frontiers
in Immunology

V(D)J recombination enables the sequential rearrangement of variable (V), diversity (D) and joining (J) gene segments in B and T cell antigen receptor (AgR) loci. This mechanism, catalysed by the RAG recombinase complex which recognises the recombination signal sequence (RSS) of each gene segment, is the essential first step in the generation of diverse AgR repertoires, transforming a couple of hundred genes into millions of different antigen specificities (1). In B cells, the immunoglobulin heavy chain (IgH) locus recombines first, with D to J_H recombination on both alleles preceding V_H to DJ_H recombination on one allele in pro-B cells (2). The joining of these genes is imprecise, due to exonuclease activity and the addition of non-templated nucleotides, partly mediated by terminal deoxynucleotidyl transferase (TdT), thereby enhancing diversity (3). Functional IgH chains are expressed on the cell surface with the surrogate light chain as the pre-B cell receptor. This promotes proliferation, differentiation to the pre-B celhn, diffed.ion to ndings concerning the potential mechanisms that underpin recombination of the Ig [illegible] in Read 2, which includes the V–J junctions; and the start position of the V read alignment. Any paired reads identical for all criteria were considered to be PCR duplicates, and only one bypasses the binary mode of chromHMM and allows the user to proceed with continuous values, preventing loss of information and overfitting, which is particularly useful for analysis of a single large locus.

For this, we divided the locus into 200 bp non-overlapping bins and calculated the enrichment of each feature over all bins using bedtools (60) multibam coverage function. As an input for EpiCSeg, we constructed a raw read-counts matrix X in which x_{ij} corresponds to the enrichment of feature i in bin number j. We ran EpiCSeg with varying numbers of states ranging from 3 to 15.

The A or E state were assigned to the promoter or RSS, if they overlapped a window extending from the centre of the gene to 500 bp up- or downstream, respectively, with the exception that if the A/E state segment did not overlap with the gene start, and its centre was downstream of the gene centre, it would be assigned to the RSS but not the promoter, or vice versa.

For the unsupervised step, we first trained a RF-C to predict whether a V gene is "active" or not. We then constructed a RF-R model to predict the recombination level of an active gene. We chose RF since it is generally accepted to be superior in tackling high dimensionality (relatively high number of features with low number of samples for the training) and co-linearity between the features (61).

Read counts for DHS-seq, ChIP-seq, and RNA-seq data were generated using Seqmonk, within four distinct windows for each V gene: "promoter," extending from 500 bp upstream of the start of the gene to its centre; "RSS," extending from the centre of the gene to 500 bp downstream of its 3 end; and "upstream" and "downstream" windows, extending from 500 bp to 3 kb up- or downstream of the gene start or end, respectively. In addition to these four windows for each of the genome-wide datasets, giving a total of 76 chromatin features, we also included three genetic features. These were: the RSS RIC score, the orientation (or strand) of the V gene, and the distance from the V gene to J 1. All of these features, except for the gene orientation, were projected between 0 and 1. These 79 features were considered as the explanatory variable for both the RF-C and RF-R.

The response variable for RF-C was the binary recombination classes (active and inactive), which were defined as described above. For RF-R, the log2-transformed median of the normalised recombination frequencies of active genes was used as the response variable.

Both the RF-C and RF-R approaches were performed with 10-fold cross-validation: 10% of genes were assigned to the test set each time, with every gene included in a test set exactly once. The number of trees generated for each fold was 1,000. For the initial RF-C including all features, the number of features tried at each step was set to 20; for all other models default parameters were used. The average importance of each feature, and SE across the 10-folds, was recorded. For the classification model, the performance was assessed by calculating the percentage of correct predictions (accuracy) across all ten test sets: this was calculated overall, as well as for the active V genes (giving a measure of the sensitivity with which we could identify an active gene) and inactive genes (which gives a measure of specificity). We also calculated the F1 score as a combined measure of sensitivity and specificity. To assess the performance of the regression model, we used the root mean squared error (RMSE) for the predicted recombination frequencies compared to the observed values across all ten test sets. The RMSE gives a measure of the SD in errors, thus 68% of our predictions are expected to have an error within this range. Since our recombination frequencies are log2-transformed, an RMSE of x corresponds to a 2^x-fold difference between the predicted recombination frequency and the observed recombination frequency.

For model selection, we focussed on the 16 most important features from the initial classification or regression model. We trained RF classification or regression models, with 10-fold cross-validation, for all possible combinations of the respective 16 features. These models were then compared using the performance metrics described above. Our analysis was performed using the R package randomForest (62).

Publicly available genome-wide datasets analysed during this study are available in the GEO repository; details including accession numbers are listed in Table 1. All were downloaded from GEO as raw short-read files (SRA) and realigned to NCBIM37/mm9 using Bowtie (49) or Bowtie 2 (63). The VJ-seq datasets generated in this study are available in the GEO repository with accession number GSE101686.[5] Some of the quantitated data from this study is also provided in Table S1 in Supplementary Material.

| Publicly available next generation sequencing datasets utilised in our study.

| Recombination frequencies for all J genes and replicates. Heatmap showing log2-transformed recombination frequencies for each V -J combination across all replicates. Read counts for each replicate were first normalised to the replicate with the median number of reads aligning to V genes. Each row represents a V gene and each column represents an individual replicate for a J gene. V genes and J gene replicates d r6/indo4TJ /1(eplio)Tjlumn rfi1 Tfsimilarity to 1 Tir [(r)1018(ese

| V recombination to J 1 varies widely across the locus. Reads associated with J 1 for each V gene were counted for all replicates, and normalised to the replicate with the median total number of J 1-associated V reads. Bars represent the median normalised read count, while each replicate is displayed as a circle. Genes are arranged geographically, from the 5 end of the locus (left) to the 3 end (J-proximal end; right). Below, names and localisation on chromosome 6 of all actively recombining V genes ( ) are shown, with genes on the forward strand (that recombine by deletion) displayed above the scale bar, and genes on the reverse strand (that recombine by inversion) displayed below. V gene families are represented by colour, with pseudogenes (pg) displayed in grey. Normalised median recombination frequencies to J 1 and recombination signal sequence (RSS) Information Content (RIC) scores are shown for all genes in each V family. The 64 pseudogenes (PG) originate from 13 out of the 20 V gene families. Normalised median recombination frequencies and RSS RIC scores for genes on the forward and reverse strands. Only genes with a RIC score > 38.81, which are considered theoretically capable of recombination, are included; -values from two-sided -tests are shown.

numerous V genes that are more highly represented at the DNA level, including 14 pseudogenes that were not detected in the RNA repertoire. Conversely, all V genes present in the expressed repertoire were detected in our assay, albeit in some cases with very low read counts. This highlights the significant contribution of transcription and posttranscriptional processes to the expressed repertoire, which would confound the aim of this study to interrogate the pre-recombination chromatin state.

To facilitate further investigation of the V-J1 repertoire, we performed a binomial test to distinguish V genes that are significantly recombining ($p_{adj} < 0.01$, Figure 3A; Table S1 in Supplementary Material). Out of 162 genes, 105 (64.8%) passed the binomial test and were labelled "active" to denote "actively recombining"; these genes were detected with a minimum of 59 reads, and included 15 pseudogenes, and are hereafter referred to as active genes. The remaining 57 genes had insufficient evidence of activity, with the median read count for each below 59, and were labelled "inactive." These inactive genes included eight V genes that are considered to be functional, suggesting that they contribute little to the primary repertoire. The usage of active genes was weakly correlated with the RIC score ($R^2 = 0.42$; Figure 3B); however, genes with a similar RIC score could

| Recombination signal sequence (RSS) Information Content (RIC) score can only partially explain variation in V gene activity. Distribution of V-J1 recombination frequencies for all 162 V genes. A one-sided binomial test was used to gauge the significance of their recombination frequency, allowing each gene to be labelled as active (fdr adjusted *p*-value <0.01) or inactive. Dependence of active genes' recombination frequency on RSS RIC score. Linear regression model (dashed line) reveals that only 17.7% of the variation in recombination frequency of active genes can be explained by the RIC score. RSS RIC scores of active and inactive V genes; *p*-value from a two-sided Wilcoxon rank sum test.

recombine at markedly different frequencies. Moreover, some inactive genes have RIC scores that are comparable to those of active genes (Figure 3C). Importantly, a linear regression model revealed that only 17.7% of the variability in V gene usage could be explained by RIC score alone (Figure 3B), highlighting the need to explore whether other mechanisms, such as chromatin features, contribute to shaping the repertoire.

Igκ

Colocalisation of Chromatin Features with Vκ Genes

In order to assess the contribution of chromatin features to V gene recombination, we used published genome-wide datasets from mouse pro-B cell models that are developmentally stalled prior to recombination of the Igh locus (48,73). There are numerous pro-B cell datasets available, and the regulatory state of the Ig locus has already begun to be established by this stage (39,40,74). Our analysis aims to determine the importance of these early regulatory events in priming the locus for recombination, thus shaping the primary repertoire. Moreover, Ig locus gene-specific studies (75,76), as well as the small number of available pre-B cell datasets (37), revealed similar enrichment of CTCF, YY1, and histone H3 acetylation in pro-B and pre-B cells. The chromatin features we chose to assess included DHS, germline transcription, and ChIP for several histone modifications and TFs (Table 1).

We first measured the distance from the centre of each V gene to the summit of the closest peak for each DHS- and ChIP-seq dataset that had at least 35 peaks over the locus, both upstream (towards the promoter) and downstream (towards the RSS)c 0.Si6)]TJ 0 00i4 Tc 5 T* [(Su-10s41.)215 Tc[(g)8(en)4(.es [(0.S)()

| Chromatin features associate with both the promoters and recombination signal sequences (RSSs) of active V genes. Scatter plots and density plots showing the log10-transformed distances of the closest ChIP-seq peaks both up- (+) and downstream ( ) from the centre of active (orange) and inactive (dark grey) V genes. Yellow (promoter) and blue (RSS) shading indicates the range of distances within which 80% of the start and end sites of V genes, respectively, are located. Grey shading indicates a distance of >1 kb from the V gene, based on the median V gene length (525 bp). Average enrichment, relative to background, of chromatin features across all active (orange) and inactive (dark grey) V genes. Genes have been scaled such that 0 and 100 represent the start and end of the gene, respectively. Yellow and blue shading indicates the location of the promoters and RSSs, respectively. Median recombination frequencies and RSS Information Content (RIC) scores of active genes, excluding 10 genes that had <70% mappability. Genes are categorised based on the number of ChIP-seq peaks within 1 kb of the gene , or the localisation of those peaks upstream (promoter) or downstream (RSS) of the gene centre . -values for V -J 1 recombination frequency are fdr adjusted, based on a two-sided Wilcoxon rank sum test. values indicate the number of genes in each category.

| V gene chromatin state is associated with recombination frequency. Number of active and inactive genes in each state. -value based on a Fisher's exact test. Violin plots with boxplot superimposed (black) showing median recombination frequencies of active genes in the Bg state compared to active genes associated with the A and/or E state. A and/or E state genes are considered altogether , or categorised based on the localisation of the state to their promoter or recombination signal sequence (RSS) . 10 genes that had <70% mappability were excluded. Fdr-adjusted -values based on two-sided WilcokasedRSS)

| Relationship between ChIP enrichment and recombination frequency for important RSS , promoter , and upstream features in RF models. Enrichment of chromatin features over the locations in which they were found to be important (projected between 0 and 1 for each gene: identical to the input for RF-R models), for active genes with low (= 23; 117–973 reads), medium (= 24; 1,006–1,801 reads) and high (= 24; 1,880–9,137 reads) relative frequency of recombination. Only genes with a high quality RIC score (> 14) were considered. Fdr-adjusted -values driven by two-sided Wilcoxon rank sum test. All data are included for statistical testing, but to better visualise the data, some outliers are not displayed.

Conversely, we noted a slight, negative association between the recombination frequency and the binding of some TFs upstream of the gene, including PAX5, PU.1, and CTCF, with a significantly greater enrichment of PAX5 upstream of genes that recombine at a low level compared to those that recombine at a medium level (Figure 9C).

We have adapted the VDJ-seq assay for the Igk locus to quantitatively profile the V-J repertoire and to enable an in-depth analysis of the local drivers of recombination. Using cutting-edge random forest machine learning approaches to integrate genetic and chromatin features, we have distinguished genes that are actively recombining from those that are not, and have predicted the relative usage of active V genes in primary recombination. We have found that local chromatin features, including PU.1 and IKAROS binding, and H3K4 methylation, explain much of the variation in recombination among V genes.

The accuracy with which we can predict both V gene activity and frequency of usage, even when the influence of the RIC score is excluded, is striking. Since we used pro-B cell genome-wide datasets, focussing on early events that prime the Igk locus for recombination, the regulatory status of the Igk locus may not fully reflect its state in pre-B cells immediately prior to V-J 1 recombination. This suggests that early priming events are crucially important and that to a large extent, the recombination potential of each V gene has been established by the pro-B cell stage. Nevertheless, we cannot exclude the possibility that features ranked unimportant here may become enriched at the Igk locus later in development, or that additional pre-B cell specific features including IRF8, AIOLOS, and BRWD1 (27, 78, 79) may play a local regulatory role in recombination. Profiling of the locus in a Rag/ model with a rearranged V_HDJ_H transgene wou 331.0G13(e)-6(a Tc 0 T0565 314.1919 Ta9(t)6(N)24.9(e)-7.9(v)8.588 -)-15.9(y could be a concern, although the use of 10-fold cross-validation mitigates this possibility. Nevertheless, the clear relationship that we observed between enrichment and recombination for several features, including IKAROS and IRF4, highlights the value of this approach in providing a shortlist of chromatin features that are potential drivers of recombination. The contribution of other features that did not display such a clear relationship with recombination, such as H3K4me3 at the RSS and MED1 bind-

B4(ce, D r)1e 9 1. Fugmann SD, Lee AI, Shockett PE, Villey IJ, Schatz DG. RAG proteins CA. [(5 412(f 3.108 -2(h)3.9(e m)4(o)11.(r)13(en(p)-5(263 a)3(s)60(t c)6(h)3(a)90 Tw

8. and V(D)J recombination: complexes, ends, and transposition. Annu Rev Immunol (2000) 18:495–527. doi:10.1146/annurev.immunol.18.1.495

2. Jung D, Giallourakis C, Mostoslavsky R, Alt FW. Mechanism and control of V(D)J recombination at the immunoglobulin heavy chain locus. Annu Rev Immunol (2006) 24:541–70. doi:10.1146/annurev.immunol.23.021704.115830

3. Benedict CL, Gilfillan S, Thai TH, Kearney JF. Terminal deoxynucleotidyl transferase and repertoire development. Immunol Rev (2000) 175:150–7. doi:10.1111/j.1600-065X.2000.imr017518.x

4. Hendriks RW, Middendorp S. The pre-BCR checkpoint as a cell-autonomous proliferation switch. Trends Immunol (2004) 25(5):249–56. doi:10.1016/j.it.2004.02.011

5. Herzog S, Reth M, Jumaa H. Regulation of B-cell proliferation and differentiation by pre-B-cell receptor signalling. Nat Rev Immunol (2009) 9(3):195–205. doi:10.1038/nri2491

6. Brekke KM, Garrard WT. Assembly and analysis of the mouse immunoglobulin kappa gene sequence. Immunogenetics (2004) 56(7):490–505. doi:10.1007/s00251-004-0659-0

7. Li YS, Hayakawa K, Hardy RR. The regulated expression of B lineage associated genes during B cell differentiation in bone marrow and fetal liver. J Exp Med (1993) 178(3):951–60. doi:10.1084/jem.178.3.951

8. Victor KD, Vu K, Feeney AJ. Limited junctional diversity in kappa light chains.

8. Junctional sequences from CD43+B220+ early B cell progenitors resemble those from peripheral B cells. J

[illegible] Mansson R, Heinz S, Miyazaki K, Miyazaki M, et al. Global changes in the nuclear positioning of genes and intra- and interdomain genomic interactions that orchestrate B cell fate. Nat Immunol (2012) 13(12):1196–204. doi:10.1038/ni.2432

40. Stadhouders R, de Bruijn MJ, Rother MB, Yuvaraj S, Ribeiro de Almeida C, Kolovos P, et al. Pre-B cell receptor signaling induces immunoglobulin kappa locus accessibility by functional redistribution of enhancer-mediated chromatin interactions. PLoS Biol (2014) 12(2):e1001791. doi:10.1371/journal.pbio.1001791

41. Pan X, Papasani M, Hao Y, Calamito M, Wei F, Quinn WJ III, et al. YY1 controls Igkappa repertoire and B-cell development, and localizes with condensin on the Igkappa locus. EMBO J (2013) 32(8):1168–82. doi:10.1038/emboj.2013.66

42. Ji Y, Resch W, Corbett E, Yamane A, Casellas R, Schatz DG. e in vivo pai14ln, R9(n)814ln, Rn, Rt, aat(l r)13(e)-9(cep)11(t)6(o)12(a lo)-8.ci(e)6(.)]TJ /T1_3 1 Tf -0.6 0 Tc 01106 Tw , Cl 013:41968Ÿ91. doi:1001171/118..B-cele20.03 01066

80. G