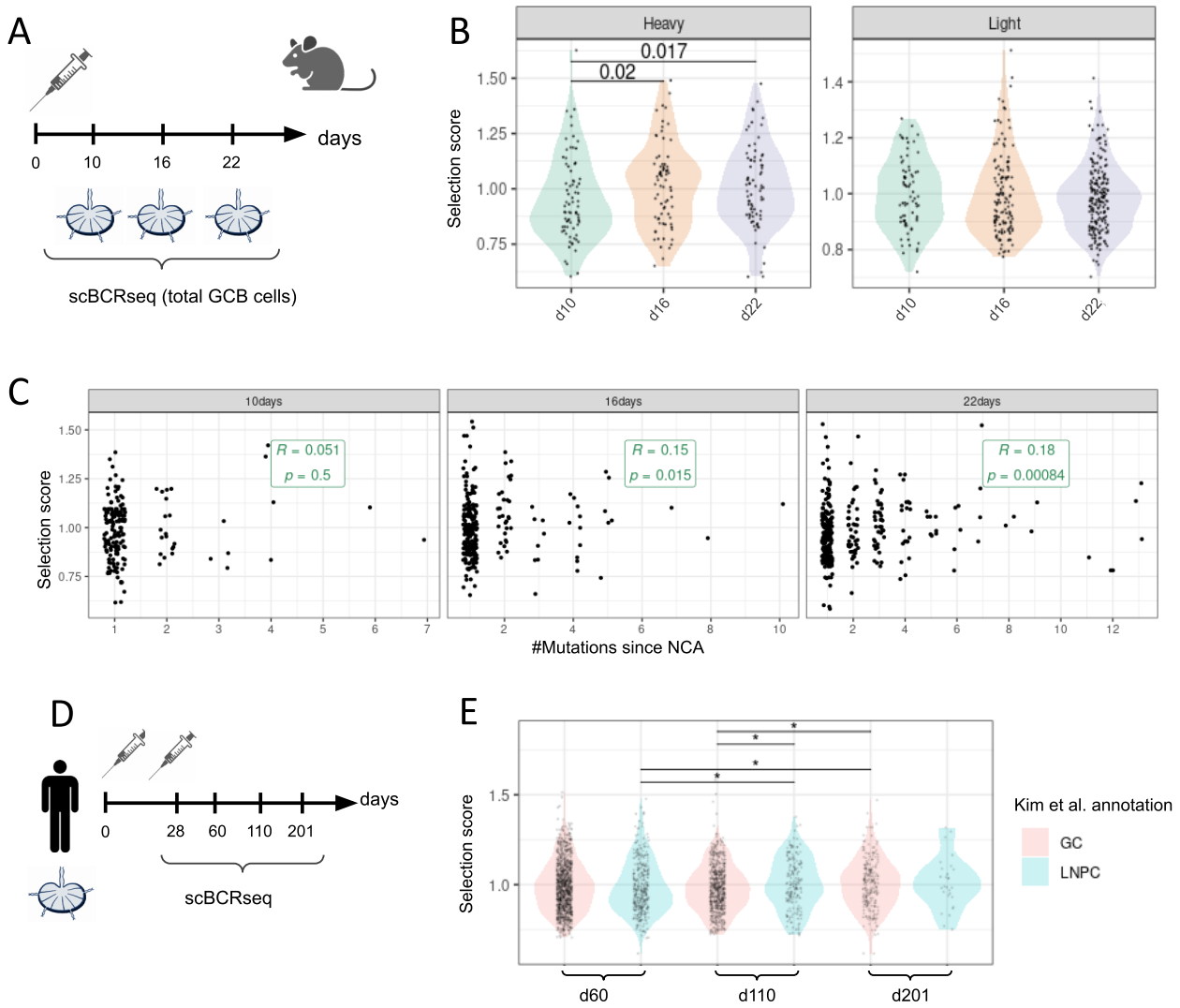


Figure 6 | Evolution of selection pressure in germinal centres over time. **A)** Schematic overview of immunization and sampling from mouse NP-KLH3 dataset. Cells from draining lymph nodes of two mice were analyzed at each timepoint. **B)** Dot plot analysis of selection scores in IgH (left) and IgL (right) at different time points after immunization. Each dot represents a sequence 1 mutation away from its parent. p-values denoting Wilcoxon rank sum test results are indicated. No multiple testing correction. **C)** Scatter plot of selection score by number of mutations from clonotype NCA (considered as BCR phylogenetic tree pseudotime) in the light chain at given time points after immunization. The selection score is significantly correlated (Spearman correlation) to distance from tree root at intermediate and late timepoints. **D)** Schematic overview of vaccination and sample collection time points in human Covid dataset**. E)** Dot plot distribution of selection scores in GC B cells (pink) and PC (green) at the indicated time points after vaccination. *p<0.05, wilcoxon rank sum test, non-adjusted.

910 # Supplementary Tables and Figures

911

	Snakemake Rules	Called Script	Description
Single cell BCR seq preprocessing	assigning_vdj_gene_annotation	AssignGenes.py (Immcantation)	V(D)J gene segments are annotated for each contig.
	standardized_database_generation	MakeDb.py (Immcantation)	Tab-delimited database file is created to store sequence alignment information.
	sequence_filtering	02_Filtering_BCR_seq.R	Most expressed contig per heavy and light chains are selected based on UMI counts. Sequences with poor alignement are filtered out.
	adding_germline	CreateGermlines.py (Immcantation)	Germline sequences from alignment data are added to AIRR tables. Junction nucleotides are still missing at this stage.
	mutation_freq_computation	03_Mutation_Analysis.R	
	clonotype_assignment	launch_reports_compilation.R	Automated Rmarkdown report is generated.
		01_prepareData.R	BCR contigs are linked to cell barcodes. Data is split by chain.
		02_cellSimilarityChainBased.R	Distance matrices are computed for heavy and light chain sequences. Hierarchical clustering is performed on each matrix.
		03_ClonesAssignment.R	Heavy and light chain data are merged back for clonotype assignation.
		04_cloneVisualizationSummary.R	Output figures and tables are produced within the html report and output data is saved in AIRR format with clonotype information.
	NCA_Annotation	UCA_MRCA_annotation.R	One subfolder per clonotype is generated, the name being informational on the clone Size and V(D)J composition. For each clonotype, ancestral sequences are reconstructed. These ancestral sequences include complete germline sequence with junction sequence and Nearest Common Ancestor (NCA). NCA reconstruction is based on a strict consensus of least mutated sequences for each nucleotide. Germline nucleotide is taken in positions lacking consensus. The script also checks ancestral sequence productivity. Ancestral and experimental sequences are then trimmed to equal length. The sequence length for trimming is set so as to maximize the number of different sequences. Sequences shorter than the chosen length are dropped.
Phylogeny reconstruction	BCR_Parsimony_Forest	01_Generate_Parsinoy_Forest_customParse_Chains_parallel.sh	This part of the script directly calls *dnapars* and *GcTree* functionalities. (See https://github.com/matsengrp/gctree) Identical BCR contigs are collapsed into unique sequences. BCR lineage reconstruction is performed with the maximum parsimony method (dnapars). Galton-Watson branching process parameters are computed with GcTree.
		02_Parsing_Forest_Outputs.py	Parsimony forest (-p format) is parsed and saved

			into single tree files (fasta, nk, .p). The original parsimony forest is extended to cover a larger tree space. The additional trees are generated from parsimonious ones with a prune and regraft approach, where branches can be moved upstream in the tree. Node abundance information and Galton-Watson process parameters are used to compute *GcTree* likelihood.
	Tree_selection		At this stage, the tree exploration space is made of the parsimony forest and the prune and regraft derived trees. Trees featuring unproductive parents are dropped from the exploration space. If several trees have the same topology, the one with the best S5F mutation parsimony is selected. Node position and all leaves lineage are retrieved from the newick formatted trees. Pairs of trees with anticorrelated node positions and discordant lineages are discarded.
Selection forces inference	compute_theoretical_probas	01_Simulation_MutProbabilities_Phylo.R	Parental nodes sequences are extracted and used to initialize a model of BCR evolution in the absence of selection. If the selected tree topologies featured only one tree (*i.e.* intermediary sequences reconstruction is unambiguous), virtual nodes can be taken as a parent. Within a model, mutations are applied sequentially until the maximal parent to daughter distance is reached. Sequence productivity is tested at each round of mutation. 2000 replicates of these models are run and the probabilities of a codon being non-synonymously mutated are calculated from the collection of *in silico* generated daughter sequences. A QC plot is saved at this stage to compare WRC (hotspot) and SYC (coldspot) mutation probabilities.
	local_selection_score	01_Selection_Score.R	Selection scores are computed at the codon and sequence level (see formulas in methods) and summarized in the output table. This script also launches the report compilation. An Rmarkdown report is generated for each clonotype.
		02.a_Local_Selection_Scores.R	This chunck generates interactive plots to explore mutation selections scores and effects at the sequence level, on cells which can be selected based on scRNAseq metadata.
		02.b_Global_selection_groups.R	This chunck generates plots for selection scores comparison across cell groups defined in the metadata within each clone.
		02.c_GloalSelection_pseudotime.R	Selections scores are plotted against tree pseudo time (mutations since germline and NCA).

Analysis	evolution_of_scores	Plotting_Selected_Trees_Scores.py	Selection Scores and columns of interest in the metadata are plotted on the selected trees. Each node corresponds to a pie chart with the cell phenotypes/BCR features (selection score, isotype, etc.).
	results_aggregation	launch_reports_compilation.R	An Rmarkdown report summarizing selection scores distribution across clonotypes is generated.
		01_CloneSize_Correlation.R	Correlation between selection scores and clone size are computed and plotted for each selection score.
		02_Clonal_Diversity.R	Clonotype composition and diversity visualization are included in the Rmarkdown report.
		03_SelectScore_groups.R	This chunck generates plots for selection score comparison across cell groups defined in the metadata. All sequences across clones are gathered for this analysis.
		04_AA_properties.R	A summary of local selection scores and a visualization of amino acid changes in term of size and chemical properties are included in the html report.

Supplementary Table 1 | Overview of scripts in the SeQuoIA pipeline (related to Fig.1).

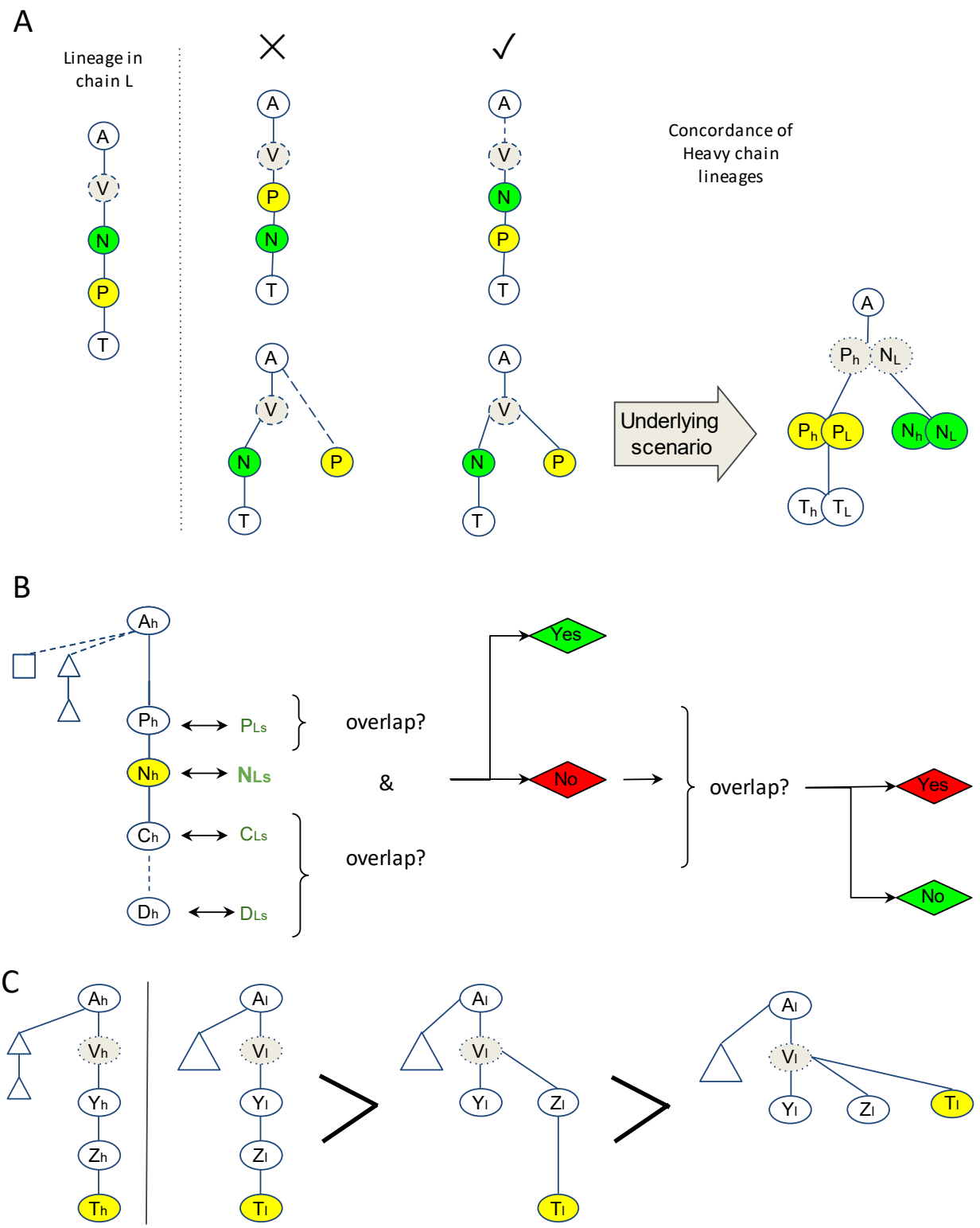
Name	Localization	User involvement	Format	Usage
*.fasta	Rawdata > VDJ	Mandatory	> BCR_contig_id sequence	experimental sequences
*annotation.csv	Rawdata > VDJ	Mandatory	cell ranger vdj-like format. cell_id contig_id columns are mandatory.	Linking chain sequences to single cell id
metadata.extension	Rawdata > scRNAseq	Optional but highly recommended	txt/csv accepted must contain "barcode" or "cell_id" column. Otherwise, rownames will be taken as cell ids.	Comparison of selection scores between groups of cells (annotated during the analysis of scRNAseq data).
config_BCRseq.yaml	Workflow	Pre-filled. Species variable is mandatory.	fields are pre-filled but are editable: -computational resources -species (mouse or human) -metadata column of interests (optional)	computational resources for parallelization, species for alignment, relevant metadata
mouse and human reference genomes	Reference > Genomes Alignments	Built-in but editable	fasta	germline sequences for BCRseq alignment
mutability.csv *substitution*.csv H (human) M (mouse)	Reference > Substitution model	Built-in	mutability and substitution matrices in csv format (cf SHazaM)	model of BCR evolution
Sequence_Patterns.csv	Reference	Built-in but editable	csv prefilled, column content can be changed additional columns can be added but will not be taken into account	plotting codon selection scores. Colors correspond to BCR sequence amino acid properties and PTM.

Supplementary Table 2 | Overview of data inputs into the *SeQuoIA* pipeline (related to Fig.1).

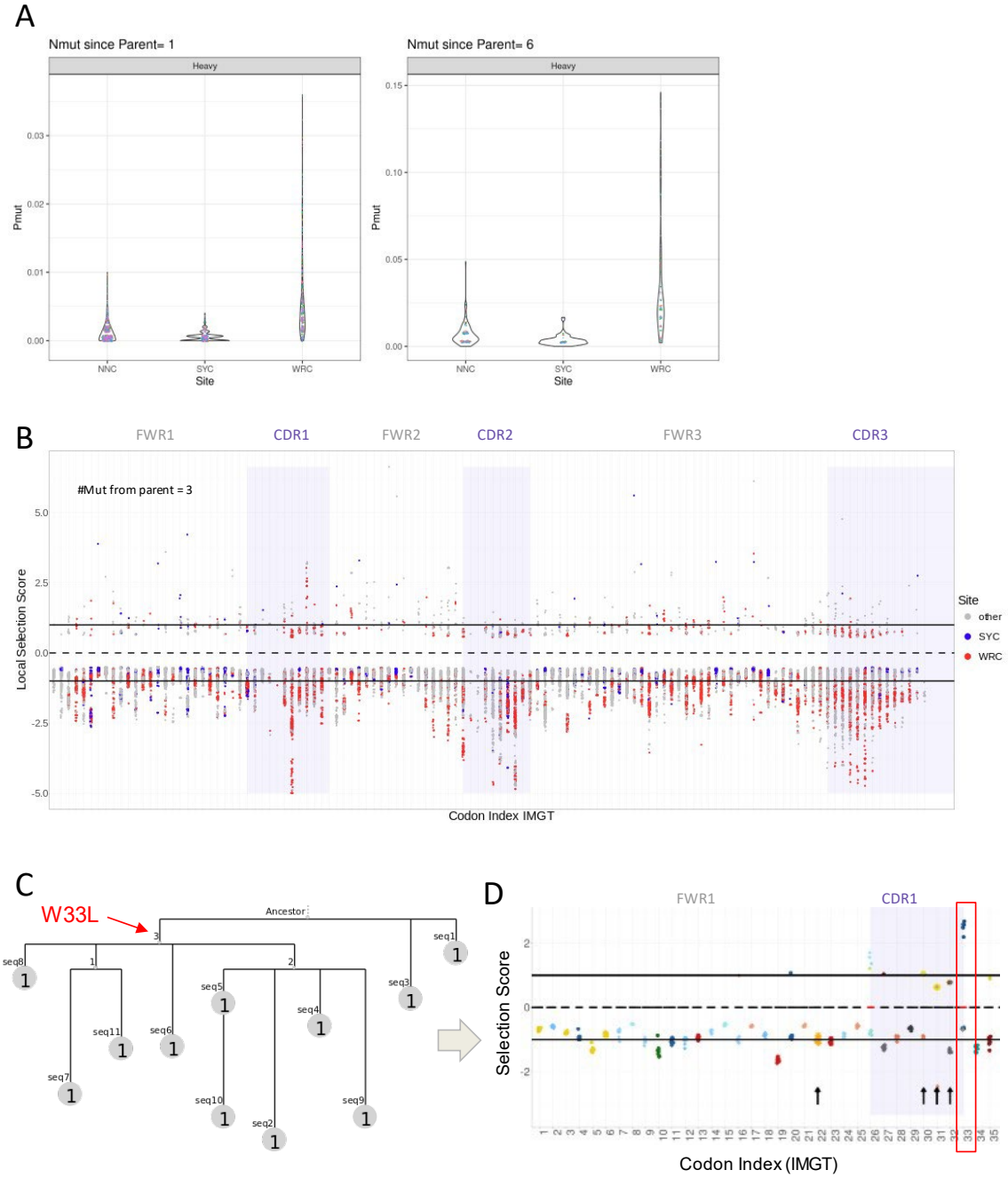
Source publication	Accession numbers	Technology	Species	tissue	#donors	Description	Associated figures	Short name
Binet et al 2024 DOI: 10.1101/2024.07.26.605240	GSE275122 GSE275124 GSE275123	10x 5’	mouse	LN	3	Mice were immunized with NP-KLH and sacrificed at day 16 after immunization. Umi data available for QC filters. GC or GC-derived B cells were sorted with the AID-Cre-YFP reporter.	Fig.2 Fig.3B, Fig.3D, Fig.4A, Fig.4B, Fig.5B, Fig.5C, Supp.Fig.2, Supp.Fig.3, Supp.Fig.4A, Supp.Fig.4B, Supp.Fig.6, Supp.Fig.9, Supp.Fig.10A-B, Supp.Fig.11, Supp.Fig.12A	NP-KLH1
					2	Mice were immunised with NP-KLH and sacrificed at day 16. prePC and PCs were enriched by FACs prior to sequencing. GC or GC-derived B cells were sorted with the AID-Cre-YFP reporter.	Fig.5F	NP-KLH2
					2/ Time point	Mice were immunised with NP-KLH and sacrificed at day 7, 16, or 21. Umi data available for QC filters. GC or GC-derived B cells were sorted with the AID-Cre-YFP reporter.	Fig.6A, Fig.6B, Fig.6C	NP-KLH3
Sprumont et al. 2023 DOI: 10.1016/j.cell.2023.10.022	Upon request	Plate based	mouse	LN	6	Mice were infected with influenza and sacrificed at day 18. GC and GC-derived PCs with measurable affinity for HA were sorted. No GEX library.	Fig.3E, Fig.5D, Fig.5E, Supp.Fig.4D, Supp.Fig.10C, Supp.Fig.10D	Mouse Influenza
King et al 2021 DOI: 10.1126/sciimmunol.abe6291	E-MTAB-8999 E-MTAB-9003 E-MTAB-9005	10x 5’	human	Tonsil	5	scRNAseq was performed on pediatric tonsilectomies. No cell enrichment.	Fig.3C, Fig.4C, Supp.Fig.3E, Supp.Fig.4A, Supp.Fig.4B, Supp.Fig.7	Human tonsil

Kim et al 2022 DOI: 10.1038/s41586-022-04527-1	PRJNA731610 PRJNA741267	10x 5’	human	LN/ Blood	15	Donors with confirmed SARS-Cov2 infection provided fine needle aspirates of draining axillary lymph nodes. GC B cells, memory and PC were sorted. Elisa was performed on LN and blood derived monoclonal antibodies, with recombinant SARS-Cov2 spike protein.	Supp.Fig.8 Supp.Fig.14C	Human covid
Chen et al 2022 DOI: 10.1038/s41586-020-2262-4	GSE140795	Rep-SHM Seq	mouse	PP	10	Mice repertoires were compared between WT and GF mice PP repertoires. Sequences from GF mice were used to compute the SHM targeting model.	Supp.Fig.5	

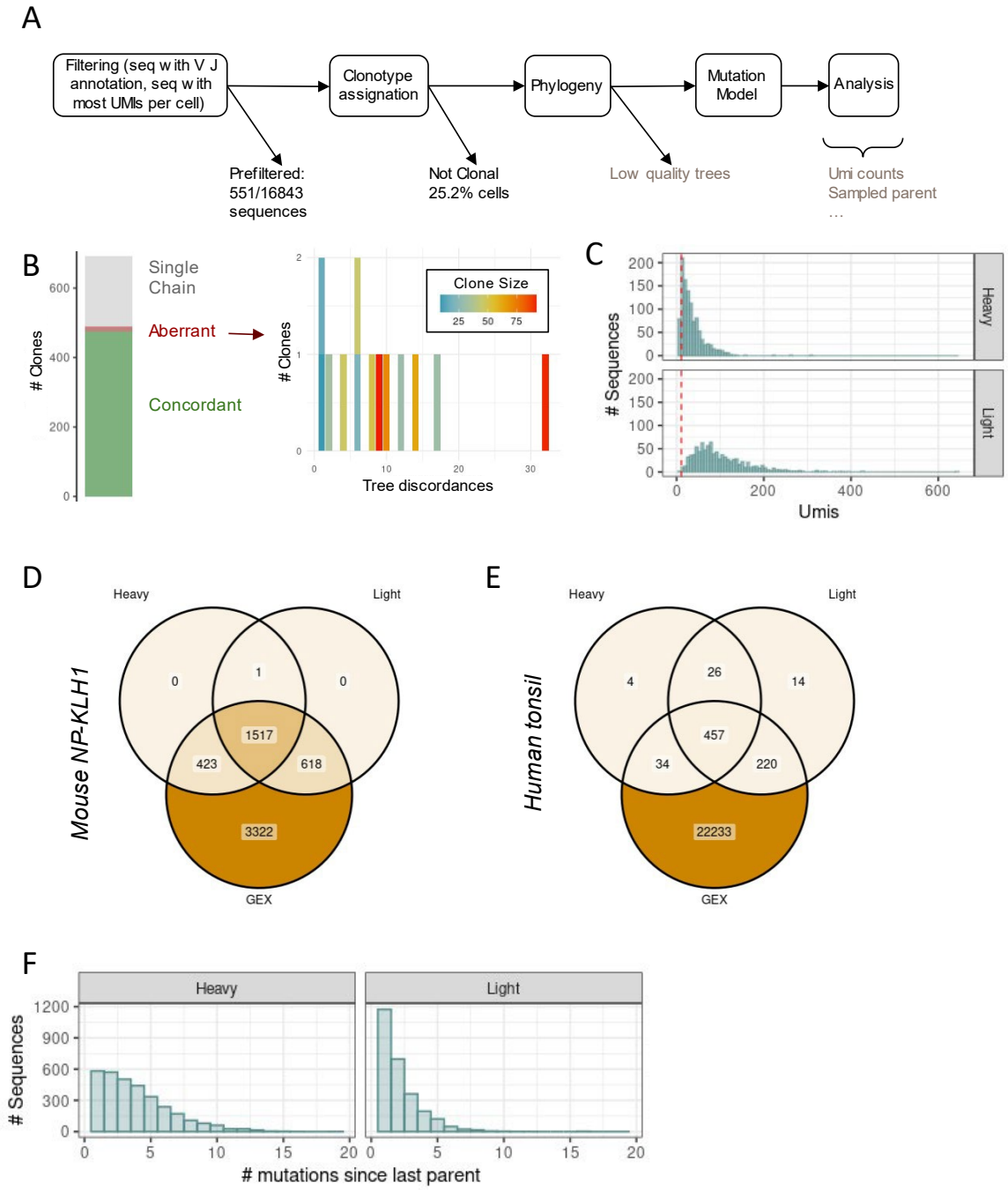
Supplementary Table 3 | Description of datasets used in this study.



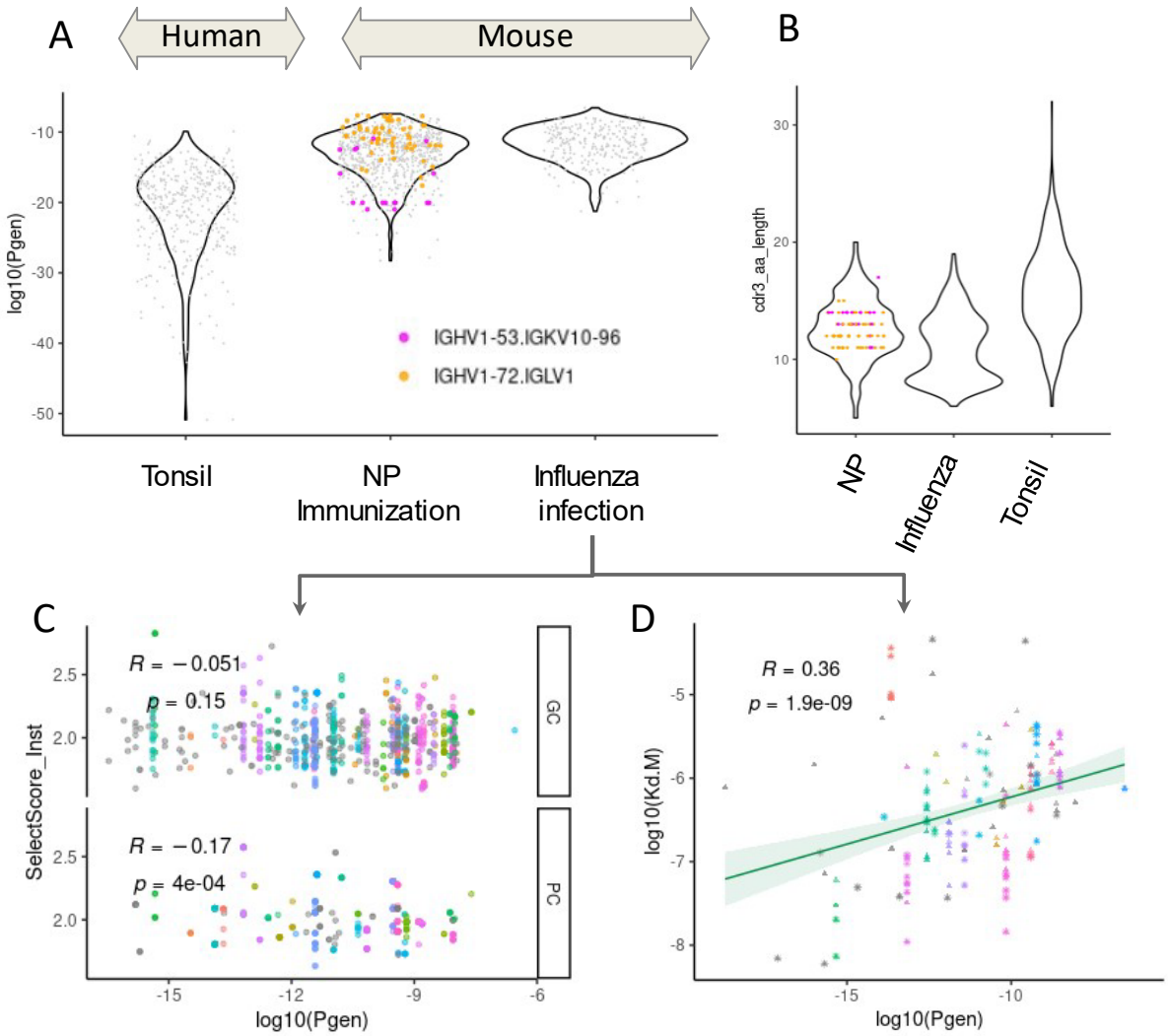
Supplementary Figure 1 | Details on tree selection implementation. **A)** Schematic representation of tree selection based on concordance criterion. SeQuoIA first checks whether BCR relationships are consistent in heavy chain and light chain trees. **B)** The order of paired light chains obtained from the heavy chain trees and the actual order in the light chain trees has to be conserved. If so, SeQuoIA checks that a paired node is not scattered in different sub-lineages in heavy and light chain trees, by checking that ascendance and descendance have common elements in both trees. If a sequence is both within the ascendance and descendance without a continuum, a penalty is applied to the pair of trees. **C)** Lineage overlap optimization among consistent tree pairs. P = Parent, N = node, C = Child, D = descendant, A = ancestor, T = terminal node, V = virtual node. H = heavy, L = light



Supplementary Figure 2 | Local selection scores. All analyses are performed on mouse NP-KLH1 dataset **A)** Dotplot of raw mutation probabilities quality control of known hotspots and coldspots as provided by the SeQuoIA pipeline at the end of the mutation modeling step. **B)** Codon selection score distribution across heavy chain, all clonotypes combined. Solid lines materialize putative positive (upper part) and negative (lower part) selection. **C)** Phylogenetic tree of a clonotype featuring sequences with and without W33L mutation in its heavy chain. **D)** Local selection score distribution in the same clonotype. The acquisition of W (blue) mutation is associated with positive selection score whereas its maintenance does not appear to be the result of negative selection.



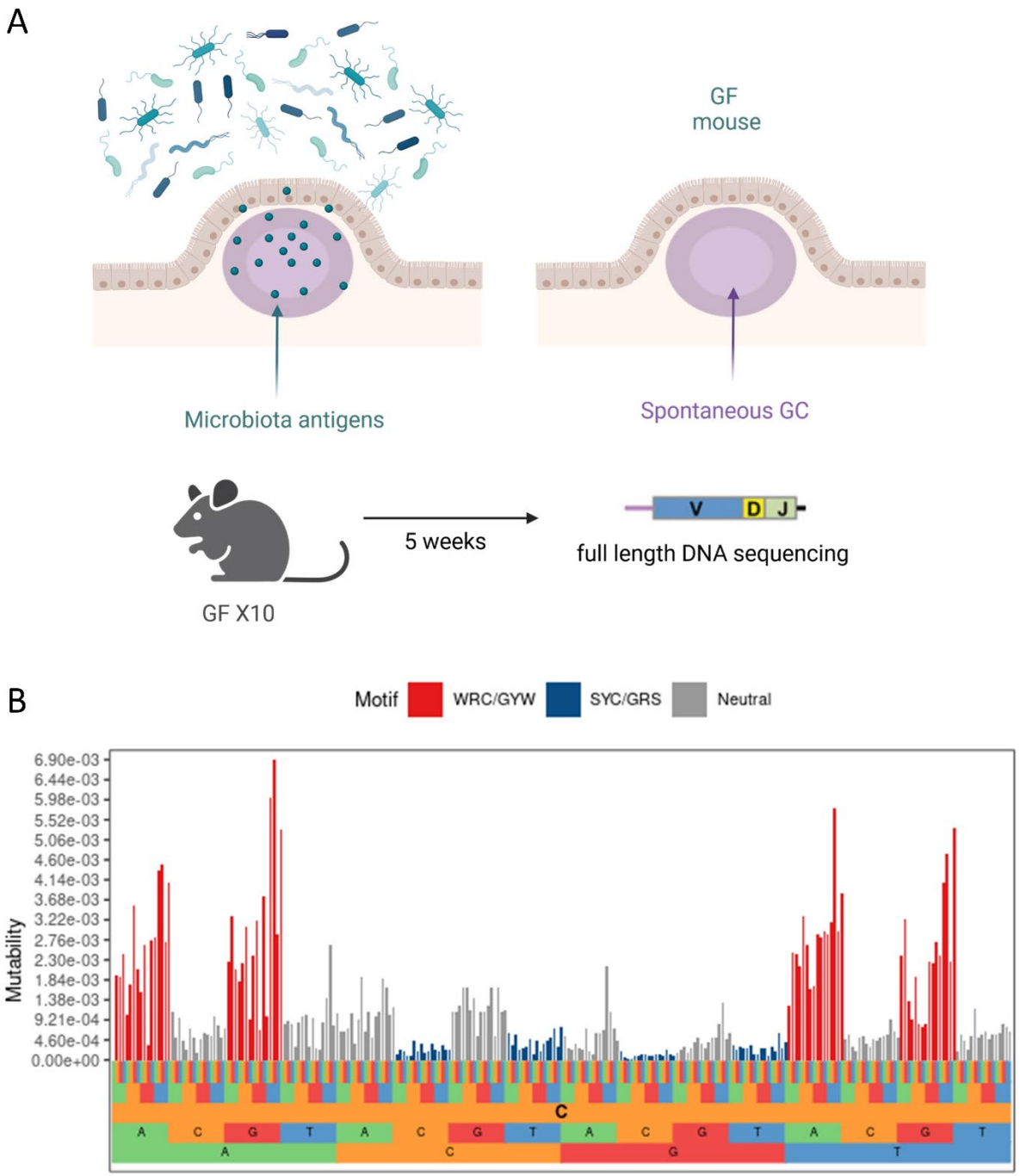
Supplementary Figure 3 | Features of integrative single cell datasets. **A)** BCR data loss after quality controls. Filtering steps in black are automated whereas filtering steps in gray are part of user's choice. **B)** Quality of tree pairs within clonotypes. Right panel depicts clone size and the number of discordances. Clonotypes with concordant tree pairs were kept for following analyses. Trees with low discordance to clonotype size ratio were not discarded to compromise between data quality and data loss. **C)** BCR library UMI counts distribution for non-PC and non-proliferative cells. Cells with low UMI counts (< 15) were discarded. **D) E)** Single-cell gene expression (GEX) and BCR information availability in mouse (D) and human (E) datasets after filtering steps. **F)** Mutation from last parent in tree distribution. Low mutation counts is associated with enhanced modeling accuracy and is informative on instantaneous selection. All figures correspond to mouse NP-KLH1 dataset, unless stated otherwise.



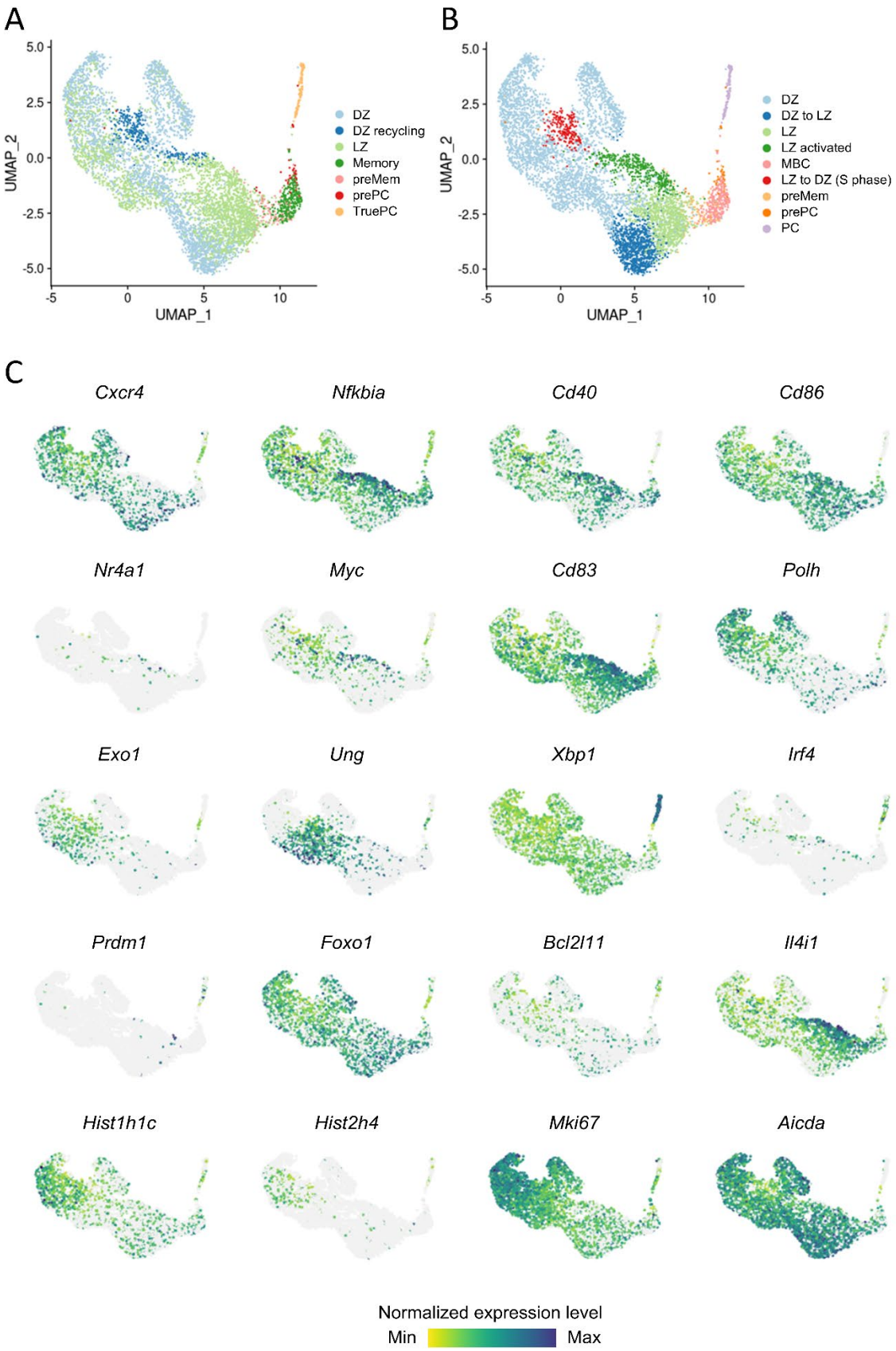
Supplementary Figure 4 | Application of IGoR within the SeQuoIA pipeline **A)** Pgen distribution of NCAs across datasets. Clonotypes recurrent in NP immunization model are highlighted. **B)** CDR3 length (nucleotides) comparison between datasets. **C)-D)** Spearman correlation of Pgen and affinity. Dots correspond to sequences with measured affinities and colors to clonotypes (within which all sequences have the same Pgen) in Sprumont *et al*. dataset.

We compared the intrinsic generation probabilities (Pgen) of human and murine clonotypes to investigate whether chimeric clones were due to low of preGC selection. Murine clonotypes from both pauciclonal and polyclonal responses were associated with higher Pgen. Reasons for this could be either a mouse-specific intrinsic bias in VDJ recombination, or permissive pre-GC selection where rare clones would not be especially favoured. In murine models, competition for follicular niches and antibody feedback should be limited as there is no infection history. In line with the first hypothesis, CDR3 lengths were smaller in mice than humans, thus limiting junctional diversity. IGOr Pgen inference was validated in figure C, where Pgen are inversely correlated to the affinity, suggesting that most probable combinations are not present due to a selection process.

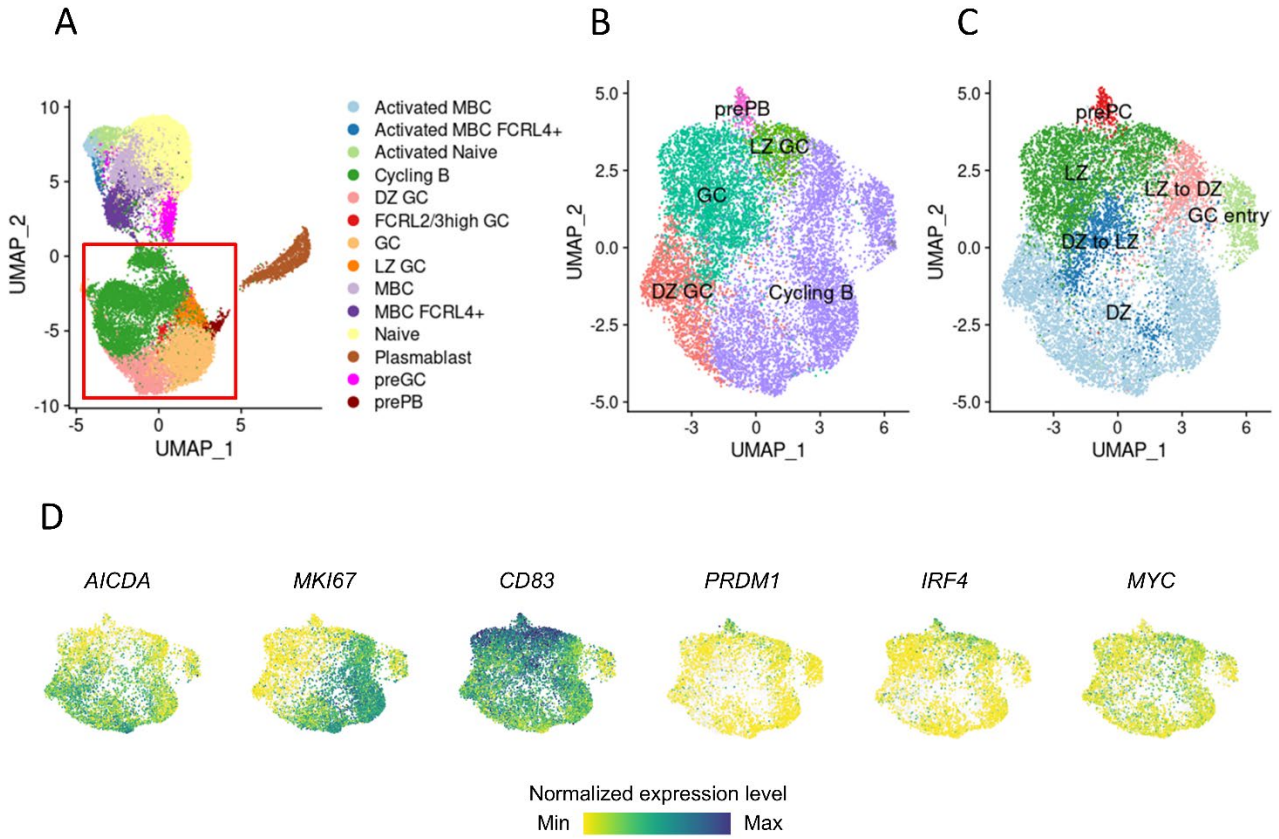
968



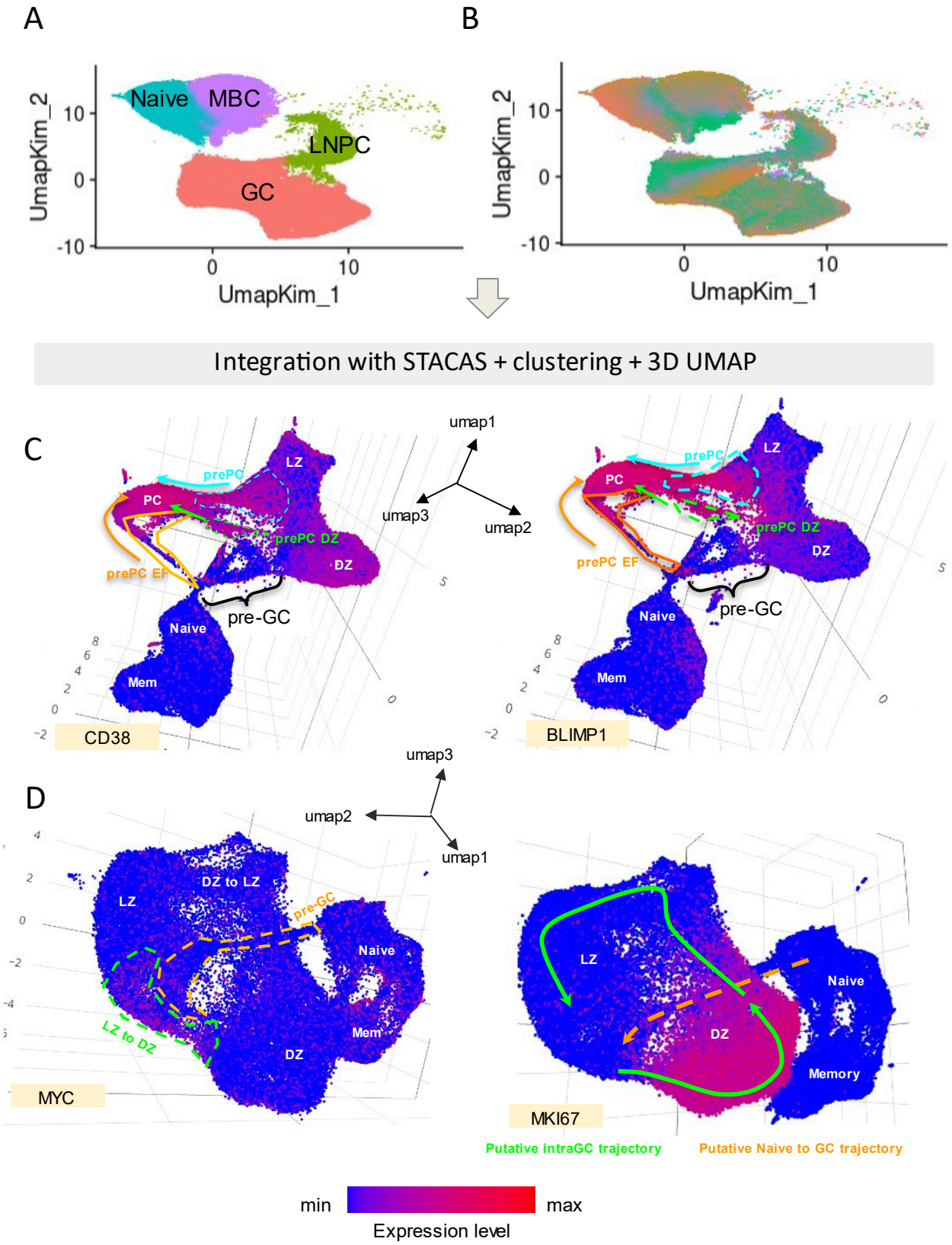
969
970 **Supplementary Figure 5** | Construction of a new murine AID mutation model. **A)** Schematic representation of scBCRseq experiment
971 in PP dataset from Chen *et al.* Only Germ-free mice (GF) sequences were kept to build the mutation model since microbiota antigen
972 binding does not confer a selective advantage to Peyer's patches GC B cells. **B)** SHM targeting patterns inferred with ShazaM tool.
973 Letters in the lower parts correspond to 5-mer context of targeted nucleotide. Well described mutation hotspots and coldspots are
974 depicted in red and blue respectively, as a quality control.
975



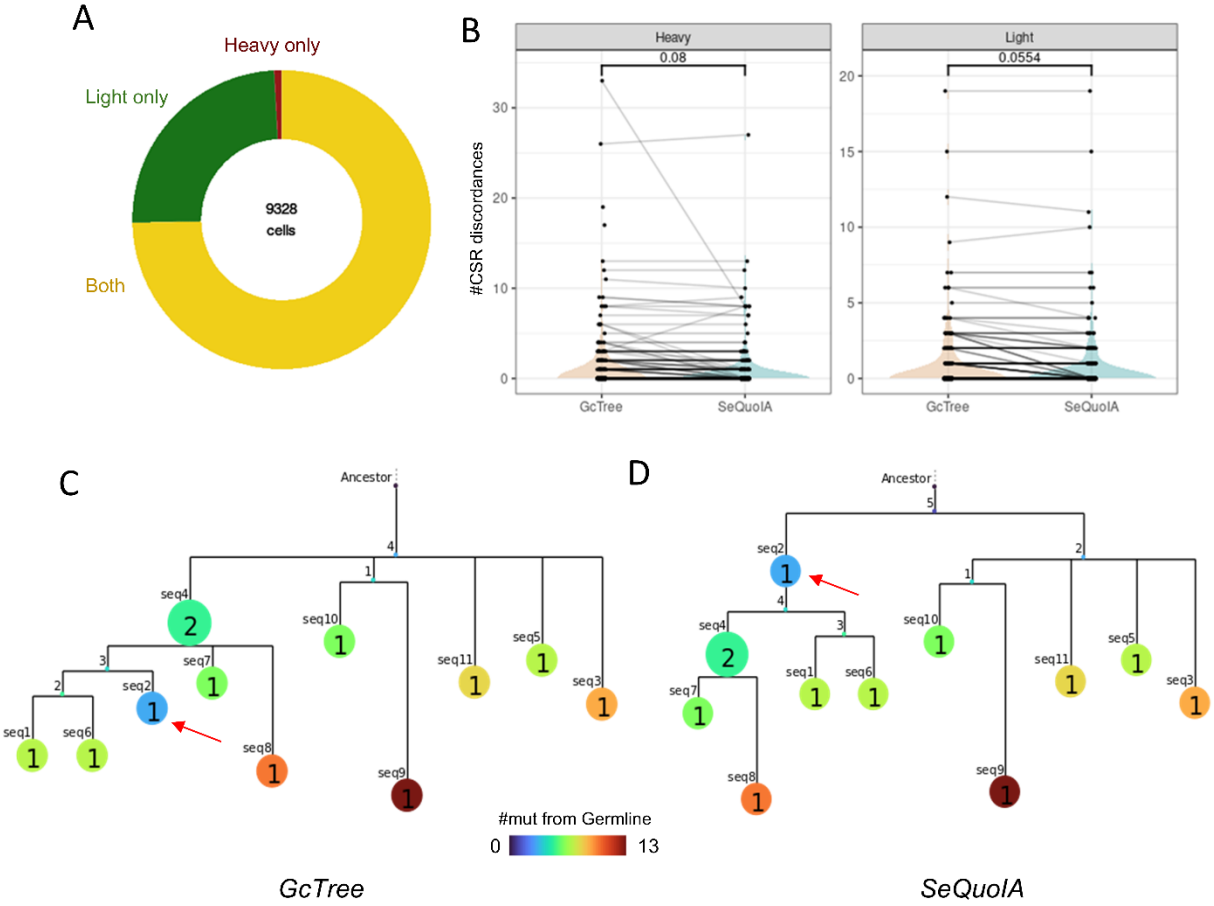
Supplementary Figure 6 | scRNA-seq annotation of NP-KLH1 dataset. **A)** UMAP projection of scRNAseq data colored by original annotations from Binet et *al*. **B)** and by our cluster-based annotation. **C)** Feature plots of marker genes. Gray dots correspond to cells with zero counts for the considered gene.



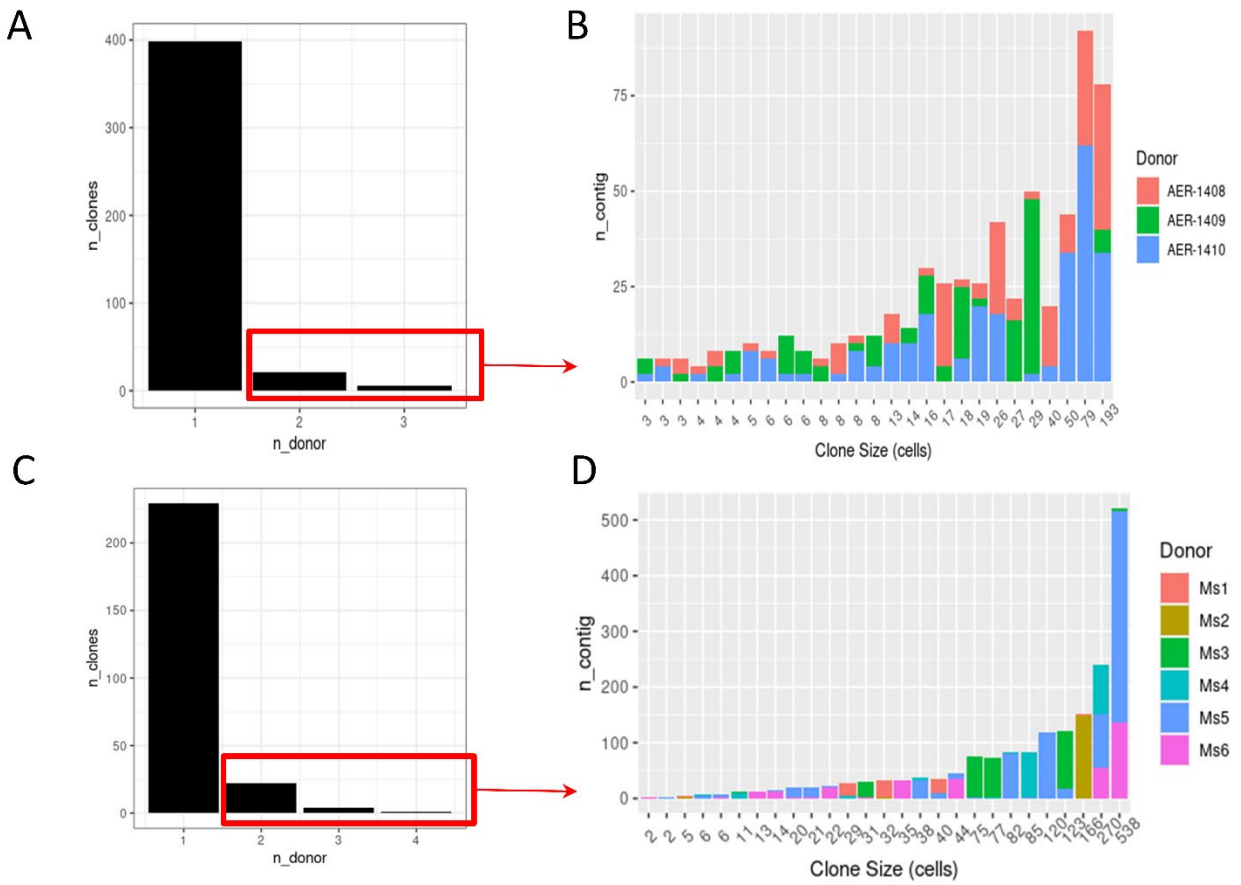
Supplementary Figure 7 | scRNAseq annotation of human tonsil dataset. **A)** UMAP projection of tonsillar B cells after cell cycle regression. Cell color corresponds to original annotation from King et *al*. Cycling and GC B cells were subsetted (red square) for further annotation **B)** UMAP projection of subsetted cells with original annotations. **C)** Same as B) with cluster-based reannotations. **D)** Feature plots of marker genes expression.



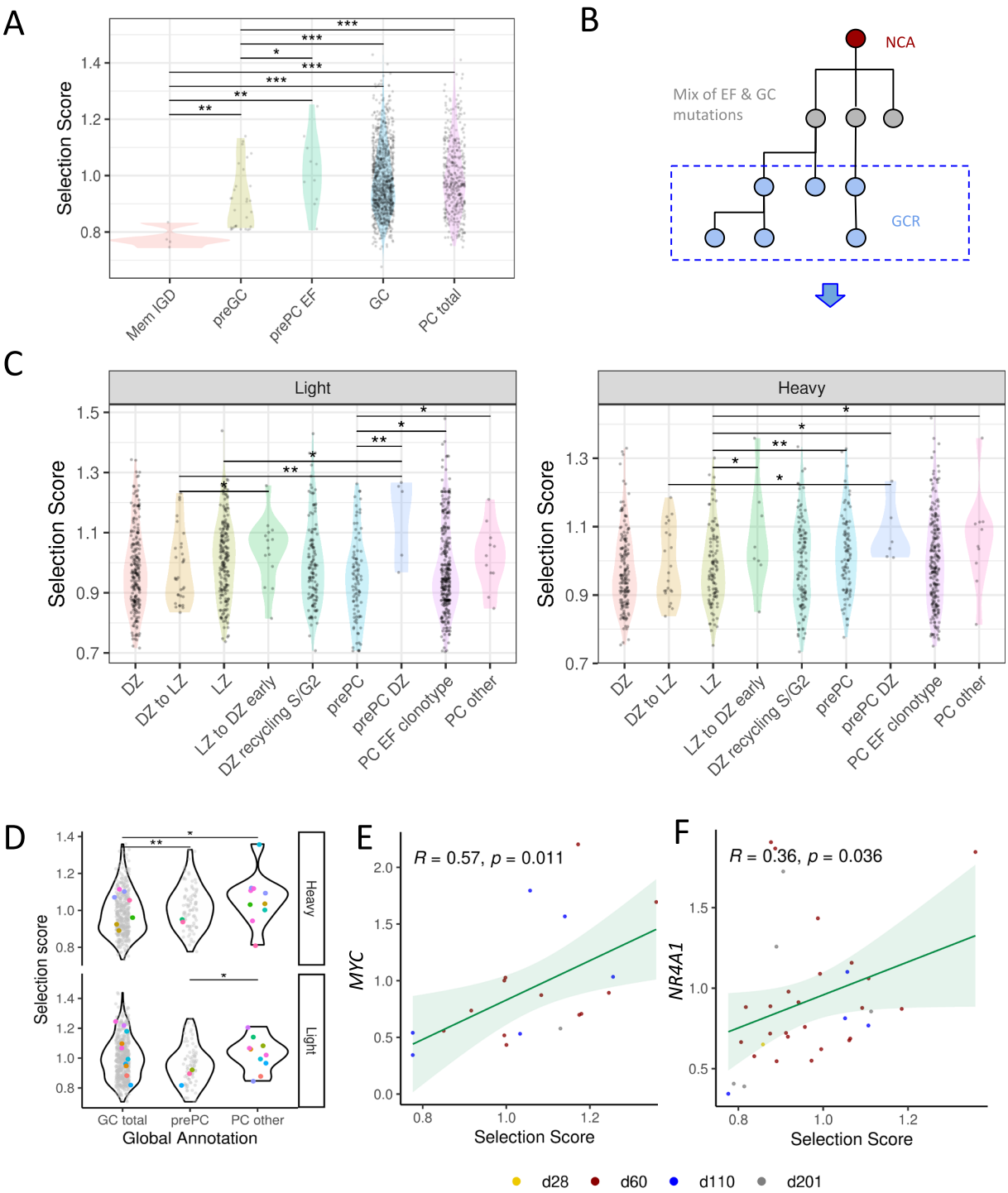
Supplementary Figure 8 | scRNAseq annotation of human Covid dataset. **A)** original UMAP projection and annotations of scRNAseq data from Kim et *al*. **B)** Original UMAP projection coloured by donor. **C)** prePC annotation. Three intermediary clusters bridging PCs to other clusters are highlighted in the 3D UMAP. CD38 expression levels support GC-experienced cells identification. BLIMP1 expression supports PC and pre-PC identification. **D)** LZ to DZ cluster annotation and heterogeneity. NB: putative trajectories were conjectured based on current knowledge of EF and GC dynamics.



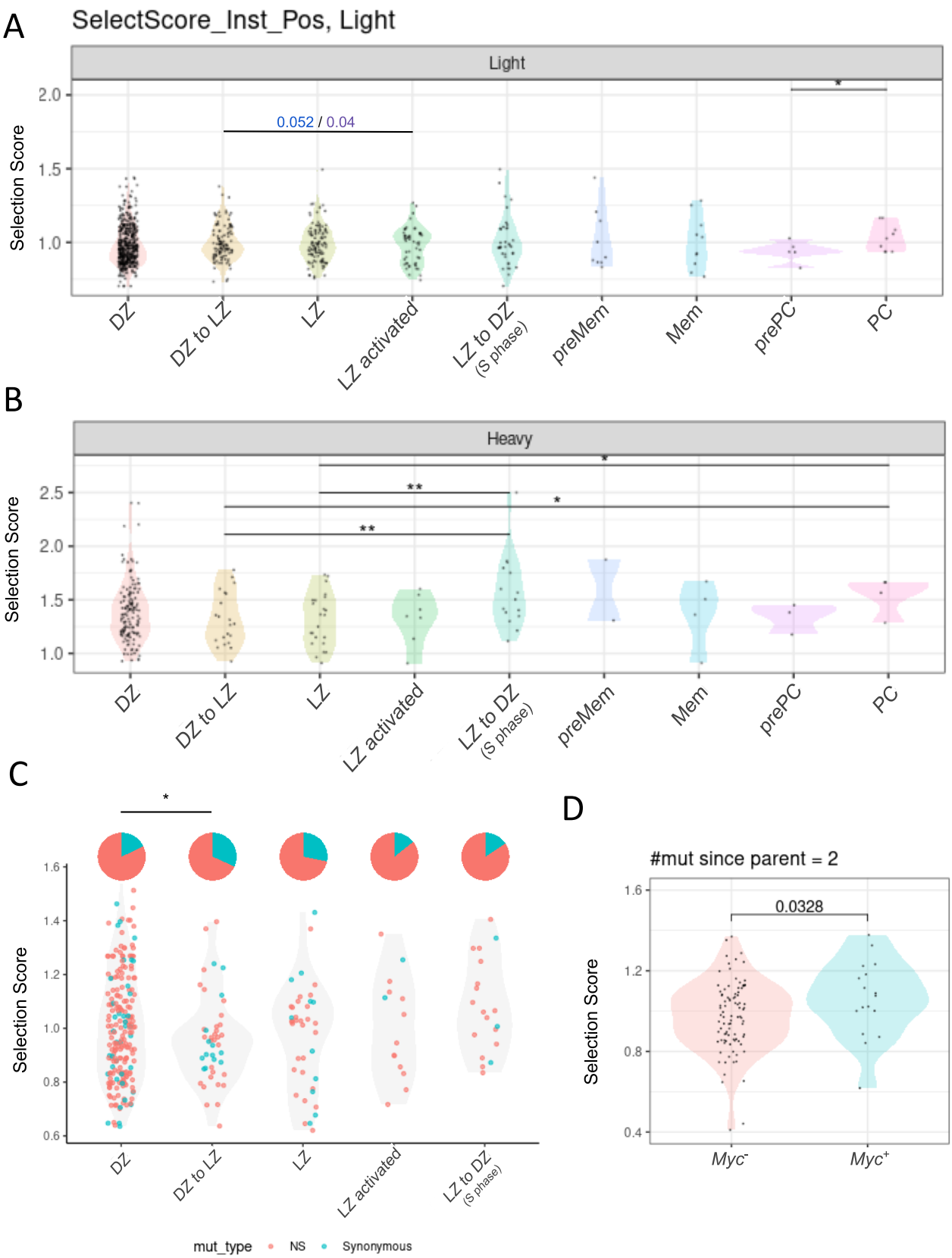
Supplementary Figure 9 |Complementary to Figure 2. **A)** Chain availability at the single cell level in mouse NP-KLH1 dataset. **B)** Optimization of CSR order in phylogenetic tree selection. P values denote paired Wilcoxon rank sum test. **C)** Mutation load of BCR sequences in a tree selected with GcTree. The red arrow highlights an aberration in mutational load. **D)** Same as C) with the tree selected by SeQuoIA for the same clonotype. The aberration (red arrow) has been corrected with SeQuoIA.



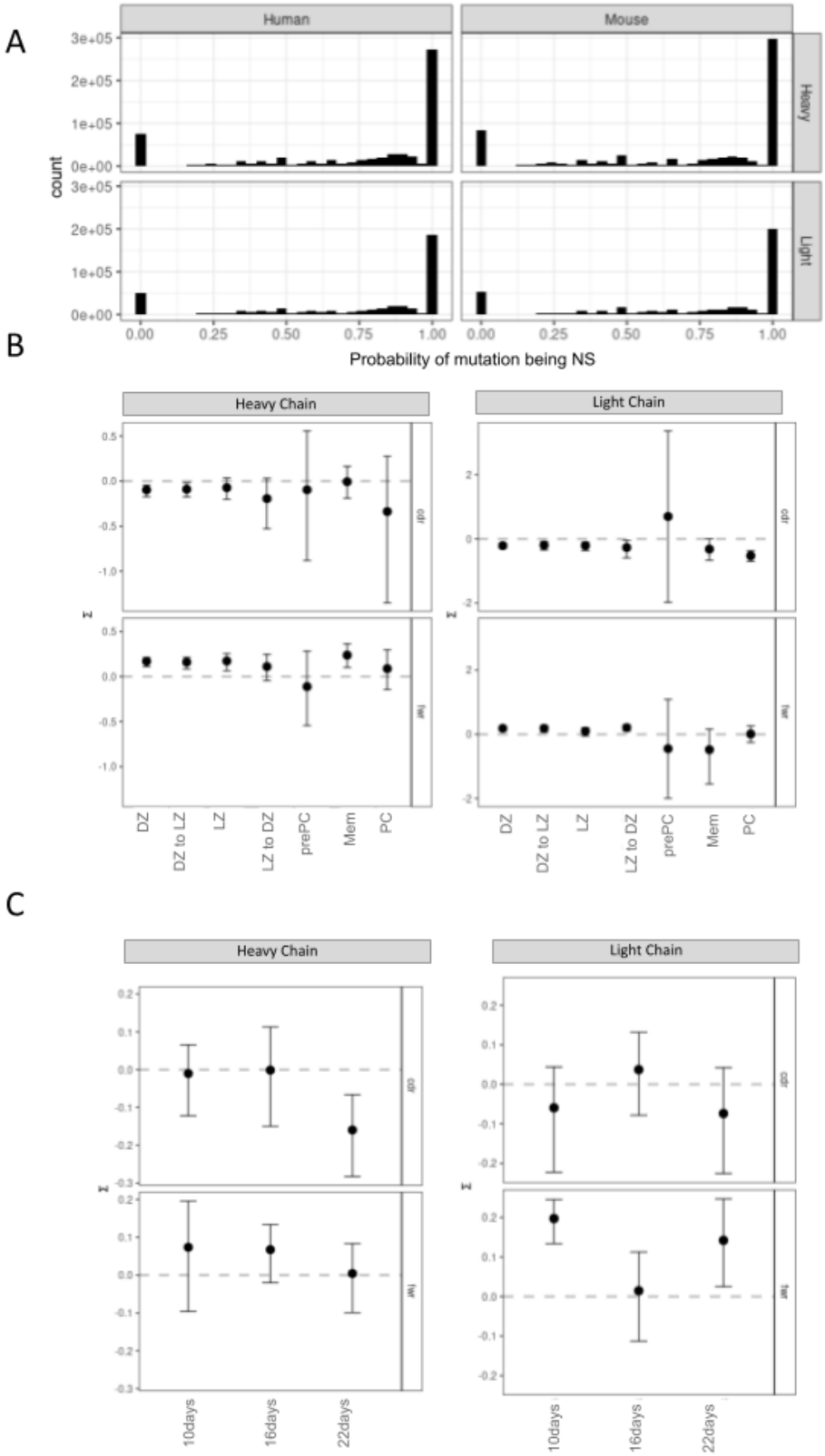
Supplementary Figure 10 | Clonotype assignation errors and optimization **A)** Number of clones featuring sequences from 1 or more donors. Clonotype assignation was performed on full BCR dataset with empirical 80% CDR3 similarity threshold in a pauciclonal dataset (mouse NP-KLH1). **B)** Barplot of sequence origin in clonotypes shared among several donors. **C) D)** Same as A-B in a polyclonal mouse influenza dataset.



Supplementary Figure 11 Application of SeQuoIA in human Covid dataset **A)** Heavy chain selection score distribution over general B cell subsets in LNs **B)** Schematic of tree-based enrichment of B cell undergoing GCR among total mutated B cells in human LNs. This strategy allows notably to discard Myc+CD38- B cell entering GCs clustering with LZ to DZ Myc+CD38+ cells **C)** Heavy and light chain selection score distribution across scRNAseq GC B cells clusters. NB: d60 time point was selected for this analysis as it makes up the major part of the whole dataset. Clonotypes with more than 5 aberrations in phylogenetic trees were excluded. PCs sharing the same clonotypes as EF prePCs were annotated as a separate cell subset. Remaining PCs should thus be enriched in GC-derived cells. Each dot represent a BCR of a given phenotype 1 or 2 mutations away from its parent. **D)** Unique BCR sequence selection score (heavy and light chains) comparison between putative GC-derived PCs and total GCB cells. Clonotypes present in PC compartment are highlighted with different colors. **E)F)** Spearman correlation between gene expression level and heavy chain selection score. All p-values denote wilcoxon rank-sum tests. All timepoints were taken for this analysis and indicated with different colours.

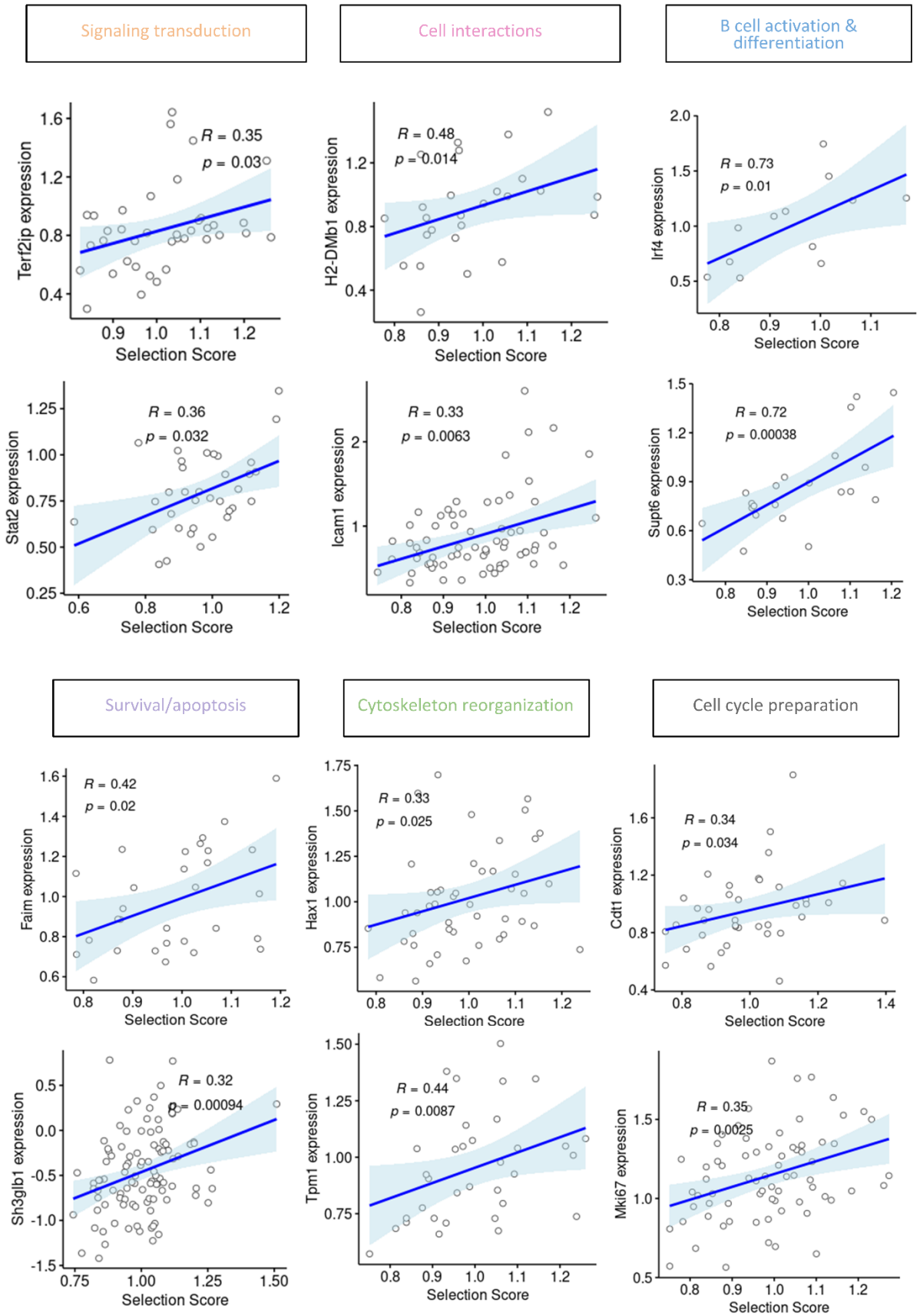


Supplementary Figure 12 | Complementary to Figure 3 **A)** Light Chain selection score distribution across LN B cell subsets. p-values in blue correspond to Kolmogorov-Smirnov distribution comparison tests and p-values in purple correspond to a proportion (of scores >1) test. **B)** Same as fig. 3.B considering only non synonymous mutations. **C)** Synonymous and NS mutation ratios and heavy chain selections scores across subsets. P-value corresponds to one-sided binomial test on proportions (of NS mutations). **D)** Heavy chain selection scores comparison between *Myc*+ and *Myc*- cells in the LZ. p-values in black denote wilcoxon rank sum test on means. All figures correspond to mouse NP-KLH1 dataset.

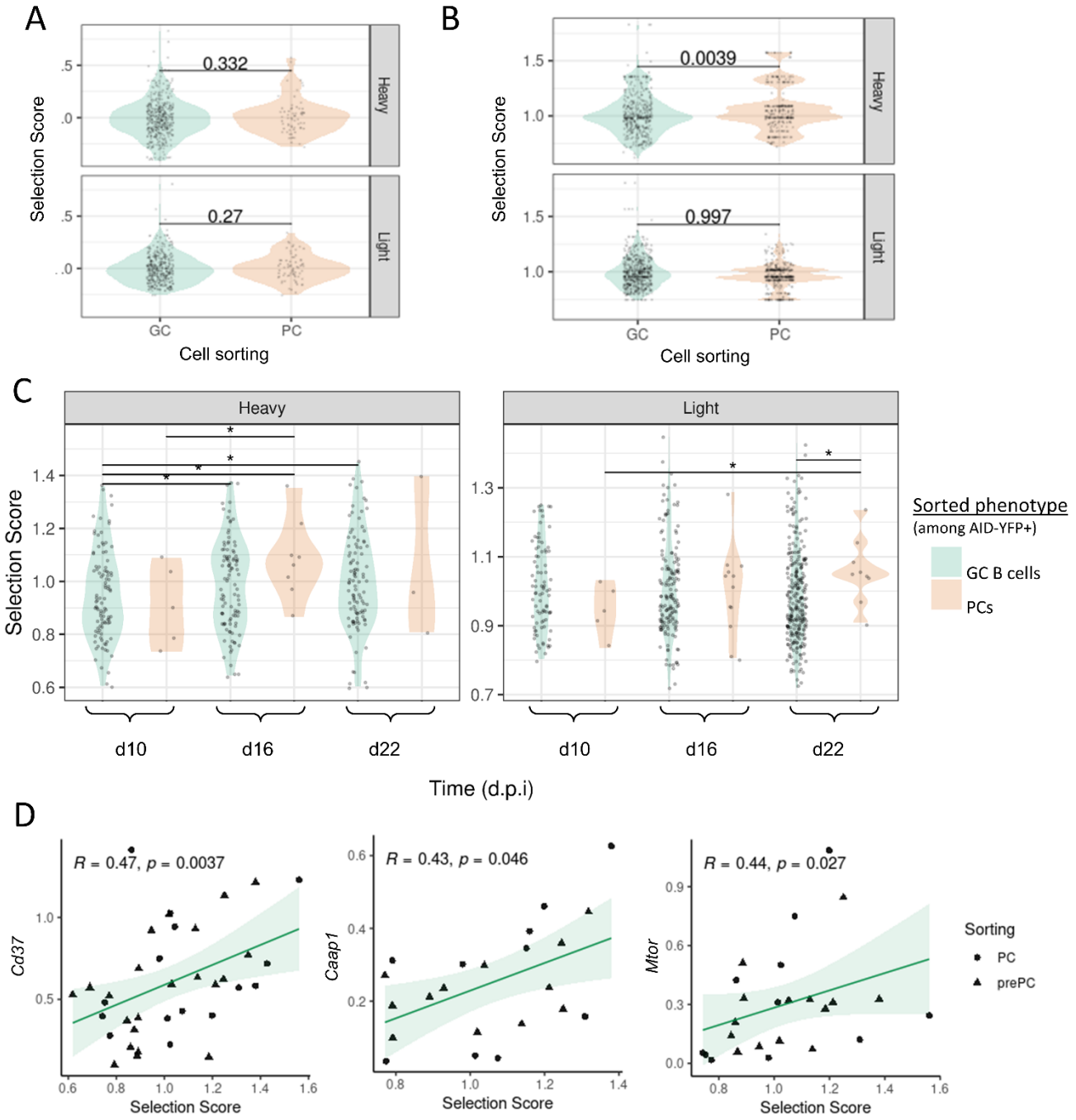


Supplementary Figure 13 | Performance of BASELINe. **A)** NS mutation probabilities distribution across codon sites in the absence of selection in human tonsil (left) and mouse NP-KLH1 (right) datasets **B)** BASELINe selection quantification across B cell subsets in mouse NP-KLH1 dataset. **C)** BASELINe selection quantification across post-immunization time points in mouse NP-KLH3 dataset.

1041

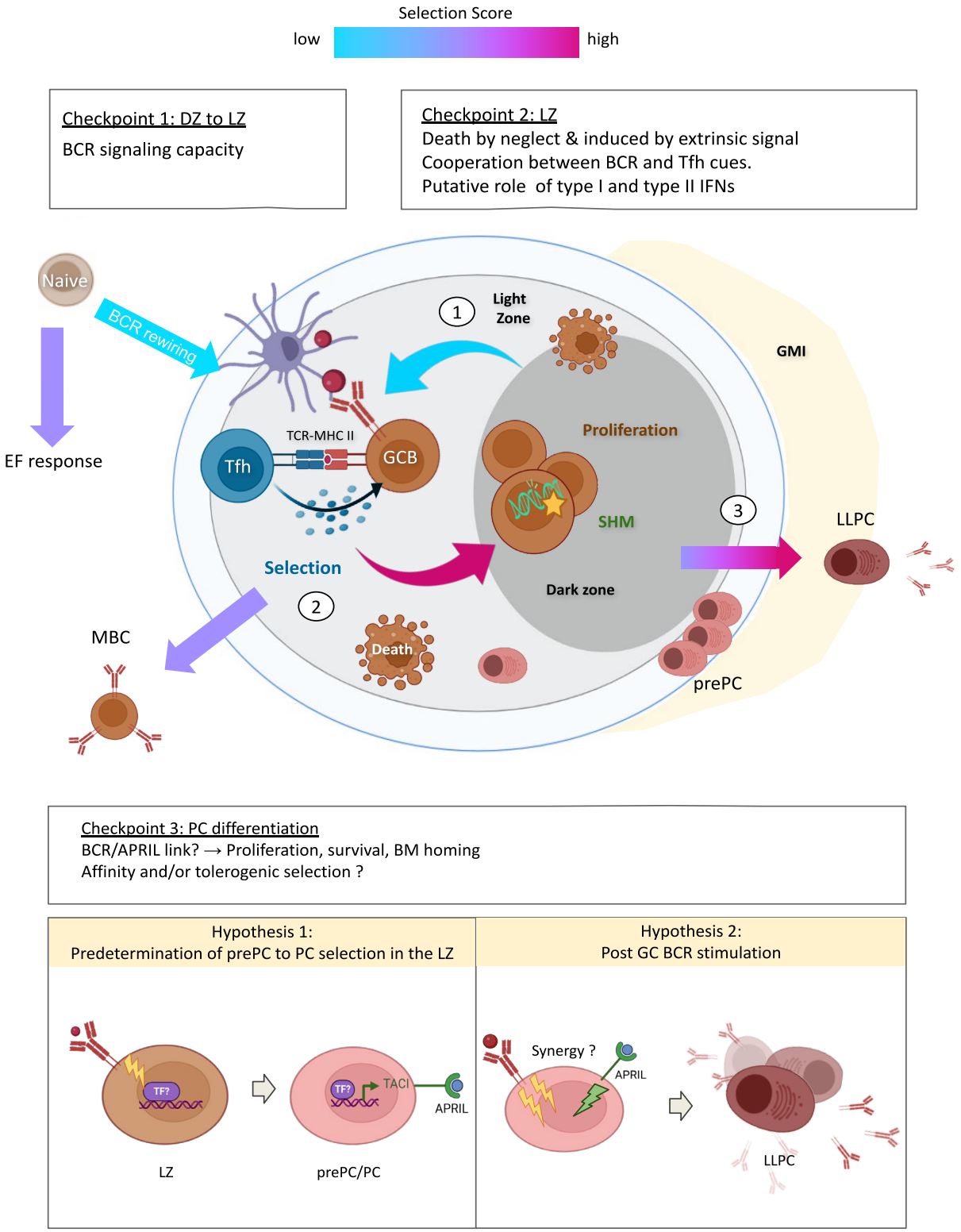


1042
1043 **Supplementary Figure 14** | Candidate genes that could provide a selective advantage to GCB cells based on BCR properties. Spearman
1044 correlation between light chain BCR selection score (1 mut away from parent) and LZ gene expression from GO pathways from Figure
1045 4.
1046



Supplementary Figure 15 | PC selection study across datasets. **A)** PC and GC B selection scores in mouse influenza dataset. Unique sequences are taken. P-values denote wilcoxon ranked sum test. **B)** Same as A) with total BCR sequences so as to consider proliferative effect **C)** PC and GC B cell selections scores distribution at different timepoints in NP-KLH3 dataset. Unique Ig sequences were taken. P-values denote wilcoxon rank sum tests (uncorrected) **D)** Spearman correlation between prePCs/PCs selection scores and selected genes in NP-KLH3 dataset (enriched for these two cell types). NB: score correlation was extended to 2 mutations since parent for Caap1

1056



1057
1058 **Supplementary Figure 16** | Working model of GC or GC-derived B cell selection. Circled numbers correspond to successive selection
1059 bottlenecks identified with selection score distribution across scRNAseq clusters. GMI = GC/medulla interface. LLPC = long lived
1060 plasma cell.
1061