mAbs 7:3, 516–524; May/June 2015; Published with license by Taylor & Francis Group, LLC

Reconstructing and mining the B cell repertoire with ImmunediveRsity

Bernardo Cortina-Ceballos[1,§], Elizabeth Ernestina Godoy-Lozano[1,§], Hugo Sámano-Sánchez[1,§], Andrés Aguilar-Salgado[1], Martín Del Castillo Velasco-Herrera[2], Carlos Vargas-Chávez[2], Daniel Velázquez-Ramírez[1,†], Guillermo Romero[1], José Moreno[1,‡], Juan Téllez-Sosa[1], and Jesús Martínez-Barnetche[1,*]

[1]Centro de Investigación Sobre Enfermedades Infecciosas; Instituto Nacional de Salud Pública (CISEI-INSP); Cuernavaca, Morelos, México; [2]Winter Genomics; D.F., México

[†]Present affiliation: Universidad del Istmo; Juchitán de Zaragoza; Oaxaca, México

[‡]Present affiliation: Dirección de Investigación; Hospital Juárez de México; D.F., México

[§]These authors equally contributed to this work.

Keywords: high-throughput sequencing, Ig repertoire, CDR3, data mining

Abbreviations: HEL, hen egg lysozyme; CDRH3, heavy chain complementarity determining region 3; Rep-Seq, repertoire sequencing; SHM, somatic hypermutation.

The B cell antigen receptor repertoire is highly diverse and constantly modified by clonal selection. High-throughput DNA sequencing (HTS) of the lymphocyte repertoire (Rep-Seq) represents a promising technology to explore such diversity *ex-vivo* and assist in the identification of antigen-specific antibodies based on molecular signatures of clonal selection. Therefore, integrative tools for repertoire reconstruction and analysis from antibody sequences are needed. We developed ImmunediveRity, a stand-alone pipeline primarily based in R programming for the integral analysis of B cell repertoire data generated by HTS. The pipeline integrates GNU software and in house scripts to perform quality filtering, sequencing noise correction and repertoire reconstruction based on V, D and J segment assignment, clonal origin and unique heavy chain identification. Post-analysis scripts generate a wealth of repertoire metrics that in conjunction with a rich graphical output facilitates sample comparison and repertoire mining. Its performance was tested with raw and curated human and mouse 454-Roche sequencing benchmarks providing good approximations of repertoire structure. Furthermore, ImmunediveRsity was used to mine the B cell repertoire of immunized mice with a model antigen, allowing the identification of previously validated antigen-specific antibodies, and revealing different and unexpected clonal diversity patterns in the post-immunization IgM and IgG compartments. Although ImmunediveRsity is similar to other recently developed tools, it offers significant advantages that facilitate repertoire analysis and repertoire mining. ImmunediveRsity is open source and free for academic purposes and it runs on 64 bit GNU/Linux and MacOS. Available at: https://bitbucket.org/ImmunediveRsity/immunediversity/

Introduction

Adaptive immunity relies on a highly diverse lymphocyte antigen receptor repertoire, which is dynamically shaped by endogenous and exogenous antigens by means of clonal selection. T and B lymphocyte receptor repertoires are generated by somatic recombination of germline V, D and J segments in an antigen-independent manner.[1] Each lymphocyte bears a unique antigen receptor that can be clonally selected. In the case of B lymphocytes, antigen receptors can be further diversified in an antigen-dependent manner by somatic hypermutation of clonally expanded lymphocytes.[2,3] Thus, the basic unit of the B cell repertoire is the idiotype, which refers to a unique antigen receptor structure (H and L chain pair) with unique epitope specificity. One or more idiotypes derived from the same VDJ recombination event and $H^{+}L$ pairing represents a clonotype. The degree of idiotypic diversification within a clonotype is indicative of antigen-mediated selection.

*Correspondence to: Jesús Martínez-Barnetche; Email: jmbarnet@insp.mx
Submitted: 12/11/2014; Revised: 02/20/2015; Accepted: 02/24/2015
http://dx.doi.org/10.1080/19420862.2015.1026502

High-throughput DNA sequencing (HTS) of lymphocyte antigen receptor repertoire (Rep-Seq)[4] provides a technological solution to measure repertoire diversity, both to track the clonal responses to antigenic challenge and to infer antigen receptor specificity (repertoire mining), that in conjunction with other high-throughput approaches could affect our view of immunity to infection, vaccine efficacy, lympho-hematological malignancies, among others.[5] Consequently, there is an increasing demand for automated data processing tools from raw data to a full virtual repertoire reconstruction and its comparative analysis within different experimental conditions. There are a number of excellent tools such as iHMMune-align,[6] JoinSolver[7] and IgBLAST[8] for primary structural analysis of Ig sequences, as well as tools that provide detailed structural features of lymphocyte receptor sequences in massive datasets such as High-V_Quest[9] and VDJfasta.[10] However none of them integrate raw data processing, reconstruction of the ontogenic relations in terms of clonal origin and diversification by somatic hypermutation to a full qualitative and quantitative representation of the sampled repertoire.

Accurate characterization of the B cell repertoire faces the challenge of correctly identifying sequences belonging to a common clonal origin, as well as the identification of the structural relationships of idiotypes within each clonal linage, accounting for sequencing errors. Clonal structure metrics, together with mutation analysis indicative of selection, represent the basis for a quantitative statistical description of repertoire diversity and could provide the means for comparative analysis between different immunological conditions, as well as for the identification of putative antigen-specific antibodies by repertoire mining.[11,12]

Results

ImmunediveRsity is a flexible stand-alone pipeline based primarily in R programming language,[13] originally designed for 454-Roche unpaired IgH sequence data derived from libraries prepared from total RNA by 5′ RACE-PCR;[14] thus, H and L chain paring information is lost during the process. Different scripts and functions (R, Python and Perl) bundled with GNU software provide a full automated analysis per sequencing library. It uses Acacia[15] for noise correction; IgBLAST[8] for V and J segment assignment; HMMER3 (http://hmmer.janelia.org/) for complementarity-determining region 3 (CDRH3) Junction identification and USEARCH[16] for clustering sequences into Heavy Chain Clonotypes, hereafter referred to as clonotypes, and their corresponding clonally related Heavy Chain lineages, referred hereafter as lineages (**Fig. S1**). The overall algorithm used by ImmunediveRsity is depicted in **Figure 1** and described in detail in **Figure S2**. For Illumina MiSeq users, a pre-processing algorithm for paired end reads assembly using PANDAseq[17] (**Fig. S3**) is provided (***imm-Illumina***).

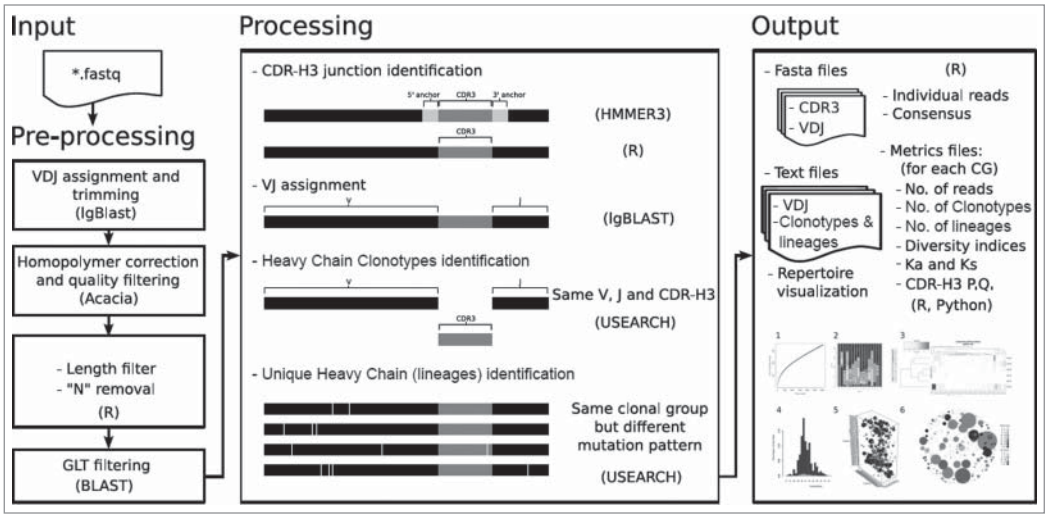

Figure 1. The overall algorithm used by ImmunediveRsity. Input file is a **.fastq*. Pre-processing consists on VDJ assignment and non-VDJ sequence trimming (5′ UTR, signal sequence and IGHC), homo-polymer correction (particularly required for 454-Roche or Ion Torrent reads), quality, size and IGH germline transcript (GLT) filters. Processing: ImmunediveRsity first identifies the CDRH3 with HMMER3 and the V and J segment rearrangement with IgBLAST. The CDRH3 reads belonging to the same V and J assignment are clustered iteratively according to sequence identity and length to define clonotypes. A second clustering step according to sequence identity is performed within the full V region of reads belonging to each clonotype to identify the lineages with different somatic hypermutation patterns. The output consists (1) *Fasta* files: containing the CDRH3 sequences for each read and clonotype, as well as the sequence for each consensus lineage with a unique identifier. (2) Text files, describing V, D and J assignments for each read and the relation of each read to a given clonotype and lineage. (3) Metrics files: for each clonotype is given frequency, the number of synonymous (Ks) and non-synonymous mutations (Ka) mutations, diversity indices, CDRH3 physico-chemical characteristics (P.Q.) and (4) Repertoire visualization: A series of predefined vectorized graphics: (1) Rarefaction curves, (2) Aminoacid composition per specific length of CDRH3, (3) Heat-map of VJ rearrengment frequencies, (4) CDRH3 spectratyping, (5) 3D cloud VDJ rearrangement frequency and (6) Network representation of the overall structure of the antibody repertoire. clonotypes (CG), lineages (Id).

Output

ImmunediveRsity generates 4 types of output: (1) *Fasta* files: containing the CDRH3 sequences for each read and clonotype, as well as the sequence for each lineage consensus. (2) Text files, describing V, D and J assignments for each read and the relation of each read to a given clonotype and lineage. (3) Metrics files: metrics of repertoire structure based on frequency, degree of diversification and somatic hypermutation that can aid the exploration of the effects of antigen-driven selection or repertoire alterations in a given disease. Such

metrics include normalized clonal and lineages frequencies, global entropy measurements such as Shannon-Weaver index[18] and Gini coefficient[19,20] (**Fig. 1** and **Fig. S2**). Such metrics can be calculated according to IGHV usage, potentially revealing hidden trends in antigen-driven clonal diversification that otherwise would not be detected only by a relative frequency analysis. Also, entropy is calculated to reveal the degree of lineage diversification within each clonotype, irrespectively of their IGHV segment usage (**Fig. S2**). Finally, the number of synonymous (Ks) and non-synonymous mutations (Ka) per lineage is calculated to indicate potential lineages under antigen-driven selection. (4) Repertoire visualization (**Figs. S4–12**): A series of predefined vectorized graphics providing frequency of V, D and J segment usage (**Figs. S4–6**), CDRH3 digital-spectratyping (**Fig. S7**), amino-acid composition at given CDRH3 length (**Fig. S8**), a heat-map of hierarchical clustering of V family usage (**Fig. S5**), rarefaction curves describing clonotype and lineage richness at a standardized sampling effort[21,22] (**Figs. S9 and S10**) and read quality before and after filtering (**Fig. S11**). In an attempt to capture the B cell repertoire complexity, an integrative graph representing a network of clonotype with their respective lineages is generated in the context of a previously described HEL-immunization experiment in mice[12] using iGraph[23] (**Fig. 2A**). These graphs can be customized to plot parameters other than hypermutation, such as diversity indices (see Methods and **Fig. S12**) or CDRH3 physicochemical properties. Finally, ImmunediveRsity provides a collection of scripts (Post-processing multi-library analysis toolbox) aimed to aid with comparisons within multiple library experiments (**Figs. 1, 4, Fig. S2**). A tool for sampling equal number of reads or clonotypes is particularly useful for such task. A tool for searching convergent CDRH3 in different individuals[22,24,25] is also provided (***find_CDR3***).

Performance of ImmunediveRsity

To test ImmunediveRsity, we used 3 benchmark data sets: (1) A mouse benchmark composed by 5,359 reads generated by

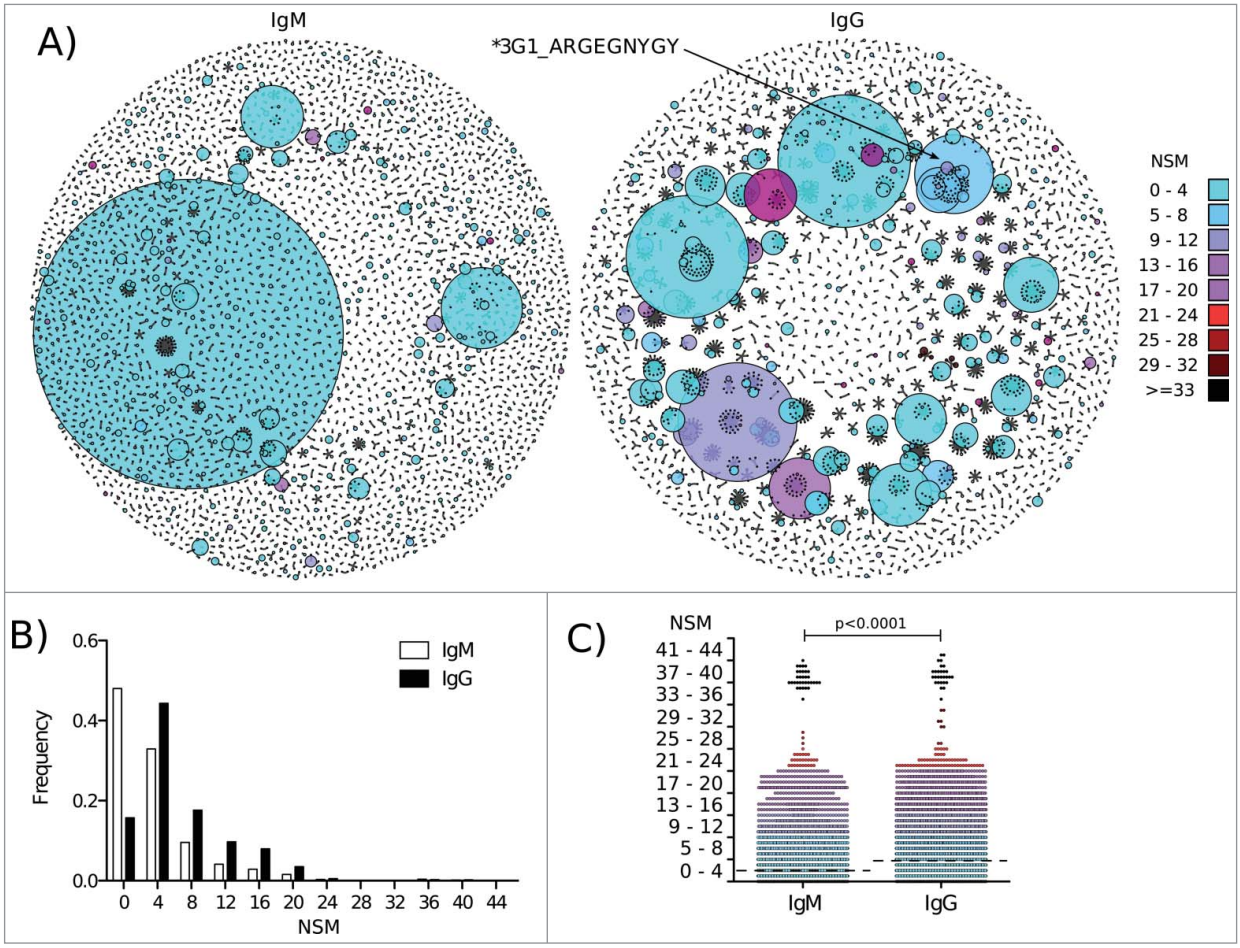

Figure 2. iGraph network representation of the sampled antibody repertoire in mouse spleen 15 days after immunization with HEL. (**A**) Left: IgM compartment. Right: IgG compartment. Each clonotype is represented by the agglomeration of lineages (nodes; represented by circles), the diameter of the circle represents the relative frequency and the color code according to the number of non-synonymous mutations. The CDRH3 sequence (*3G1 ARGEGNYGY) of recombinant HEL-specific antibody as described[12] is shown. Fading of certain clonotypes allows visualization of other clonotypes in the background. (**B**) Quantitative analysis of SHM in the IgM vs. IgG compartment in the same dataset as in **A**. (**C**) Statistical analysis of SHM in the IgM vs. IgG compartment. Median for each compartment is shown (dotted line). U-Mann-Whitney test. Frame shifted sequences were removed before the analysis. NSM, non-synonymous mutations.

sequencing a PCR amplicon library generated from a cloned 5′ RACE-PCR product derived from the spleen of a MD4 transgenic mouse (see supplementary material), which bears a monoclonal (IGHV6-3*01-IGHD4-1*01-IGHJ2*01) B cell compartment.[26] This benchmark was also used to assess ImmunediveRsity's clonotype and lineage assignment performance, such that the identification of a single clonotype and a single lineage was expected. The MD4 amplicon contains one G homopentamer and one A homotetramer within the CDRH3 region, and 3 additional homotetramers in FWR2, 3 and 4, respectively, providing a means to assess homopolymeric sequencing errors and Acacia correction performance. (2) A human benchmark composed by 1,044 sequences of a single clonotype (IGHV1-3*01-IGHD3-10*01-IGHJ3*02) manually identified from a human library (see supplementary material). To construct a reference dataset, sequences were aligned, indels were manually corrected and 10 lineages were identified according to their mutation patterns and based on the error pattern observed in the MD4 sequence data. As for the mouse benchmark, the manually curated human data set was used to evaluate ImmunediveRsity's lineage assignment accuracy using the corresponding human raw sequences as input. (3) The previously described Stanford 22 dataset[27,28] was used to assess ImmunediveRsity's clonotype assignment capacity. It consists of 13,141 human IgH sequences referred to as being derived from independent V(D)J recombination events (non- identical V, D, and J segments and non-identical V-D and D-J junction sequences). Although, absolute certainty regarding the independence of a V(D)J recombination event cannot be ascertained, the Stanford 22 data set was used as a proxy of non-clonally related sequences.

Using the MD4 mouse dataset as a proxy for sequence error calibrator, a major clonotype containing 99.5% of the reads was obtained. A second clonotype with 10 reads (0.18%) was found. Closer examination of the sequences in this clonotype determined that a G insertion in the G homopentamer located in the CDRH3 anchor created this artifact. The remaining 14 reads (0.26%) generated 8 clonotypes with low frequency and singletons (**Table 1**). At the lineage level, ImmunediveRsity identified 21 lineages, with a single lineage containing 99.3% reads, and the remaining were composed of $\leq$5 reads. Thus, based on the errors identified in the MD4 mouse data set, a minimum of 6 reads was defined as threshold to consider a true lineage (well supported). This corresponds roughly to 1 read per 1000 reads of coverage. The MD4 or other amplicons should be used as guides for calibration in terms of homopolymer content, but it does not represent all the possible sources of sequencing errors. Therefore, users are encouraged to optimize their own amplicons and means to calibrate sequencing runs.

Using this threshold, results for each dataset are shown in **Table 1**. For the human data set of 1,044 human immunoglobulin raw sequences from a single clonal origin with 10 lineages, ImmunediveRsity identified correctly a single clonotype. However, it identified 469 lineages as well, most of them singletons. Using the $\geq$6 read threshold, the number of lineages identified was 7. Therefore, if the experimental aim is to describe the clonal structure of the repertoire, we recommend not to use coverage filters. However, if the goal is to identify and analyze structural properties of an antibody subset, one should choose a minimal coverage to call a true lineage. For the Stanford22 dataset, after removal of one read with a duplicated identifier and 11 with duplicated sequences, ImmunediveRsity identified 11,779 different clonotypes and 12,421 lineages (**Table 1**), indicating that either the Stanford 22 data set contains clonally related sequences or that the CDRH3 clustering identity threshold is too low, allowing to merge non-clonally related sequences into clonotypes.

As IgBLAST is time consuming and must be executed twice, ImmunediveRsity allows parallelization according to the number of cores, which can also be customized by the user. Processing of a raw dataset of 48,350 reads described in **Figure 3** (HEL-immunized mice) takes 136 minutes or 92 minutes in 1/8 or 7/8 Intel core i7 CPU's of a 3.8 GHz and 8 Gb RAM PC computer; and 173 min in 4/4 Intel core i5 CPU's of a 2.3 GHz and 4 Gb RAM MacBook computer.

To test ImmunediveRsity performance with data derived from MiSeq Illumina sequencing, we used the human data set described by Schanz et al.[29] The whole dataset contained 2.88×10^6 paired reads, and was processed with the ***imm-Illumina*** script, which uses PANDAseq[17] for paired read assembly. Random subsampling of 1×10^4 and 1×10^5 assembled pairs were used as input for ImmunediveRsity taking 18 and 280 minutes, respectively, in a 7/8 Intel core i7 CPU's of a 3.8 GHz 8 Gb RAM PC computer.

Table 1. Overview of the reference sequencing sets

Set	Sequenced reads	After filters[1]	Observed clonotypes	Expected clonotypes	Well supported clonotypes[2]	Observed lineages (without singletons)	Expected lineages	Well supported lineages[3]
MD4	5,359	99.6%	10	1	1	21(7)	1	1
IGHV1-3	1,044	95.2%	1	1	1	469 (52)	10	7
Stanford22[4]	13,141	100%	11,779	13,141	NA	12,421	13,141	NA

[1]Percent of reads that pass the pre-processing filters.
[2]Number of clonotypes whose corresponding lineages are composed $\geq$ 6 reads.
[3]Lineages composed of $\geq$ 6 reads.
[4]The publicly available Stanford22 set was published as a set of non-clonally related immunoglobulin sequences; [28] we removed one read with a duplicated identifier and 11 with duplicated sequences.
NA, not applicable.

Critical parameters for optimization according to user needs and data type

As described in the previous sections, it is highly recommended that users include as reference a monoclonal amplicon to optimize ImmunediveRsity. Critical parameters that may require tweaking by the user are CDRH3 identity (specified by the ***id*** parameter. Default = 0.97), low frequency reads cutoff for clonotypes (specified by the ***CGfreq_cut*** parameter. Default = 6) and lineages (specified by the ***Ifreq_cut*** parameter. Default = 6). These filters are applied by default and the output is directed to different directories (WellSupportedCGs and WellSupportedIs, respectively), however unfiltered results are still available in the library output directory.

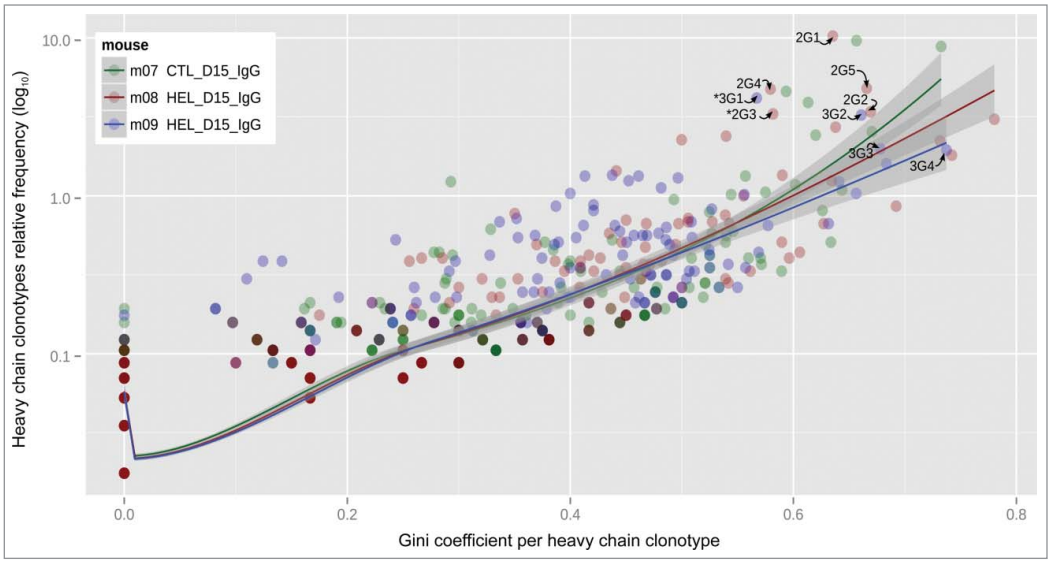

Figure 3. Comparison of clonotype intra-clonal inequality (Gini coefficient) between control (PBS) and 2 HEL-immunized mice 15 days post-immunization. clonotypes derived from control mice are shown as green dots. Clonotypes from HEL-immunized mice are shown in red (m8) and purple (m9) dots. x axis; Gini coefficient per clonotype, y axis; clonotype relative frequency. The CDRH3 sequences of anti-HEL recombinant antibodies as described[12] are shown. Fading dots allow the visualization of overlapping dots.

Differences in the B cell repertoire after an immune challenge

Immunization with protein antigens induces the germinal center reaction characterized for an extensive antigen-specific B cell proliferative response, somatic hypermutation and class switching.[30] In mice, the germinal center reaction induced by immunization peaks after 10–12 days.[31] As a proof of principle of ImmunediveRsity performance to digitally reconstruct the antibody repertoire with experimental data, we used our previously described IGHV sequence data derived from libraries (IgM and IgG class) from 2 BALB/c mice spleens at 3, 7 and 15 days post-immunization with hen egg lysozyme (HEL) and one mouse inoculated with PBS in each time point as controls.[12] The sequencing metrics of the 18 sequenced libraries are shown on **Table S1**. As expected, the proportion of somatic hypermutation was higher in the IgG than in the IgM compartment (**Fig. 2B, C**).

Fifteen days after immunization, 2 HEL-specific IGHV B cell clones, functionally validated in a previous work,[12] were identified as clones with high relative frequency, but lower Gini coefficients compared to the corresponding high relative frequency clones from control mice, (**Fig. 3**), suggesting that within the higher frequency range (y axis), antigen-specific clones are the ones having lower Gini coefficient values. Lower Gini coefficients make biological sense based on the assumption that affinity maturation would allow the selection of different lineages leading to a more even distribution than in the non-selected clonotypes. Further research is needed to clarify if Gini coefficient per clonotype is a useful measure to identify clonotypes undergoing antigen-mediated selection.

The germinal center reaction is initiated by a small number (1-8) of antigen-selected IgM expressing-B cells.[31,32] Due to the extensive oligoclonal proliferation of founder cells and their concomitant differentiation to IgG switched high affinity antibody secreting cells, a reduction in clonal diversity would be expected after immunization. We used ImmunediveRsity to sample a fixed number of reads per library (5,700) and to track clonal and idiotypic diversity induced by HEL immunization 3, 7 and 15 days post-immunization in the IgM and IgG compartment. Sampling was necessary to compare equal numbers of reads. The number of reads was based on the library with the lowest number of reads (**Table S1**). No changes in clonal and idiotypic entropy (Shannon-Weaver index), or in clonal and idiotypic inequality (Gini coefficient) were detected at day 3 and 15 post-immunization (**Fig. 4**). However, a reduction in clonal and idiotypic diversity was observed in the IgM compartment at day 7 post-immunization. In contrast, clonal and idiotypic diversity was markedly increased in the IgG compartment (**Fig. 4A**). Consistently but in the opposite direction, clonal and idiotypic inequality in the IgM compartment increased at day 7 post-immunization, whereas, in the IgG compartment, they decreased (**Fig. 4B**). The entropy measures suggest an increase in the circulation of non-clonally related IgG+ B cells at day 7.

Discussion

The lymphocyte antigen receptor repertoire is a fascinating biological system that represents the basis for acquired immunity. Hence, its analysis is critical to understand responses to vaccination and autoimmune diseases. Antibody repertoire diversity is generated by antigen-independent somatic rearrangements of

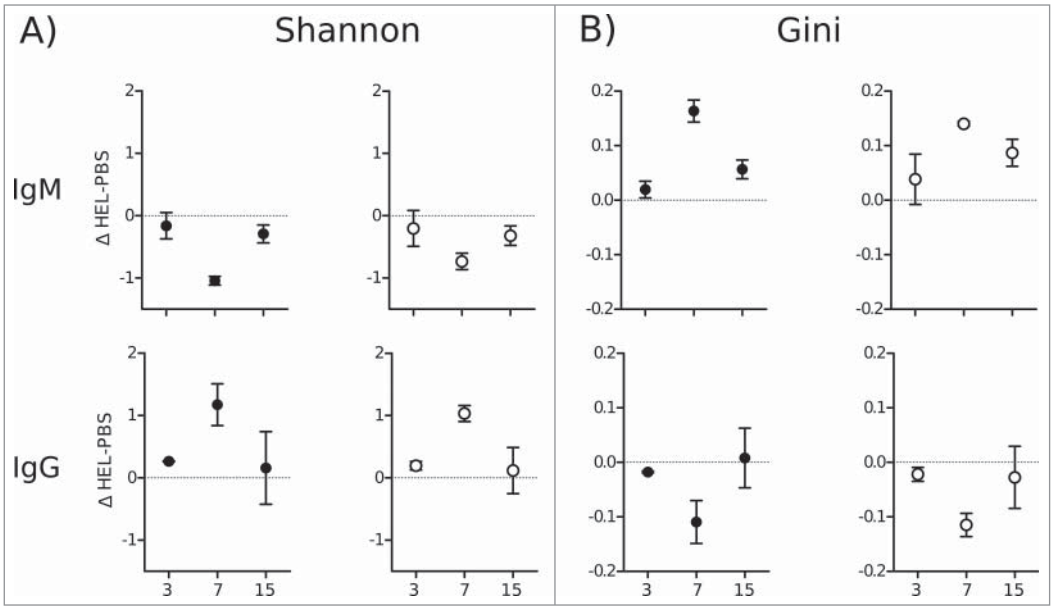

Figure 4. Clonal diversity and somatic diversification after immunization. A change in clonal (closed symbols) and lineage (open symbols) diversity measured by the average Shannon-Weaver index of HEL-immunized ($n = 2$) minus PBS-injected mouse ($n = 1$) at day 3, 7 and 15 post-immunization for the IgM (upper panel) and IgG (lower panel) compartments. (**B**) The corresponding change in clonal (closed symbols) and lineage (open symbols) inequality measured by the average Gini coefficient. For **A** and **B**, 5,700 reads per library were randomly sampled using the post-processing multi-library analysis toolbox. Sequencing metrics of the libraries used to estimate diversity measurements are described in **Table S1**.

germline V(D)J segments and antigen-dependent somatic hypermutation.[33,34] Potentially, the repertoire diversity can reach astronomical proportions. However, it is clear that structural and functional constraints, as well as clonal selection operating at different developmental stages influence its ultimate size and shape.[35] Owing to the capacity to screen a larger sample of the lymphocyte repertoire, HTS is offering the possibility to approach its complexity, and how it is affected during normal and pathological immune response.[5] Additionally, the information generated by HTS on antibody repertoires can be exploited to address higher order statistical properties of biological structures in general.[36]

Aiming toward a faithful digital reconstruction of clonal diversification as a result of somatic hypermutation, repertoire sequencing imposes particular challenges regarding data analysis. Four such challenges are: (1) All HTS platforms possess certain degree of quality to throughput trade off that can potentially overestimate true diversity.[37] To minimize the impact of sequencing errors, ImmunediveRsity attempts to correct or discard noisy reads. However, this process is not exhaustive, as exemplified by the MD4 experiment (**Table 1**). It is thus important that users tweak clustering parameters to obtain a reliable repertoire reconstruction. (2) Clonal relations for large data sets are difficult to assign due to the nature of V(D)J recombination process, which can be confounded by CDRH3 somatic hypermutation. (3) Incomplete knowledge of germline structure and population diversity at immunoglobulin loci, including allelic copy number variation.[38,39] (4) The influence of sample size sequencing depth and BCR transcript expression variation in entropy measurements, particularly in the case of cDNA library sequencing.

The development of ImmunediveRsity was motivated by the need for a stand-alone tool for digital reconstruction of the B cell repertoire and measurement of clonal diversity parameters describing functional aspects of clonal selection and to facilitate repertoire mining. There are open source tools that allow B cell repertoire analysis from HTS datasets, including clonotype reconstruction and somatic hypermutation analysis such as IgAT,[40] AbMining ToolBox[41] and Michaeli's algorithm[42] (**Table S2**). IgAT is a basic and friendly tool that does not provide clonal assignments and runs only in Windows. AbMining ToolBox is a fast, powerful tool optimized for selection of phage displayed antibodies avoiding panning steps and increasing the yield of relevant recombinant antibodies. In many ways, the most similar tool to ImmunediveRsity is the Michaeli's algorithm, which also integrates a sequencing error correction step and allows lineage identification. However, Michaeli's algorithm is designed to run in 32-bit computers, limiting its use. Additional contributions of ImmunediveRsity are the calculation of different entropy measurements such as the Gini and Shannon-Weaver indexes per clonotype and lineage and their corresponding partitions according to IGHV gene usage. Its utilization in the analysis of the B cell repertoire in response to immunization suggests a potential application of the Gini coefficient in repertoire mining and *in silico* identification of antigen-specific clones (**Fig. 3**), but further experimentation is required. Moreover, entropy measurements provided the observation that clonal and somatic diversity in the IgG compartment increases at day 7 post-immunization instead of decreasing (**Fig. 4**), supporting the notion that germinal centers are open structures that can be colonized by unrelated B cells.[43-46] The inverse relationship between Shannon entropy and Gini coefficient indicate that both measures are redundant. However, we have explored such correlation at the clonotype level in human peripheral blood of patients infected with dengue virus and found very weak correlation (unpublished data). The significance of the correlation between both measures warrants further research.

An additional feature that contributes to an integrative analysis of the B cell repertoire is ImmunediveRsity's rich output in

terms of a diverse and informative graphical output and a detailed description per clonotype sequence and structure. In conclusion, ImmunediveRsity is a highly modular and customizable tool for B cell repertoire analysis that can accelerate data integration and discovery of biologically relevant processes, as well as the identification of antibodies with potential biotechnological applications.

Methods

Input

ImmunediveRsity accepts **.fastq* files in reverse complement; thus, in principle it is suitable for most HTS platforms. Currently, it can only process VH libraries from human and mouse. It should be noted that somatic hypermutation can occur along the whole variable region, which is 400 bp on average. Platforms that generate longer reads lengths can offer better representation of somatic diversity (**Fig. 1**).

Pre-processing

As an initial step, IgBLAST[8] is performed for each individual read to map the V(D)J region and to trim the 5′ and 3′ flanks. This step is particularly important for the analysis of libraries generated by 5′ RACE-PCR to exclude germ-line transcripts (sterile transcripts) that may be present in libraries derived from total RNA, i.e., when using 5′ RACE-PCR (unpublished observations), and to trim the 5′ UTR and signal peptide sequences and possible IGHC nucleotides, which are irrelevant for antigen binding and thus for clonal selection. Moreover, the distal flank to the sequencing primer usually will have the lowest quality. As 454-Roche and Ion Torrent sequencing platforms are prone to indels in homopolymeric regions,[47] ImmunediveRsity can call Acacia,[48] an error-correction tool. Sequences below a median read quality of Q28 or below 200 bp long are discarded. These parameters can be easily customized (**Fig. 1**).

CDRH3 Junction identification

The CDRH3 Junction region is delimited by the 3′ end of the V gene and the 5′ end of the J gene, including the D gene and the N and P-nucleotides.[49] The CDRH3 region is the most variable and defines the clonal origin. The CDRH3 region has conserved regions (anchors) in both flanks, hence ImmunediveRsity uses HMMER3 with a HMMER DNA profile trained to match either human or mouse CDRH3-coding sequence for each read. HMMER3 retrieves the CDRH3 by means of 2 anchors, which are trimmed on the basis of the conserved motifs Tyr-Phe/Tyr/His-Cys or Phe-Phe-Cys in the 3′ end of the IGHV segment (IMGT positions 102-104) and the motif Trp-Gly in the IGHJ segment. Additionally, it considers the potential presence of indels due to homopolymeric errors in the sequence coding for Tyr-Phe-Cys. The presence of D segments in tandem, which was previously described,[50,51] are also considered since our second step returns the largest CDRH3 with the different possible motifs. Reads without an identifiable CDRH3 are discarded (**Fig. 1**).

V(D)J assignment

A critical step toward clonal origin identification is the determination of V, D and J segment usage. Different tools are publicly available to achieve this task. The main challenge relies on correctly identifying the D segment, because of its usually short length (from 11 to 37 nt)[52] and mutations content by exonuclease activity during the recombination process. To assign the V, D and J segments, ImmunediveRsity uses IgBLAST[8] by aligning each read to the current set of functional germ line sequences from ImMunoGeneTics database (50 V genes and 239 alleles, 23 D genes and 30 alleles, and 6 J genes and 13 alleles for human)[53] (**Fig. 1**). However, this database can be upgraded as new genes and alleles are described.[39]

Heavy Chain Clonotype (clonotype) identification

Although biologically a B cell clonotype corresponds to a group of B cells sharing a unique antigen receptor originated by a unique VDJ and VJ recombinatorial event leading to the respective H and L chain pair, ImmunediveRsity interprets the clonotypes as single chained (in this case IGHV) objects. The clonotypes are defined by CDRH3 clustering with USEARCH.[54] USEARCH is fast, exhaustive and offers simplicity in setting the desired parameters. As the human D gene sequences can be as small as 11 bp,[52] as in the case of IGHD7-27, accurate D gene assignation in mature B-lymphocytes is difficult to achieve.[28] For that reason, we define that 2 reads belong to the same clonotype if the following is true: (1) The V and J gene assignment is the same in both sequences; (2) Junction regions of the 2 sequences have a nucleotide identity $\geq$ 97%. This parameter (***id***) can be adjusted according to the user needs and calibration results; and (3) The trimmed length of the shorter Junction sequence $\geq$ 97% of the length of the larger Junction sequence (**Fig. 1**). CDRH3 length identity and length parameters used for clustering were determined by estimating the accuracy obtained by sequencing an IgH amplicon derived from an anti-HEL transgenic mouse (MD4) that is virtually monoclonal[26] and a manually-curated human immunoglobulin set (IGHV1-3) (see Results. Performance of ImmunediveRsity. **Table 1**), but can be easily customized.

Unique heavy chain (lineage) identification

As for clonotypes, ImmunediveRsity obviates L chain pairing and interprets a lineage as an object derived from a unique VDJ recombinatorial event and a unique SHM pattern. Lineages are identified on a second iterative clustering step with USEARCH, using the complete VDJ region of each read within each clonotype.[54] Reads corresponding to the same clonotype may have different lengths due to the sequence quality decay during the sequencing process. Shorter reads generated by early sequencing termination or shortened by quality filters would be still retained in the analysis. Two reads belong to the same lineage if: (1) Both reads have the same clonotype assignment; (2) The V(D)J regions of the 2 sequences have a nucleotide identity $\geq$ 99.5%; and (3) The length of the shorter V(D)J region is $\geq$ 60% of the larger read length (**Fig. 1**). These criteria are based on the mouse MD4

and the human IGHV1-3 sets (see **Performance of ImmunediveRsity,** benchmark data sets, **Table 1**).

Entropy and Gini coefficient measurement

The Shannon-Weiner index was calculated by the following formula: $\boldsymbol{H'} = -\sum_1^n \boldsymbol{p_i ln(p_i)}$, where p is the proportion of *i* elements (clonotypes or lineages). The Gini coefficient was calculated according to the following formula:

$$G = \frac{1}{n}\left\{\frac{2\sum_1^n [(n+1)xi]}{n\sum_1^n xi} - (n+1)\right\}$$

where *n* is the number of elements (lineages per clonotype), *xi* is the sorted proportion of each lineage within clonotype.

Post-processing multi-library analysis toolbox

ImmunediveRsity provides a collection of scripts aimed to aid comparisons within multiple library experiments (**Figs. 1, 4, Fig. S1**). A tool for sampling an equal number of reads or clonotypes is particularly useful for such task. A search tool for convergent CDRH3 in different individuals[22,24,25] is also provided.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Acknowledgments

We would like to thank Rosa Elena Gómez Barreto for support on 454-Roche sequencing and Humberto Valdovinos Torres for support with animal immunization and library preparation. The input of Robert Edgar in improving this manuscript is greatly acknowledged.

Funding

This work was funded by FOSISS-CONACyT grant # 142120 to JMB and SEP-CONACyT grant # 133765 to JTS.

Supplemental Material

Supplemental data for this article can be accessed on the publisher's website.

References

1. Tonegawa S. Somatic generation of antibody diversity. Nature 1983; 302:575-81; PMID:6300689; http://dx.doi.org/10.1038/302575a0
2. Cohn M, Mitchison NA, Paul WE, Silverstein AM, Talmage DW, Weigert M. Reflections on the clonal-selection theory. Nat Rev Immunol 2007; 7:823-30; PMID:17893695; http://dx.doi.org/10.1038/nri2177
3. Jackson KJ, Kidd MJ, Wang Y, Collins AM. The shape of the lymphocyte receptor repertoire: lessons from the B cell receptor. Front Immunol 2013; 4:263; PMID:24032032; http://dx.doi.org/10.3389/fimmu.2013.00263
4. Benichou J, Ben-Hamo R, Louzoun Y, Efroni S. Rep-Seq: uncovering the immunological repertoire through next-generation sequencing. Immunology 2012; 135:183-91; PMID:22043864; http://dx.doi.org/10.1111/j.1365-2567.2011.03527.x
5. Georgiou G, Ippolito GC, Beausang J, Busse CE, Wardemann H, Quake SR. The promise and challenge of high-throughput sequencing of the antibody repertoire. Nat Biotechnol 2014; 32:158-68; PMID:24441474; http://dx.doi.org/10.1038/nbt.2782
6. Gaeta BA, Malming HR, Jackson KJ, Bain ME, Wilson P, Collins AM. iHMMune-align: hidden Markov model-based alignment and identification of germline genes in rearranged immunoglobulin gene sequences. Bioinformatics 2007; 23:1580-7; PMID:17463026; http://dx.doi.org/10.1093/bioinformatics/btm147
7. Souto-Carneiro MM, Longo NS, Russ DE, Sun HW, Lipsky PE. Characterization of the human Ig heavy chain antigen binding complementarity determining region 3 using a newly developed software algorithm, JOINSOLVER. J Immunol 2004; 172:6790-802; PMID:15153497; http://dx.doi.org/10.4049/jimmunol.172.11.6790
8. Ye J, Ma N, Madden TL, Ostell JM. IgBLAST: an immunoglobulin variable domain sequence analysis tool. Nucleic Acids Res 2013; 41:W34-40; PMID:23671333; http://dx.doi.org/10.1093/nar/gkt382
9. Alamyar E, Duroux P, Lefranc MP, Giudicelli V. IMGT((R)) tools for the nucleotide analysis of immunoglobulin (IG) and T cell receptor (TR) V-(D)-J repertoires, polymorphisms, and IG mutations: IMGT/V-QUEST and IMGT/HighV-QUEST for NGS. Methods Mol Biol 2012; 882:569-604; PMID:22665256; http://dx.doi.org/10.1007/978-1-61779-842-9_32
10. Glanville J, Zhai W, Berka J, Telman D, Huerta G, Mehta GR, Ni I, Mei L, Sundar PD, Day GM, et al. Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. Proc Natl Acad Sci U S A 2009; 106:20216-21; PMID:19875695; http://dx.doi.org/10.1073/pnas.0909775106
11. Saggy I, Wine Y, Shefet-Carasso L, Nahary L, Georgiou G, Benhar I. Antibody isolation from immunized animals: comparison of phage display and antibody discovery via V gene repertoire mining. Protein Engin Design Selection 2012; 25:539-49; http://dx.doi.org/10.1093/protein/gzs060
12. Valdes-Aleman J, Tellez-Sosa J, Ovilla-Munoz M, Godoy-Lozano E, Velazquez-Ramirez D, Valdovinos-Torres H, Gomez-Barreto RE, Martinez-Barnetche J. Hybridization-based antibody cDNA recovery for the production of recombinant antibodies identified by repertoire sequencing. MAbs 2014; 6:493-501; PMID:24492293; http://dx.doi.org/10.4161/mabs.27435
13. Zhao JH, Tan Q. Integrated analysis of genetic data with R. Hum Genomics 2006; 2:258-65; PMID:16460651; http://dx.doi.org/10.1186/1479-7364-2-4-258
14. Matz M, Shagin D, Bogdanova E, Britanova O, Lukyanov S, Diatchenko L, Chenchik A. Amplification of cDNA ends based on template-switching effect and step-out PCR. Nucleic Acids Res 1999; 27:1558-60; PMID:10037822; http://dx.doi.org/10.1093/nar/27.6.1558
15. Bragg L, Stone G, Imelfort M, Hugenholtz P, Tyson GW. Fast, accurate error-correction of amplicon pyrosequences using Acacia. Nat Methods 2012; 9:425-6; PMID:22543370; http://dx.doi.org/10.1038/nmeth.1990
16. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 2010; 26:2460-1; PMID:20709691; http://dx.doi.org/10.1093/bioinformatics/btq461
17. Masella AP, Bartram AK, Truszkowski JM, Brown DG, Neufeld JD. PANDAseq: paired-end assembler for illumina sequences. BMC Bioinformat 2012; 13:31; PMID:22333067; http://dx.doi.org/10.1186/1471-2105-13-31
18. Shannon CE. The mathematical theory of communication. 1963. MD Comput 1997; 14:306-17; PMID:9230594
19. Ceriani L, Verme P. The origins of the Gini index: extracts from Variabilità e Mutabilità (1912) by Corrado Gini. J Econom Inequal 2012; 10:421-43; http://dx.doi.org/10.1007/s10888-011-9188-x
20. Bashford-Rogers RJ, Palser AL, Huntly BJ, Rance R, Vassiliou GS, Follows GA, Kellam P. Network properties derived from deep sequencing of human B-cell receptor repertoires delineate B-cell populations. Genome Res 2013; 23:1874-84; PMID:23742949; http://dx.doi.org/10.1101/gr.154815.113
21. Hughes JB, Hellmann JJ. The application of rarefaction techniques to molecular inventories of microbial diversity. Methods Enzymol 2005; 397:292-308; PMID:16260298; http://dx.doi.org/10.1016/S0076-6879(05)97017-1
22. Weinstein JA, Jiang N, White RA 3rd, Fisher DS, Quake SR. High-throughput sequencing of the zebrafish antibody repertoire. Science 2009; 324:807-10; PMID:19423829; http://dx.doi.org/10.1126/science.1170020
23. Csardi G, Nepusz T. The igraph software package for complex network research. InterJ Complex Syst 2006; 1695.
24. Krause JC, Tsibane T, Tumpey TM, Huffman CJ, Briney BS, Smith SA, Basler CF, Crowe JE, Jr. Epitope-specific human influenza antibody repertoires diversify by B cell intraclonal sequence divergence and interclonal convergence. J Immunol 2011; 187:3704-11; PMID:21880983; http://dx.doi.org/10.4049/jimmunol.1101823
25. Jackson KJ, Liu Y, Roskin KM, Glanville J, Hoh RA, Seo K, Marshall EL, Gurley TC, Moody MA, Haynes BF, et al. Human responses to influenza vaccination show seroconversion signatures and convergent antibody rearrangements. Cell Host Microbe 2014; 16:105-14; PMID:24981332; http://dx.doi.org/10.1016/j.chom.2014.05.013
26. Goodnow CC, Crosbie J, Adelstein S, Lavoie TB, Smith-Gill SJ, Brink RA, Pritchard-Briscoe H, Wotherspoon JS, Loblay RH, Raphael K, et al. Altered immunoglobulin expression and functional silencing of self-

reactive B lymphocytes in transgenic mice. Nature 1988; 334:676-82; PMID:3261841; http://dx.doi.org/10.1038/334676a0
27. Boyd SD, Marshall EL, Merker JD, Maniar JM, Zhang LN, Sahaf B, Jones CD, Simen BB, Hanczaruk B, Nguyen KD, et al. Measurement and clinical monitoring of human lymphocyte clonality by massively parallel V-D-J pyrosequencing. Sci Translat Med 2009; 1:12ra23-12ra23; PMID:20161664; http://dx.doi.org/10.1126/scitranslmed.3000540
28. Jackson KJ, Boyd S, Gaeta BA, Collins AM. Benchmarking the performance of human antibody gene alignment utilities using a 454 sequence dataset. Bioinformatics 2010; 26:3129-30; PMID:21036814; http://dx.doi.org/10.1093/bioinformatics/btq604
29. Schanz M, Liechti T, Zagordi O, Miho E, Reddy ST, Gunthard HF, Trkola A, Huber M. High-throughput sequencing of human immunoglobulin variable regions with subtype identification. PloS one 2014; 9:e111726; PMID:25364977; http://dx.doi.org/10.1371/journal.pone.0111726
30. Victora GD, Nussenzweig MC. Germinal centers. Ann Rev Immunol 2012; 30:429-57; PMID:22224772; http://dx.doi.org/10.1146/annurev-immunol-020711-075032
31. Kelsoe G. Life and death in germinal centers (redux). Immunity 1996; 4:107-11; PMID:8624801; http://dx.doi.org/10.1016/S1074-7613(00)80675-5
32. Hermans MH, Wubbena A, Kroese FG, Hunt SV, Cowan R, Opstelten D. The extent of clonal structure in different lymphoid organs. J Exp Med 1992; 175:1255-69; PMID:1569396; http://dx.doi.org/10.1084/jem.175.5.1255
33. Schatz DG, Ji Y. Recombination centres and the orchestration of V(D)J recombination. Nat Rev Immunol 2011; 11:251-63; PMID:21394103; http://dx.doi.org/10.1038/nri2941
34. Maul RW, Gearhart PJ. AID and somatic hypermutation. Adv Immunol 2010; 105:159-91; PMID:20510733; http://dx.doi.org/10.1016/S0065-2776(10)05006-6
35. Schroeder HW, Jr., Zemlin M, Khass M, Nguyen HH, Schelonka RL. Genetic control of DH reading frame and its effect on B-cell development and antigen-specifc antibody production. Critical Rev Immunol 2010; 30:327-44; PMID:20666706; http://dx.doi.org/10.1615/CritRevImmunol.v30.i4.20
36. Mora T, Walczak AM, Bialek W, Callan CG, Jr. Maximum entropy models for antibody diversity. Proc Natl Acad Sci U S A 2010; 107:5405-10; PMID:20212159; http://dx.doi.org/10.1073/pnas.1001705107
37. Kircher M, Kelso J. High-throughput DNA sequencing–concepts and limitations. Bioessays 2010; 32:524-36; PMID:20486139; http://dx.doi.org/10.1002/bies.200900181
38. Kidd MJ, Chen Z, Wang Y, Jackson KJ, Zhang L, Boyd SD, Fire AZ, Tanaka MM, Gaeta BA, Collins AM. The inference of phased haplotypes for the immunoglobulin H chain V region gene loci by analysis of VDJ gene rearrangements. J Immunol 2012; 188:1333-40; PMID:22205028; http://dx.doi.org/10.4049/jimmunol.1102097
39. Watson CT, Steinberg KM, Huddleston J, Warren RL, Malig M, Schein J, Willsey AJ, Joy JB, Scott JK, Graves TA, et al. Complete haplotype sequence of the human immunoglobulin heavy-chain variable, diversity, and joining genes and characterization of allelic and copy-number variation. Am J Hum Genet 2013; 92:530-46; PMID:23541343; http://dx.doi.org/10.1016/j.ajhg.2013.03.004
40. Rogosch T, Kerzel S, Hoi KH, Zhang Z, Maier RF, Ippolito GC, Zemlin M. Immunoglobulin analysis tool: a novel tool for the analysis of human and mouse heavy and light chain transcripts. Front Immunol 2012; 3:176; PMID:22754554; http://dx.doi.org/10.3389/fimmu.2012.00176
41. D'Angelo S, Glanville J, Ferrara F, Naranjo L, Gleasner CD, Shen X, Bradbury AR, Kiss C. The antibody mining toolbox: an open source tool for the rapid analysis of antibody repertoires. MAbs 2014; 6:160-72; PMID:24423623; http://dx.doi.org/10.4161/mabs.27105
42. Michaeli M, Barak M, Hazanov L, Noga H, Mehr R. Automated analysis of immunoglobulin genes from high-throughput sequencing: life without a template. J Clin Bioinformat 2013; 3:15; PMID:23977981; http://dx.doi.org/10.1186/2043-9113-3-15
43. Schwickert TA, Lindquist RL, Shakhar G, Livshits G, Skokos D, Kosco-Vilbois MH, Dustin ML, Nussenzweig MC. In vivo imaging of germinal centres reveals a dynamic open structure. Nature 2007; 446:83-7; PMID:17268470; http://dx.doi.org/10.1038/nature05573
44. Bende RJ, van Maldegem F, Triesscheijn M, Wormhoudt TA, Guijt R, van Noesel CJ. Germinal centers in human lymph nodes contain reactivated memory B cells. J Exp Med 2007; 204:2655-65; PMID:17938234; http://dx.doi.org/10.1084/jem.20071006
45. Schwickert TA, Alabyev B, Manser T, Nussenzweig MC. Germinal center reutilization by newly activated B cells. J Exp Med 2009; 206:2907-14; PMID:19934021; http://dx.doi.org/10.1084/jem.20091225
46. Bergqvist P, Stensson A, Hazanov L, Holmberg A, Mattsson J, Mehr R, Bemark M, Lycke NY. Re-utilization of germinal centers in multiple Peyer's patches results in highly synchronized, oligoclonal, and affinity-matured gut IgA responses. Mucosal immunol 2013; 6:122-35; PMID:22785230; http://dx.doi.org/10.1038/mi.2012.56
47. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature 2005; 437:376-80; PMID:16056220
48. Bragg L, Stone G, Imelfort M, Hugenholtz P, Tyson GW. Fast, accurate error-correction of amplicon pyrosequences using Acacia. Nat Meth 2012; 9:425-6; PMID:22543370; http://dx.doi.org/10.1038/nmeth.1990
49. Yousfi Monod M, Giudicelli V, Chaume D, Lefranc MP. IMGT/JunctionAnalysis: the first tool for the analysis of the immunoglobulin and T cell receptor complex V-J and V-D-J JUNCTIONs. Bioinformatics 2004; 20 Suppl 1:i379-85; PMID:15262823; http://dx.doi.org/10.1093/bioinformatics/bth945
50. Sanz I. Multiple mechanisms participate in the generation of diversity of human H chain CDR3 regions. J Immunol 1991; 147:1720-9; PMID:1908883
51. Larimore K, McCormick MW, Robins HS, Greenberg PD. Shaping of human germline IgH repertoires revealed by deep sequencing. J Immunol 2012; 189:3221-30; PMID:22865917; http://dx.doi.org/10.4049/jimmunol.1201303
52. Lefranc M-P, Lefranc G. The Immunoglobulin FactsBook. San Diego: Academic Press; 2001.
53. Lefranc MP, Giudicelli V, Ginestoux C, Jabado-Michaloud J, Folch G, Bellahcene F, Wu Y, Gemrot E, Brochet X, Lane J, et al. IMGT, the international ImMunoGeneTics information system. Nucleic Acids Res 2009; 37:D1006-12; PMID:18978023; http://dx.doi.org/10.1093/nar/gkn838
54. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 2010; 25:2460-1; http://dx.doi.org/10.1093/bioinformatics/btq461
55. Miqueu P, Guillet M, Degauque N, Dore JC, Soulillou JP, Brouard S. Statistical analysis of CDR3 length distributions for the assessment of T and B cell repertoire biases. Mol Immunol 2007; 44:1057-64; PMID:16930714; http://dx.doi.org/10.1016/j.molimm.2006.06.026