RAG Recombinase as a Selective Pressure for Genome Evolution

D. Passagem-Santos†, M. Bonnet†, D. Sobral, I. Trancoso, J.G. Silva, V.M. Barreto, A. Athanasiadis, J. Demengeot*, and J.B. Pereira-Leal*

Instituto Gulbenkian de Ciência, Oeiras, Portugal

†These authors contributed equally to this work.

*Corresponding authors: E-mails: jleal@igc.gulbenkian.pt; jocelyne@igc.gulbenkian.pt.

Accepted: October 25, 2016

Abstract

The RAG recombinase is a domesticated transposable element co-opted in jawed vertebrates to drive the process of the so-called V(D)J recombination, which is the hallmark of the adaptive immune system to produce antigen receptors. RAG targets, namely, the Recombination Signal Sequences (RSS), are rather long and degenerated sequences, which highlights the ability of the recombinase to interact with a wide range of target sequences, including outside of antigen receptor loci. The recognition of such cryptic targets by the recombinase threatens genome integrity by promoting aberrant DNA recombination, as observed in lymphoid malignancies. Genomes evolution resulting from RAG acquisition is an ongoing discussion, in particular regarding the counter-selection of sequences resembling the RSS and the modifications of epigenetic regulation at these potential cryptic sites. Here, we describe a new bioinformatics tool to map potential RAG targets in all jawed vertebrates. We show that our REcombination Classifier (REC) outperforms the currently available tool and is suitable for full genomes scans from species other than human and mouse. Using the REC, we document a reduction in density of potential RAG targets at the transcription start sites of genes co-expressed with the *rag* genes and marked with high levels of the trimethylation of the lysine 4 of the histone 3 (H3K4me3), which correlates with the retention of functional RAG activity after the horizontal transfer.

Key words: Bioinformatic RSS classifier, Cryptic RSS, motif evolution, Recombination Classifier.

Background

DNA-mediated transposons, also known as class 2 transposable elements (TE), are one of the most common genomic elements in the eukaryotes genomes (Bulyk et al. 2009), shaping the genomic landscape of their host by alteration of gene function, induction of chromosomal rearrangements, and generation of new coding and noncoding genetic material (Feschotte and Pritham 2007). One outstanding case of domesticated TE (Alzohairy et al. 2013; Huang et al. 2016) is the Recombination Activating Genes (RAG), present in all jawed vertebrates. The *rag1* and *rag2* genes are just a few kilobases apart in the genome, which is interpreted as a signature of their entry in the genome by horizontal transfer in the form of a transposon, and their encoded proteins associate to form the RAG1/RAG2 recombinase, which has been shown to have transposase activity (Clatworthy et al. 2003). Over 400 million years ago, this recombinase was co-opted in jawed vertebrates to drive the process of the so-called V(D)J recombination, which is the hallmark of the adaptive immune system.

V(D)J recombination gives rise to the large collections of distinct Ig and TCR antigen receptors in B and T lymphocytes, respectively. Each antigen receptor locus is composed of multiple gene segments that are flanked by Recombination Signal Sequences (RSS). These are specifically recognized by RAG1/2 recombinase, which during a typical reaction generates DNA double strand breaks (DSB) at each of a pair of RSS (Akira et al. 1987). The joining of the DSB is then accomplished by the nonhomologous end-joining machinery and results in one of many possible genes encoding a unique functional chain of an antigen receptor. V(D)J recombination is a tightly controlled reaction, both at the level of the transcriptional and post-translational regulation of recombinase, which is only active in the G1 phase of developing lymphocytes, and the accessibility of the RSS as a substrate for recombination, which is

determined by epigenetic marks. Notably, the plant homeodomain of RAG2 has the ability to bind H3K4me3 (Shimazaki et al. 2009), a mark abundant in transcribed genes and enriched in the sequences surrounding transcriptional start sites (Lauberth et al. 2013). Moreover, the permissive H3K4me3 has been shown to enhances V(D)J recombination through its interaction with the recombinase (Matthews et al. 2007; Gopalakrishnan et al. 2013).

RSS display a consensus structure encompassing a heptamer and a nonamer, separated by either a 12- or a 23-nucleotide spacer sequence exhibiting few consensus nucleotide positions (Lewis 1994). These sequences are thus considerably degenerated, which highlights the ability of the recombinase to interact with a wide range of target sequences. One outcome of the low complexity of the RSS is the presence of millions of DNA sequences with similar features—called cryptic RSS (cRSS) (Lewis and Wu 1997; Dik et al. 2007)—strewn all over the genomes of jawed vertebrates. cRSS may be recognized by the recombination machinery (Lewis et al. 1997; Cowell et al. 2002), although the actual frequency of RAG-mediated mistargeted recombination events remain speculative and is only recently being appreciated as potentially very high. For long, only few cRSS have been formally described outside V(D)J loci both in healthy individuals and in tumorigenic contexts (Lieber et al. 2006; Schlissel et al. 2006; Onozawa and Aplan 2012). More recently, high throughput techniques coupled with DNA break bait strategies identified hundreds of RAG "off-target" breaks all over the genome(Hu et al. 2015), and targeted approaches in tumors identified frequent illegitimate recombination events in specific loci (Papaemmanuil et al. 2014; Mijuskovic et al. 2015).

The sequence of the RAG proteins, the modular structure of antigen receptor genes and their bona fide RSS are conserved among all jawed vertebrates (Sadofsky 2001), indicating the molecular intricacy of the presumably advantageous adaptive immune system. When the RAG transposon was horizontally transferred to animal lineages and how it subsequently implied genomes evolution is an ongoing discussion fuelled by the genome sequencing projects (Huang et al. 2014). In this work, we aimed at understanding the selective pressures acting on the genomes of jawed vertebrates regarding the off-targets of RAG, given that developing lymphocytes expose the genome to the RAG1/2 recombinase activity. One hypothesis is that cRSS are epigenetically silenced, however, the sets of genes necessary for B and T cell development that are co-expressed with RAG obviously cannot be epigenetically silenced in developing lymphocytes. Another scenario is that the vertebrate genomes evolved specific DNA sequence changes as a defense mechanism against RAG presence. This latter hypothesis received experimental support in a recent study reporting a genome wide analysis of mouse and human DNA sequences bound by RAG1 protein and revealing that such sequences are enriched in transcription start site and biased to low cRSS content (Teng et al. 2015).

Here, we hypothesize that natural selection may have removed cRSS after the domestication of RAG in those sequences more likely to be accessible to the RAGs, namely, gene co-expressed and/or H3K4me3 marked at the developmental time of RAG expression, as these cannot be protected by epigenetic mechanisms.

To test this hypothesis, we developed and validated a new bioinformatics tool to map RSS in all jawed vertebrates, more sensitive able to detect cRSSs. This tool confirmed a decreased RSS density at the transcription start sites (TSS) of genes bound by Rag1 in the mouse (Teng et al. 2015), and revealed the same trend in genes co-expressed with the *rag* and genes marked with high levels of the H3K4me3, with the latter providing the highest correlation. By analyzing vertebrates and invertebrates genomes, we show that depletion of RSS is a reproducible trait specific of V(D)J positive species. We conclude that the entry of the *rag* genes as well as their retention as an active recombinase had an impact in the structure of genomes outside of antigen receptor loci.

Results and Discussion

Recalibration of the RIC for Use in Multiple Species

In 2002, Kelsoe and collaborators developed a statistical model that weighs the nucleotide composition at each RSS position as well as the correlation between positions, using a reference set of mouse RSS associated to functional V, D, or J gene segments. The model computes a RSS Information Content (RIC) score, highest for the most common V(D)J-associated RSS, and is to date the only available tool to characterize the nucleotide sequence of RSS, specifically in the mouse and human genomes(Cowell et al. 2002; Merelli et al. 2010).

Our ultimate goal being to detect RSS and cRSS in multiple organisms, we first tested the RIC's power to do so, based on a multi-species datasets that we compiled (see Methods). As shown in table 1, the RIC has high sensitivity (percentage of true RSS classified as RSS) for mouse and human RSS, but is less sensitive for other species' RSS, failing to identify over 10% of the authentic RSS (also see supplementary file S1, Supplementary Material online for origin and number of sequences). This lower sensitivity is even more pronounced for cRSS, for which the majority of sequences fails the RIC thresholds. These results show that in its current formulation, the RIC is not ideally suited for multi-species studies, consistent with the fact that the RIC algorithm was originally trained using mouse authentic RSS. Retraining the RIC with human sequences resulted in significantly different scores, as was observed by Merelli et al. (2010). We thus retrained the RIC algorithm (hereafter rRIC) with a pooled non-redundant multi-species dataset (see Methods). As the overall pairwise mutual information did not change significantly

Table 1
REC and RIC Sensitivity in Different Datasets

Dataset	12-RSS		23-RSS	
	RIC	REC	RIC	REC
Mouse	0.96	0.99	0.96	1.00
Human	0.92	0.97	0.97	0.99
Other species	0.90	0.98	0.89	0.99
cRSS	0.16	0.39	0.42	0.89

(supplementary file S2, Supplementary Material online), we used the original position–correlation matrix, only retraining the specific nucleotide frequencies. Note that this implies that the initial position correlations identified by the RIC do not change for larger and multispecies datasets, illustrating the high conservation of the RSS motif. As our final goal was full genome scans for detection of potential cRSS, we adjusted the thresholds of the rRIC for maximum specificity (i.e., not allowing false positives within a negative set of sequences gathered from the literature; see methods for more details). The thresholds for the original RIC were −38.81 and −58.45 for 12- and 23-RSS, respectively, whereas the thresholds defined for rRIC are −33.39 and −51.96 for 12- and 23-RSS, respectively. The rRIC produces less false negatives than the RIC algorithm (4th vs. 5th and 9th vs. 10th columns in supplementary file S3, Supplementary Material online). As expected, this improvement is marked for V(D)J associated RSS in species other than mouse or human (64 false negatives for RIC versus 33 for rRIC out of 529 12-RSS, and 62 for RIC versus only 6 for rRIC out of 581 23-RSS, *P*-values of 2.58×10^{-4} and 7.25×10^{-10} for 12- and 23-RSS, respectively). The rRIC also performed better for cryptic 23-RSS (86 for RIC versus 19 for rRIC out of 126 23-RSS, *P*-values of 2.72×10^{-12} and 2.15×10^{-14} for 12- and 23-RSS, respectively). Of note, the score of the human cryptic 12-RSS LMO2 (Dik et al. 2007), which failed the threshold using the original RIC score whereas exhibiting *in vitro* recombination (Raghavan et al. 2001; Trancoso et al. 2013), was detected as a positive RAG-target using the rRIC model (score of −30,96). Importantly, adding non-mouse RSS to the training set did not have a significant impact on the sensitivity regarding mouse data (supplementary file S3, Supplementary Material online). Finally, a 10-fold cross-validation confirmed that the rRIC outperforms the RIC (supplementary file S4, Supplementary Material online). However, retraining did not increase sufficiently the sensitivity for 12-RSS demanding thus a complementary characterization of RSS sequences.

A New Recombination Classifier, the REC

We hypothesized that beyond nucleotide sequence, additional RSS features like DNA biophysical properties could bring additional predictive power. This reasoning is based on the evidence that purine–pyrimidine nucleotides repeats prone to adopt a non-B-DNA conformation (namely, Z-DNA) are the target of misdirected RAG recombination that may for instance translocate the Bcl2 gene to the *Igh* locus (Raghavan et al. 2004). Strikingly, the heptamer of the canonical RSS represents close to an ideal motif for the formation of Z-motif. Moreover, the junction between B and Z-DNA is formed by the extrusion of a single base pair (Ha et al. 2005), thus markedly exposed to modification enzymes, presumably including the recombinase. This lead us to use an algorithm developed to model transcription factors binding sites using biophysical properties of the DNA (CRoSSeD; Meysman et al. 2011), together with the biophysical properties pre-computed in the DiProDB database (Friedel et al. 2009), testing their ability to predict cRSS. This approach had a predictive power comparable to the rRIC, but was able to identify RSS that failed the rRIC thresholds (supplementary file S3, Supplementary Material online). These results suggested that combination of the rRIC and the CRoSSeD predictors would improve RSS detection sensitivity.

We thus derived a new ensemble classifier, the REcombination Classifier (REC), that combines the information from rRIC and CRoSSeD. Each tested sequence is classified by two scores (one for each classifier) and is labeled as positive if both scores pass the corresponding threshold (fig. 1*A*). We defined the combination of thresholds such as to maximize the sensitivity whereas ensuring no false positive in the training set (see methods; vertical and horizontal lines on fig. 1*B* and *C*). With this strategy, the obtained thresholds are, for the rRIC component, −35.32 for 12-RSS and −54.19 for 23-RSS, and for the CRoSSeD component 0.999681 and 0.99974 for 12- and 23-RSS, respectively. The REC shows improved predictive power for cRSS and RSS from multiple species and to a smaller extent for authentic human and mouse RSS (see table 1 and supplementary file S3, Supplementary Material online for raw counts). The thresholds for rRIC and REC were adjusted for maximum specificity, ensuring that none of the negative sequences were classified as RSS; thus the results shown in table 1 and in supplementary file S3, Supplementary Material online represent the lower boundary for REC performance regarding sensitivity. For 12-cRSS, the REC represents a small improvement in sensitivity compared with the RIC (table 1; 0.28 for RIC vs. 0.39 for REC, *P*-value 4.14×10^{-04}), which is more impressive for 23-cRSS, as only 13 sequences remain as false negatives (out of 126 total 23-cRSS, *P*-value 5.80×10^{-30}; table 1). We also assessed the REC's performance by 10-fold cross validation methods, revealing that it outperforms both the RIC and the rRIC models independently of the species (supplementary file S4, Supplementary Material online) with significantly higher predictive statistics (*P*-values of 2.2×10^{-152} for12-RSS and 6.71×10^{-158} for 23-RSS for REC vs. RIC and of 1.37×10^{-73} for 12-RSS and 1.24×10^{-158} for 23-RSS for REC *versus* rRIC). Given these results, we will use the REC as our standard estimator for genome-wide cRSS prediction.

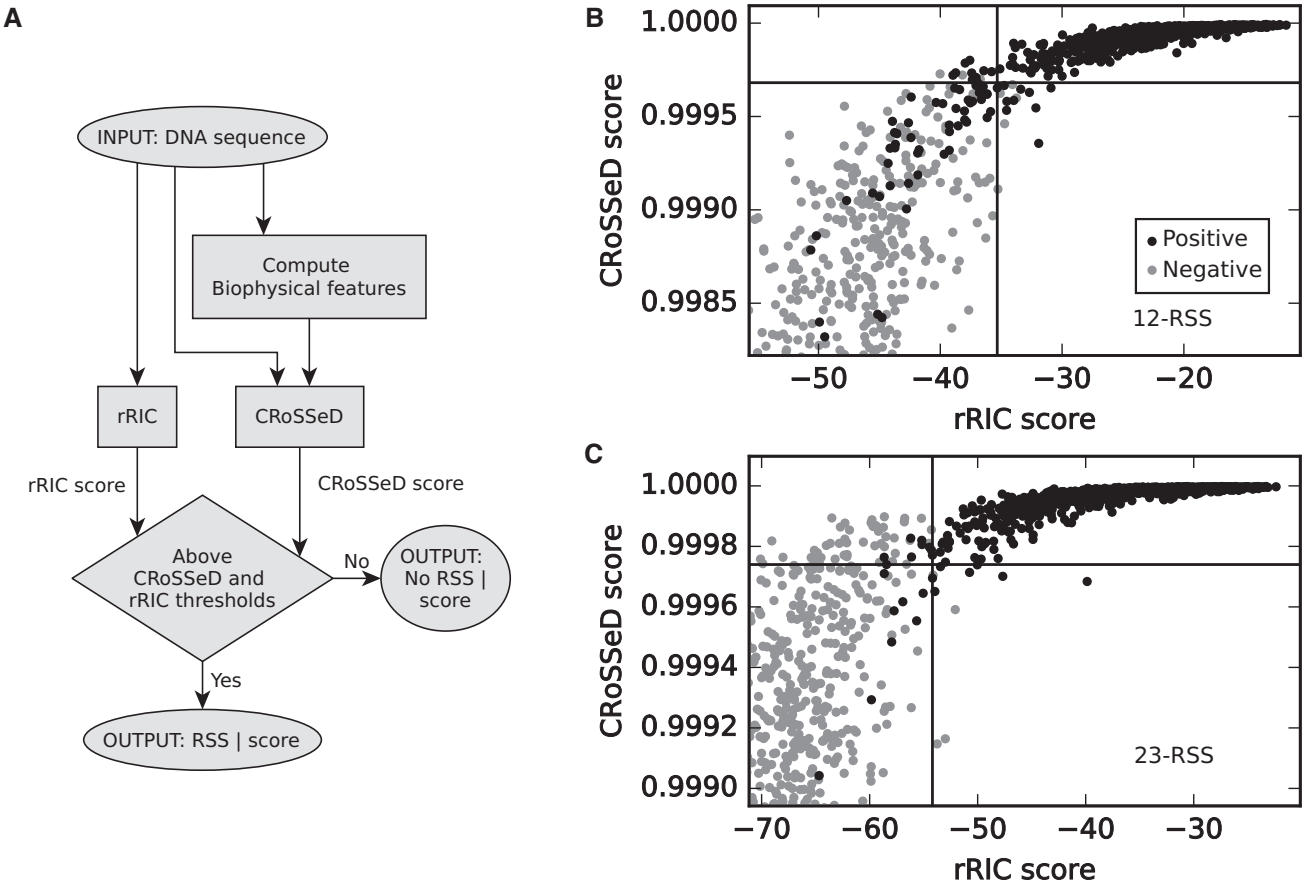

Fig. 1.—The REC scoring procedure and thresholds. (*A*) REC's flowchart (panel A). The input for REC calculations is the DNA sequence to be tested. The rRIC score is obtained using the new positive training set. A set of biophysical features is calculated from the DNA sequences and both the features and the DNA sequence serve as input for CRoSSeD score calculation. When both scores is below the predefined thresholds the sequence is classified as a predicted cryptic RSS (RSS). If any of the two scores is below the threshold, the sequence is classified as a non-RSS. Panels B and C represent the CRoSSeD and rRIC scores of our training set for 12-RSS and 23-RSS, respectively. The horizontal lines represent CRoSSeD thresholds and the vertical lines rRIC thresholds. The grey dots represent sequences from our negative training set and the positive sequences are represented by the black dots. A sequence is classified as a RSS by REC if it lays in the upper right quadrant of the plot.

Negative Bias of Predicted cRSS Density at Transcription Start Sites of RAG Accessible Genes in Mice

We used the REC to perform a genome-wide census of predicted RSS in the mouse genome, for which there are abundant functional genomic data. The potential of a DNA sequence to be targeted by the RAG complex for V(D)J recombination is not only defined by the presence of a RSS but also by the accessibility of the region to the recombinase (Yancopoulos and Alt 1986; Schatz and Ji 2011). Thus, we hypothesized that accessible genes are those under a stronger selective pressure to reduce the number of RSS. Accessible genes can be defined as those co-expressed with both RAG proteins (hereafter coRAG+ genes) or associated with the permissive H3K4me3 epigenetic mark at the differentiation stages where RAG are expressed (hereafter markRAG+ genes). Experimentally, genes effectively bound by the RAG can be identified by Chip seq (hereafter boundRAG1+ genes), as done previously (Teng et al. 2015).

The nucleotide compositions of the different regions of a gene are intrinsically distinctive. In particular, the structures of exons, introns and sequences neighboring transcription start sites (TSS) present a different CG content (Koudritsky and Domany 2008; Amit et al. 2012) and are known to be subjected to distinct evolutionary pressures. We compared the RSS density at each of the different gene bodies (TSS, exons and introns) in the RAG+ gene sets with the corresponding gene bodies in the RAG− gene set, by calculating the ratio between RAG+ RSS density and RAG− RSS density. We first tested bulk coRAG+ and − gene sets selected as described in the methods. As shown in figure 2, the TSS regions display a low ratio, with the RAG+ TSS having a significant reduction of RSS compared with RAG− TSS (29% less for 12- and 23%

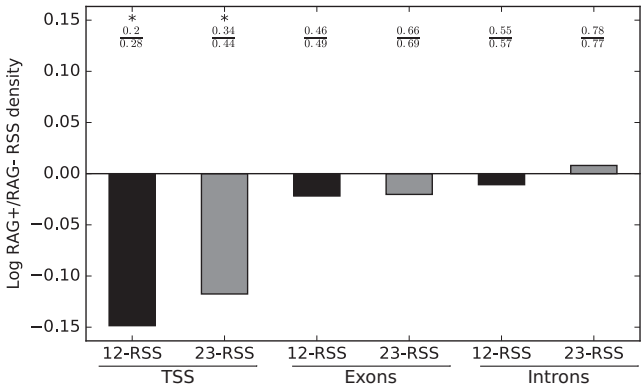

Fig. 2.—RSS densities in specific gene regions in mouse. TSS, exons and introns exhibit different patterns of RSS density in genes co-expressed with the *rag* genes. We computed the ratio of RSS density for each region between RAG+ genes and RAG− genes (plotted as log(ratio) for symmetry). Black and grey bars represent the log ratios for 12-RSS and 23-RSS densities, respectively. All RSS densities are assessed using the REC. The fraction above each bar is the actual fraction of RSS density between RAG+ and RAG− genes. * stand for *P*-values below 10^{-2}, respectively, for the Mann–Whitney U test.

less for 23-RSS; *P*-values of 3.71×10^{-08} and 6.49×10^{-10}, respectively). For example, gene Rpf1 (Ensemble ID ENSMUSG00000028187), that is active in the hematopoietic lineage had no predicted RSS both for 12- and 23-RSS. For comparison, the gene Paip1 (Ensemble ID ENSMUSG00000025451), with the highest density of RSS in the TSS area for genes co-expressed with RAG, has 20 12-RSS and 7 23-RSS, whereas the gene RIKEN (Ensemble ID ENSMUSG00000027446), that is not co-expressed with RAG, has 33 12-RSS and 13 23-RSS predicted.

As promoters of housekeeping genes (hereafter HK) present different features than promoters of tissue specific genes (Zhu et al. 2008), we further decomposed the gene set in HK genes, representing 26% of the coRAG+ gene set, and in lymphoid-specific genes (RS+, 44% of coRAG+ genes). As control, we analyzed a new set of genes specific from non-lymphoid tissues, expressed either in skeletal muscle, liver, lung or kidney (TS+) vs. not expressed in any of these tissues (TS-, see table 2). This partition revealed that both genes specific of RAG expressing tissues (RS+) and HK genes (HK) display lower RSS density around TSS than RAG− genes, with a better signal for the latters (fig. 3, 25% and 16% reduction for 12 and 23 RSS, respectively, in the RS+ set, compared with 32% and 31% reduction for the HK set, *P*-value 8.71×10^{-4} and 2.63×10^{-3} for 12- and 23-RSS, respectively, for the RS+ set and 7.02×10^{-3} and 8.84×10^{-6} for 12- and 23-RSS for the HK+ set). In contrast, RSS densities around the TSS were similar when comparing genes specific for nonlymphoid tissues (TS+ vs. TS− subsets) (fig. 3).

We next set out to test whether H3K4me3 loads around the TSS correlates with lower RSS density. We used H3K4me3 data from double negative thymocytes (see Methods) to partition all mouse genes into three categories according to their H3K4me3 load at TSS, and retained the upper third and lower third sets of genes ($H3K4me3^{high}$ and $H3K4me3^{low}$, respectively, see supplementary file S6, Supplementary Material online) for RSS density analysis (fig. 4). Genes with high loads of H3K4me3 display a strong decrease in RSS density at TSS compared with genes with low H3K4me3 loads (*P*-values $< 10^{-34}$), with a reduction of 51% for 12-RSS density and 39% for 23-RSS. This effect is stronger than the one observed when comparing coRAG+ and − genes (fig. 2 and dashed lines on fig. 4) and is likely provided by the inclusion of poised genes in the $H3K4me3^{high}$ but not in the coRAG+ set, which bear high levels of H3K4me3 at their TSS but are not (yet) expressed. Confirming that H3K4me3 best explains the specific reduction in RSS numbers at TSS, comparison of the RSS densities in coRAG+ and − genes within the $H3K4me3^{high}$ set revealed no significant differences (*P*-values > 0.05, fig. 4*A* and *B*, right bar).

A recent analysis of RAG1 bound sequences showed enrichment in TSS bearing H3K4me3 mark, as well as decreased density of RIC predicted RSS (Teng et al. 2015). Re-analysis of the same bound RAG1+ and − gene sets, now using the REC and covering TSS but also introns, exons, confirmed that the

Table 2
Definition of Analyzed Populations of Genes

Group		RAG	Liver	Kidney	Lung	Skeletal muscle
RAG+		+				
RAG−		−				
RS+		+	−	−	−	−
HK		+	+	+	+	+
TS+	Liver	−	+	−	−	−
	Kidney	−	−	+	−	−
	Lung	−	−	−	+	−
	Skeletal muscle	−	−	−	−	+
TS−			−	−	−	−

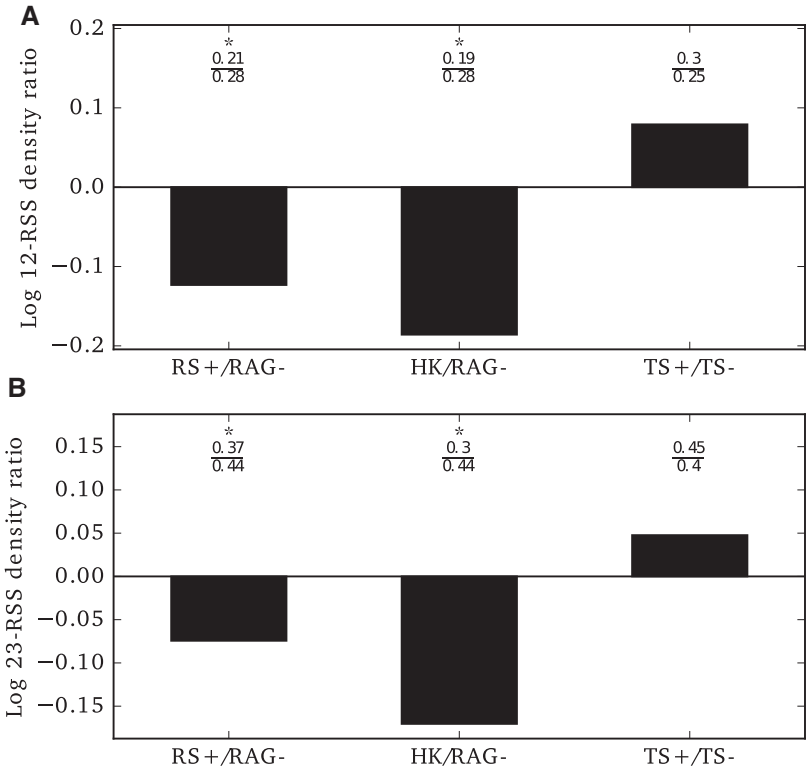

Fig. 3.—Reduction of RSS densities at TSS of housekeeping and RAG-specific genes in mouse. The log ratio between the RSS densities at TSS of housekeeping genes (HK+/RAG−; panel A for 12-RSS and B for 23-RSS) or genes specific of RAG expressing tissues (RS+/RAG−; panel A for 12-RSS and B for 23-RSS) over genes never co-expressed with the RAG complex (RAG−) were computed for 1 kb gates centered on TSS. Similarly, ratios of RSS densities at TSS of liver, lung, kidney or skeletal muscle tissue-specific genes over non tissue-specific genes (TS+/TS−; panel A for 12-RSS and panel B for 23-RSS) were calculated around TSS. The fraction above each bar is the actual fraction of RSS densities. Mann–Whitney U test were performed. * stand for *P*-values below 10^{-2}.

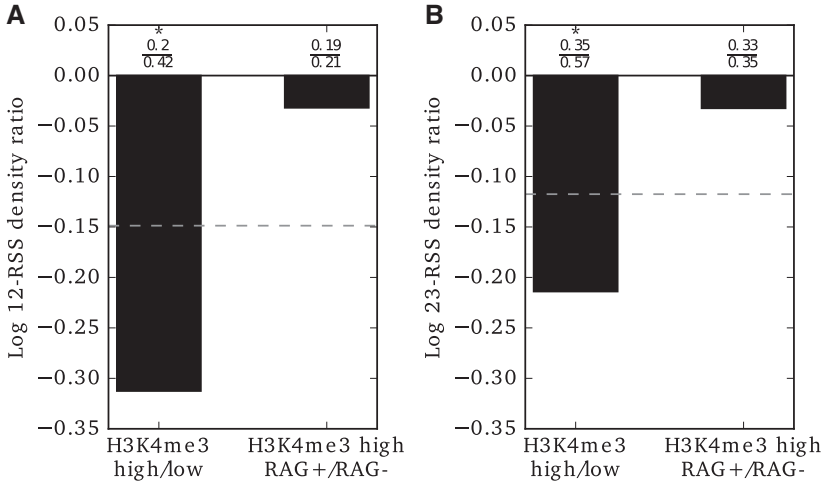

Fig. 4.—H3K4me3 loads explain RSS depletion at TSS in mouse. The log ration between RSS density at TSS of H3K4me3^{high} genes over H3K4me3^{low} genes (left bars, 12-RSS on panel A and 23-RSS on panel B) was computed. The same ratio previously calculated for RAG+ over RAG− genes is also represented (dashed lines). Within H3K4me3^{high} genes, we compute the log ration of RSS density at TSS between RAG+ over RAG- genes (right bars, 12-RSS on panel A and 23-RSS on panel B). The fraction above each bar is the actual fraction of RSS densities. Mann–Whitney U test was applied. *stands for *P*-values $< 10^{-2}$.

reported depletion of RSS is specific to the TSS region. Our re-analysis shows a reduction of 36% for 12-RSS and 31% for 23-RSS density (supplementary file S5 and fig. S3, Supplementary Material online). These scores are consistent with, but lower than, those observed when comparing gene sets partitioned for their level of H3K4me3 load (compare fig. 4 and supplementary fig. S3, Supplementary Material online). Our H3K4me3^{high} gene set is likely to be enriched in sequences intensively bound by Rag2, and conversely, our H3K4me3^{low} gene set may serve as a more stringent negative reference.

Taken together, these analyses serve to validate the REC and provide additional evidence that H3K4me3 is a major factor correlating with the reduction of RSSs at TSSs. In turn, these results support the notion that the mouse genome was under evolutionary pressure to reduce cRSS density specifically at sites that would have high levels of H3K4me3 in the tissues where an active RAG complex is expressed. With this perspective, deregulation of the H3K4me3 levels or changes in the transcriptional program in *rag*-expressing cells could allow the recombinase to bind to regions that have not been under selective pressure and thus have remained enriched in RSS, potentially leading to genomic instability. This proposition is consistent with evidence that H3K4me3 contribute to illegitimate recombination through RAG-mistargeting to cRSS (Shimazaki and Lieber 2014).

Lower RSS Density Correlates with Functional RAG Activity

As evoked above, our data confirms that the mouse genome show signature of negative selection for RSS around TSS, in genes potentially or effectively bound by the recombinases. To investigate whether this feature is specific of the evolutionary history of mouse or is conserved in all organisms with an adaptive immune system based on RAG activity, we extended our analysis to 18 species (see fig. 5 and Methods), among which 9 do and 9 are not V(D)J recombination competent. In the absence of available epigenetic data for all these species, we assumed that the epigenetic regulation of orthologues is similar in different species (Zemach et al. 2010; Long et al. 2013). We first defined a robust set of genes for mouse using the intersection of our H3K4me3^{high} gene set (6638 genes) with the bound RAG1+ gene set defined earlier (Teng et al. 2015) (4018 genes), which resulted in 2275 genes defining this novel RAG+ gene set. Our set of mouse H3K4me3^{low} genes defined the RAG− gene pool. We mapped the

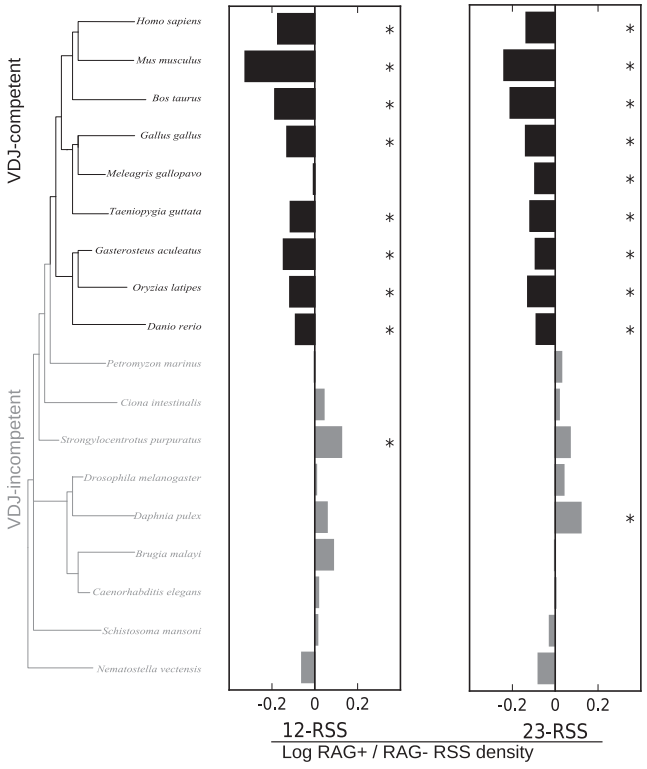

Fig. 5.—RSS deprivation from H3K4me3^{high} TSS is concomitant with the emergence of the recombinase. For each of the species represented on the tree (left panel), the log ratios of RSS densities at TSS of H3K4me3^{high} genes over H3K4me3^{low} genes was calculated for 12-RSS (middle panel) and 23-RSS (right panel), and represented by black bars for species VDJ-competent and by grey bars for species VDJ incompetent. Pearson's chi-square test was performed for each species and Benjamini–Hochberg correction was applied for the resulting *P*-values. * stands for *P*-values $< 10^{-2}$.

orthologues of all these RAG+ and RAG− genes (see methods and supplementary file S6, Supplementary Material online) to construct the equivalent gene sets for each species. Using the REC, we then computed the densities of RSS in the TSS regions and compared RAG+ and RAG− gene sets for each species (fig. 5). The species that have a V(D)J competent system show significant decrease in RSS density at TSS of RAG+ *versus* RAG- genes (top 9 lines) for at least one type of RSS, with the exception of 12-RSS in *Meleagris gallopavo*. In contrast, none of the species without a V(D)J competent system show decreased RSS density in the RAG+ when compared with the RAG− genes. Some closely related species display different levels of depletion of RSS in the RAG+ genes. These discrepancies may relate to the fact that our algorithm is trained with data biased toward certain species, and there is a small but non-negligible species-specific signal in the training set. It is also likely that imprecisions in orthologue identification and in genome annotation introduces noise in our analysis. Nevertheless, the overall trend is consistent across VDJ competent species, and trend conservation prevails over specific species differences in these type of complex analysis.

Overall, our results indicate that the depletion of RSS in the TSS of RAG+ genes is a feature that arose at the origin of jawed vertebrates, concomitantly with the emergence of RAG activity. In turn, these findings suggest that the presence of active *rag* genes creates a selective pressure in all jawed vertebrates to decrease RSS density specifically around TSS with high levels of H3K4me3 and where the RAG bind.

Conclusions

In a recent study, Teng and colleagues show that heptamers normally included in RSS are under-represented in the region surrounding the TSS of specific mouse and human genes. The sequences submitted to such selective pressure were identified through their binding of RAG1 at proximity of the TSS enriched in H3K4me3 mark. They proposed this feature to be a signature of an evolutionary response of the mouse and human genome to the RAG recombinase threat (Teng et al. 2015). The work we report here was initiated before the above-mentioned publication, and followed a different reasoning and approach. We develop a novel algorithm to directly address whether the RAG imposed a threat to the vertebrate genomes. Reasoning that only sequences exposed to the RAG with high frequency can be strongly affected by such putative evolutionary process, we dissociated *a priori* gene parts and gene sets. We used subsets of genes co-expressed or not with both RAG proteins and/or defined by their level of H3K4me3 load to show that H3K4me3 is a major factor correlated with the reduction of density of RSS at TSS from genes accessible to RAG. We reach similar conclusions as Teng and colleagues as what concern the mouse genome and produce novel evidence to sustain the evolutionary dimension of the selective pressure imposed by the cooption of the RAG recombinase by jawed vertebrates.

We developed an improved predictor of RSS, the REC, which is suitable for to use in a whole genome and multispecies context. Our analysis of RSS data from multiple species, and the observation that mutual information matrices do not change according to species reveals that the RSS is a conserved motif, which is in line with the high conservation of the RAG enzymes. An analysis of the mouse genome revealed a depletion of RSS in specific genomic segments. A comparative analysis of other genomes shows this tendency to be associated with the emergence of an adaptive immune system and the domestication of a transposable element to generate antibody/receptor diversity. This result strongly supports the notion that the presence of TE in a genome is a major shaper of this genome's evolution, not only by the direct genome alterations they introduce (Feschotte and Pritham 2007), but also by the defense mechanisms they elicit, as exemplified by previous studies on the RIP system in fungi (RIP; Galagan and Selker 2004). In the present case, the presence of a functional recombinase is linked to depletion of RSS from the genome of jawed vertebrates. Interestingly, the absence of depletion of RSS in sea urchin supports that the recombinase is not active in this species and supports the hypothesis that functional RAG activity was acquired later in evolution (Fugmann et al. 2006). In this context, we believe that depletion of RSS does not correlate with the entry of RAG1/2 in the genome, but rather with the retention of functional RAG activity after the horizontal transfer.

In addition, our results also show that depletion of RSS-like sequences occurs in regions of high H3K4 trimethylation, marking open chromatin for transcription, and tending to co-localize with TSS, as just reported by the Schatz group (Teng et al. 2015). Our evidence that this phenomenon is observed for cell lineages expressing RAG, and not in other tissues, strongly supports the claim that their depletion is a result of natural selection imposed by the activity of the RAG enzymes. As a consequence, genes that are normally expressed in lineages that do not express RAG would be particularly susceptible to ectopic RAG expression, as is observed in many malignancies (Zhang and Swanson 2008; Onozawa and Aplan 2012). Our unpublished results indicate that there are many areas of the human genome bearing high H3K4me3 loads that are predicted to contain both 12- and 23-RSS in *cis* flanking genes. Genetic variations of these sequences may render them effective cryptic RSS, in turn causing susceptibility to some cancer types.

Materials and Methods

Generation of a New Tool: The Recombination Classifier

Compilation of New Datasets

We used the ImMunoGeneTics database (IMGT) (Lefranc et al. 1999; Ruiz 2000; Lefranc 2001; Lefranc 2003; Lefranc et al.

2005, 2009) from July 2016 to build a repertoire of known functional RSS from 39 species (see supplementary file S1, Supplementary Material online). We used all the entries annotated by IMGT as functional or productive of exactly 28 and 39 nucleotides starting by the consensus CA dinucleotide to generate the positive training sets of 12- and 23-RSS, respectively. We retrieved the cRSS sequences from the literature that were functional in *in vitro* recombination assays to implement the data sets (Fuscoe et al. 1991; Lewis et al. 1997; Raghavan et al. 2001; Marculescu et al. 2002; Hu et al. 2015), and we filtered for non-redundant sequences to construct the final positive training sets, including 836 distinct 12-RSS and 1108 distinct 23-RSS from 39 species (see supplementary file S1, Supplementary Material online). We also built negative training sets for the 12-RSS and the 23-RSS, using *in vitro* recombination data from two studies (Lewis et al. 1997; Marculescu et al. 2002). In the first one, Marculescu and colleagues tested five regions of the mouse genome for the presence of cRSS, related with known cRSS (TAL2, LMO2, TAL1, BCL1 mtc, and BCL2 mbr), cloned in a pUC19 backbone vector [41]. In the second study, Lewis and coll. determined cRSS presence in pJH290 (based on pUC13) (Lewis et al. 1997). We extracted all non-redundant 28- or 39-bp subsequences from the tested genomic regions and backbone vectors starting with a CA dinucleotide that failed to exhibit recombination in the *in vitro* assays to generate the negative training sets (see supplementary file S1, Supplementary Material online). This resulted in 1266 and 1271 non-redundant sequences for the negative 12-RSS like sequence and 23-RSS like sequences, respectively. We found one 39bp sequence negative for recombination in Marculescu et al. (2002) that displayed a high RIC score (−37.30) and decided to test this sequence in our fluorescence based *in vitro* assay, the GFPi (Trancoso et al. 2013). This 39 bp sequence, when paired with a consensus 12-RSS, exhibits over 10% of recombination (see FACS plots on supplementary file S8, Supplementary Material online), which determined our choice to remove it from the negative set. As controls, additional sequences from the negative training set were tested *in vitro* and failed to show recombination in our assay (data not shown).

Fluorescence-Based In Vitro Recombination Assay

The CFP-GFPi construct and the cloning and analysis procedures were described in Trancoso et al. (2013). The RSS sequences of the consensus human 12-RSS and the 39bp sequence tested are CACAGTGATACAGACCCTAACAAAAACC (Murray et al. 2006) and CACAGTGATGTGCGGCCGCCTCCCCTCTGCACAG AAAGG (Yanisch-Perron et al. 1985), respectively.

Normalized Mutual Information

For both RSS categories, we calculated the pair-wise Normalized mutual information as defined in (Witten and Frank 2005). We did the previous calculation with the RSS dataset used in the original RIC (Cowell et al. 2002) and with our new RSS dataset. We then calculated the difference in mutual information as the matrix of mutual information of the new dataset minus the corresponding matrix of the original dataset.

RIC Score and Retrained RIC

All the calculations of the RIC score where performed using the published Perl scripts, training sets and thresholds from Cowell et al. (2002). We also retrained the RIC model using the RIC Perl scripts with our new positive training sets (rRIC), which we applied to both the positive and the negative datasets to define new thresholds. The 12- and 23-RSS thresholds correspond to the lower score computed from the corresponding positive dataset, ensuring no false positive in the negative training set of the same RSS category.

Implementing the RSS Characterization

We selected the most significant properties to use in CRoSSeD (for Conditional Random fields of Smoothed Structural Data) (Meysman et al. 2011), by finding the set of features that would maximize CRoSSeD power to separate our positive and negative training sets, using a greedy search strategy (supplementary file S9, Supplementary Material online). We retrieved all DNA related biophysical features from DiProDB (Friedel et al. 2009), which provides a value for each biophysical property for each dinucleotide, with a total of 110 DNA related properties. We calculated the values of each DNA sequence by replacing each dinucleotide for its relative value in the DiProDB database for each feature. We applied a greedy algorithm to select the best set of features to separate positive from negative sequences: for each of the 110 features we trained the CRoSSeD algorithm with all data and calculated the threshold that would maximize sensitivity whereas keeping specificity at maximum for that set of scores. We then selected and fixed the feature that exhibited the highest sensitivity, before iteration of the same process until no improvement was observed (supplementary file S9, Supplementary Material online).

Computing a New Sequence Classifier: The REC

We combined rRIC and CRoSSeD into a new classifier, the REcombination Classifier (REC). We chose threshold scores of −35.32 and −54.18 for rRIC scores of 12- and 23-RSS, respectively, and 0.999681 and 0.99974 for CRoSSeD scores of 12- and 23-RSS, respectively (vertical and horizontal lines on fig. 1*B* and *C*) to maximize sensitivity whereas imposing maximum specificity (vertical and horizontal lines on fig. 1*B* and *C*). We also determined a pair of thresholds that would maximize specificity whereas imposing maximum sensitivity (available option of the REC tool). A score is calculated for each

sequence that is the average of the normalized rRIC and CRoSSeD scores. The REC is available at http://www.evocell.org/cgl/resources.

Genomes Partitions

V(D)J Regions

We retrieved from Ensembl all annotated gene segments and pseudogenes from T and B cell receptors (TR and IG respectively) from *Homo sapiens*, *Mus musculus*, *Bos taurus*, *Gallus gallus* and *Danio rerio* (Flicek et al. 2013). We used Ensembl V73 (Flicek et al. 2013) to retrieve the genome sequences and predicted orthologue genes of this set to identify all orthologue TR and IG gene segments and pseudogenes from *Homo sapiens* (human), *Mus musculus* (mouse), *Bos taurus* (cow), *Gallus gallus* (chicken), *Meleagris gallopavo* (turkey), *Taeniopygia guttata* (zebra finch), *Gasterosteus aculeatus* (stickleback), *Oryzias latipes* (medaka), *Petromyzon marinus* (lamprey), *Ciona intestinalis* (vase tunicate), *Caenorhabditis elegans* (nematode), and *Drosophila melanogaster* (fruit fly); and we used Ensembl Metazoa V20 (Kersey et al. 2010) for *Strongylocentrotus purpuratus* (sea urchin), *Daphnia pulex* (water flea), *Brugia malayi* (agent of lymphatic filariasis), *Schistosoma mansoni* (blood fluke), and *Nematostella vectensis* (starlet sea anemone). For the Metazoa and for *Danio rerio* (zebrafish), we mapped from the predicted orthologues for *Drosophila melanogaster*. The V(D)J regions from each species were defined by a 800 bp gate centered on the coordinates of each segment. For each genome we discarded mitochondrial and V(D)J regions. In the genomes for which chromosomes where built, we discarded all DNA outside chromosomes.

Definition of Genes and Gene Bodies

For each of the previous retrieved genomes we used Ensembl database to map all protein coding genes, using their transcription start and end sites to delimitate the genes coordinates. We used Ensembl exons annotations to define exons and introns coordinates and we defined the TSS region as the 1Kb sequence centered on the TSS.

Selection of Expressed Genes

We used the Gene Expression Barcode V2.0 to obtain the list of genes expressed at the same stages as the recombinase (McCall et al. 2011). *Rag1* and *rag2* genes are expressed at the CD4 and CD8 double negative (DN) and double positive (DP) stages in T cells, and at the pre- and pro-B stages in B cells. We used published transcriptomic data from these cell stages generated with the platform Affymetrix Mouse Genome 430 2.0 Array on NCBI GEO DataSets. We analyzed the following samples: GSM312035, GSM312036, and GSM312037 from Bordon et al. (2008), GSM492807, GSM49280, GSM492809, and GSM492810 from Zetterblad et al. (2010), GSM188611, GSM188612, GSM188613, GSM188614, GSM188615, GSM188616, GSM188617, and GSM188618 from Kawazu et al. (2007), and GSM713993, GSM713994 and GSM713995 from Ren and Cowell (2011). We applied the Barcode algorithm to all samples and mapped all tags to the respective Ensembl genes. For each of the four cell types, we defined the expressed genes as the ones exhibiting at least one positive tag in each sample. The final set of genes co-expressed with the recombinase in V(D)J competent organisms was defined by the union of the co-expressed genes of each cell types (coRAG+, 6153 genes). To select genes that are never co-expressed with *rag1* and, we selected the subset of genes that have no positive tag in any of the previous samples *rag2* (coRAG−; 9433 genes).

Selection of Specific Genes and Housekeeping Genes

We used the Gene Expression Barcode V2.0 to define the expressed genes in kidney, lung, liver and skeletal muscle. We retrieved all the genes expressed in the consensus tissue (expressed in 95% of the samples) from kidney, lung, liver, and skeletal muscle. By removing all genes expressed in any of these four tissues from the RAG+ set, we defined the RAG specific set of genes (RS+, 2730 genes). We followed the same approach to define the set of genes specific of each tissue. A set of genes specific of non-RAG expressing tissues was computed by the union of the specific genes from kidney, lung, liver, and skeletal muscle (TS+; 1256 genes). The housekeeping genes (HK) were defined by the intersection of the genes expressed in all four tissues and the ones co-expressed with the recombinase (HKRAG+; 1575 genes). The set of negative genes for non-lymphoid tissues (TS−) was defined by genes neither expressed in kidney, lung, liver nor skeletal muscle (14,227 genes).

H3K4me3 Data

We used H3K4me3 data from DN thymocytes (Pekowska et al. 2011), and mapped the reads to the mouse genome (mm9) with BWA (Li and Durbin 2009), using the default parameters, and subtracted the input value for each position. For each gene, we calculated the total number of reads in 1kb gates centered on the TSS. We then computed the 33° and 67° percentiles (bottom 33% and top 33%) to discriminate between genes exhibiting low *versus* high H3K4me3 loads respectively (H3K4me3low, 6642 genes vs. H3K4me3high, 6638 genes). We calculated the densities of 12-RSS and 23-RSS at each TSS of H3K4me3low and H3K4me3high genes. Sub-sequentially, we selected the intersection between H3K4me3high and coRAG+ genes (3067 genes) and compared the distributions of RSS densities at the TSS with the same measures for the genes in the intersection of H3K4me3high and coRAG− subsets (2354 genes).

RSS Density in Different Partitions of the Genome

RSS Density in Genes

We selected all non-overlapping gene regions from the mouse genome, and computed the number of RSS of these gene regions using the REC. We repeated this process for the non-gene regions. We estimate the expected RSS count in gene regions by calculating the percentage of genome belonging to gene regions and multiply that fraction by the total amount of RSS in the genome. The same procedure was used for non-gene regions.

RSS Density along Genes

We used the previously defined gene bodies (exons, introns and TSS region) of the RAG+ and RAG− sets of genes and computed the density of 12-RSS and 23-RSS in each set of genes, using the REC. We excluded overlapping regions between RAG+ and RAG− genes and between any of the gene sets and the V(D)J regions.

Selection of Genes for Multi-Species Comparison

The positive set for mouse was selected using the intersection of our H3K4me3^{high} gene pool with the pool of RAG1 bound genes previously described (Teng et al. 2015). The negative set was obtained by removing any of these genes from our H3K4me3^{low} genes set. For the remaining species, we mapped orthologues using the ensemble orthologue database, which is based on phylogentic inference. For species without direct orthologues to mouse in ensemble, we first mapped the orthologues in Drosophila and used these to identify the orthologue genes in each species. For the number of sequences analyzed, see supplementary file S7 and table S3, Supplementary Material online.

Statistical Analysis

All statistical comparisons were done using python scipy package (Jones et al. 2001). In the cross-validation analyses to perform pairwise comparisons between distributions, we used Wilcoxon signed-rank test (supplementary file S4, Supplementary Material online).

To compare the sensitive of RIC, rRIC, and REC, we use McNemar's test (McNemar 1947). The *P*-values are presented on tables on supplementary file S3, Supplementary Material online.

To test for the depletion of RSS in different genomic regions in the mouse (figs. 2–4) we calculated the RSS density for each TSS in each group and tested if one distribution had larger values than the other using a Mann–Whitney U test, and corrected the *P*-values using Bonferroni's correction with the total number of tests in each figure.

In the multi-species analysis (fig. 5), we counted the total number of RSS in all the TSS of each group and compared it with the expected number of RSS using Person's chi-square test, and *P*-values were corrected using Benjamini–Hochberg procedure. The expected number of RSS for the RAG+ genes is calculated with the total number of observed RSS time the total size of RAG+ TSS regions divided by the total size of analyzed regions (RAG+ plus RAG− TSS).

Phylogenetic Signal

For both RSS categories, we gathered the nonredundant set of sequences annotated as functional RSS for each species. We computed the pairwise number of nucleotide differences and used the resulting matrix as input in python scipy hierarchical cluster module to compute the sequence cluster, with single-linkage and number of nucleotide differences as distance metric. We then computed the flat clusters using maximum number of clusters equal to the number of species represented in the set. We compared this set of clusters with the results obtained when using the species label to cluster the sequences and computed the adjusted Rand index (ARI). The resulting Adjusted Rand Index of 0.0003 and −0.0006 (12- and 23-RSS, respectively) indicate very weak species-specificity of RSS sequences.

Supplementary Material

Supplementary tables S1–S3 and figures S1–S6 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

We wish to thank Thiago Carvalho, Jorge Carneiro, and members of our laboratories for helpful discussions. The authors acknowledge funding by the Instituto Gulbenkian de Ciência and by the Fundação para a Ciência e Tecnologia via Grant PTDC-BIA-GEN-116830-2010. DS was funded by a Optimus alive award, MB by a postdoctoral FCT fellowship SFRH/BPD/65292/2009, and IT by a PhD fellowship SFRH/BD/51179/2010. The authors declare they do not have any competing interests.

Literature Cited

Akira S, Okazaki K, Sakano H. 1987. Two pairs of recombination signals are sufficient to cause immunoglobulin V-(D)-J joining. Science 238:1134–1138.

Alzohairy AM, Gyulai G, Jansen RK, Bahieldin A. 2013. Transposable elements domesticated and neofunctionalized by eukaryotic genomes. Plasmid 69:1–15.

Amit M, et al. 2012. Differential GC content between exons and introns establishes distinct strategies of splice-site recognition. Cell Rep. 1:543–556.

Bordon A, et al. 2008. Enforced expression of the transcriptional coactivator OBF1 impairs B cell differentiation at the earliest stage of development. PLoS One 3:e4007.

Bulyk M, Levine M, Bourque G. 2009. Transposable elements in gene regulation and in the evolution of vertebrate genomes. Curr Opin Genet Dev. 19:607–612.

Clatworthy AE, Valencia MA, Haber JE, Oettinger MA. 2003. V(D)J recombination and RAG-mediated transposition in yeast. Mol Cell 12: 489–499.
Cowell LG, Davila M, Kepler TB, Kelsoe G. 2002. Identification and utilization of arbitrary correlations in models of recombination signal sequences. Genome Biol. 3:research0072.1–research0072.20.
Dik WA, et al. 2007. Different chromosomal breakpoints impact the level of LMO2 expression in T-ALL. Blood 110:388–392.
Feschotte C, Pritham EJ. 2007. DNA transposons and the evolution of eukaryotic genomes. Annu Rev Genet. 41:331–368.
Flicek P, et al. 2013. Ensembl 2013. Nucleic Acids Res. 41:D48–D55.
Friedel M, Nikolajewa S, Sühnel J, Wilhelm T. 2009. DiProDB: a database for dinucleotide properties. Nucleic Acids Res. 37:D37–D40.
Fugmann SD, Messier C, Novack LA, Cameron RA, Rast JP. 2006. An ancient evolutionary origin of the Rag1/2 gene locus. Proc Natl Acad Sci U S A. 103:3728–3733.
Fuscoe JC, et al. 1991. V(D)J recombinase-like activity mediates hprt gene deletion in human fetal T-lymphocytes. Cancer Res. 51:6001–6005.
Galagan JE, Selker EU. 2004. RIP: the evolutionary cost of genome defense. Trends Genet. 20:417–423.
Gopalakrishnan S, et al. 2013. Unifying model for molecular determinants of the preselection Vβ repertoire. Proc Natl Acad Sci U S A. 110:E3206–E3215.
Ha SC, Lowenhaupt K, Rich A, Kim Y-G, Kim KK. 2005. Crystal structure of a junction between B-DNA and Z-DNA reveals two extruded bases. Nature 437:1183–1186.
Hu J, et al. 2015. Chromosomal loop domains direct the recombination of antigen receptor genes. Cell 163:947–959.
Huang S, et al. 2014. Decelerated genome evolution in modern vertebrates revealed by analysis of multiple lancelet genomes. Nat Commun. 5:5896.
Huang S, et al. 2016. Discovery of an active RAG transposon illuminates the origins of V(D)J recombination. Cell 166:102–114.
Jones E, Oliphant T, Peterson P. 2001. {SciPy}: Open source scientific tools for {Python}. [cited 25 Oct 2016] http://www.scipy.org/.
Kawazu M, et al. 2007. Expression profiling of immature thymocytes revealed a novel homeobox gene that regulates double-negative thymocyte development. J Immunol. 179:5335–5345.
Kersey PJ, et al. 2010. Ensembl Genomes: extending Ensembl across the taxonomic space. Nucleic Acids Res. 38:D563–D569.
Koudritsky M, Domany E. 2008. Positional distribution of human transcription factor binding sites. Nucleic Acids Res. 36:6795–6805.
Lauberth SM, et al. 2013. H3K4me3 interactions with TAF3 regulate preinitiation complex assembly and selective gene activation. Cell 152:1021–1036.
Lefranc M-P, et al. 1999. IMGT, the international ImMunoGeneTics database. Nucleic Acids Res. 27:209–212.
Lefranc M-P. 2001. IMGT, the international ImMunoGeneTics database. Nucleic Acids Res. 29:207–209.
Lefranc M-P. 2003. IMGT, the international ImMunoGeneTics database. Nucleic Acids Res. 31:307–310.
Lefranc M-P, et al. 2009. IMGT, the international ImMunoGeneTics information system. Nucleic Acids Res. 37:D1006–D1012.
Lefranc M-P, et al. 2005. IMGT, the international ImMunoGeneTics information system. Nucleic Acids Res. 33:D593–D597.
Lewis SM. 1994. The mechanism of V(D)J joining: lessons from molecular, immunological, and comparative analyses. Adv Immunol. 56:27–150.
Lewis SM, Agard E, Suh S, Czyzyk L. 1997. Cryptic signals and the fidelity of V(D)J joining. Mol Cell Biol. 17:3125–3136.
Lewis SM, Wu GE. 1997. The origins of V(D)J recombination. Cell 88:159–162.
Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760.
Lieber MR, et al. 2006. Recombinase, chromosomal translocations and lymphoid neoplasia: targeting mistakes and repair failures. DNA Repair (Amst) 5:1246–1258.
Long HK, et al. 2013. Epigenetic conservation at gene regulatory elements revealed by non-methylated DNA profiling in seven vertebrates. Elife 2:e00348.
Marculescu R, Le T, Simon P, Jaeger U, Nadel B. 2002. V(D)J-mediated translocations in lymphoid neoplasms: a functional assessment of genomic instability by cryptic sites. J Exp Med. 195:85–98.
Matthews AGW, et al. 2007. RAG2 PHD finger couples histone H3 lysine 4 trimethylation with V(D)J recombination. Nature 450:1106–1110.
McCall MN, Uppal K, Jaffee HA, Zilliox MJ, Irizarry RA. 2011. The Gene Expression Barcode: leveraging public data repositories to begin cataloging the human and murine transcriptomes. Nucleic Acids Res. 39:D1011–D1015.
McNemar Q. 1947. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 12:153–157.
Merelli I, et al. 2010. RSSsite: a reference database and prediction tool for the identification of cryptic Recombination Signal Sequences in human and murine genomes. Nucleic Acids Res. 38:W262–W267.
Meysman P, et al. 2011. Use of structural DNA properties for the prediction of transcription-factor binding sites in *Escherichia coli*. Nucleic Acids Res. 39:e6.
Mijuskovic M, et al. 2015. Off-target V(D)J recombination drives lymphomagenesis and is escalated by loss of the Rag2 C terminus. Cell Rep. 12:1842–1852.
Murray JM, et al. 2006. V(D)J recombinase-mediated processing of coding junctions at cryptic recombination signal sequences in peripheral T cells during human development. J Immunol. 177:5393–5404.
Onozawa M, Aplan PD. 2012. Illegitimate V(D)J recombination involving nonantigen receptor loci in lymphoid malignancy. Genes Chromosom Cancer 51:525–535.
Papaemmanuil E, et al. 2014. RAG-mediated recombination is the predominant driver of oncogenic rearrangement in ETV6-RUNX1 acute lymphoblastic leukemia. Nat Genet. 46:116–125.
Pekowska A, et al. 2011. H3K4 tri-methylation provides an epigenetic signature of active enhancers. EMBO J. 30:4198–4210.
Raghavan SC, Kirsch IR, Lieber MR. 2001. Analysis of the V(D)J recombination efficiency at lymphoid chromosomal translocation breakpoints. J Biol Chem. 276:29126–29133.
Raghavan SC, Swanson PC, Wu X, Hsieh C-L, Lieber MR. 2004. A non-B-DNA structure at the Bcl-2 major breakpoint region is cleaved by the RAG complex. Nature 428:88–93.
Ren M, Cowell JK. 2011. Constitutive Notch pathway activation in murine ZMYM2-FGFR1-induced T-cell lymphomas associated with atypical myeloproliferative disease. Blood 117:6837–6847.
Ruiz M. 2000. IMGT, the international ImMunoGeneTics database. Nucleic Acids Res. 28:219–221.
Sadofsky MJ. 2001. The RAG proteins in V(D)J recombination: more than just a nuclease. Nucleic Acids Res. 29:1399–1409.
Schatz DG, Ji Y. 2011. Recombination centres and the orchestration of V(D)J recombination. Nat Rev Immunol. 11:251–263.
Schlissel MS, Kaffer CR, Curry JD. 2006. Leukemia and lymphoma: a cost of doing business for adaptive immunity. Genes Dev. 20:1539–1544.
Shimazaki N, Lieber MR. 2014. Histone methylation and V(D)J recombination. Int J Hematol. 100:230–237.
Shimazaki N, Tsai AG, Lieber MR. 2009. H3K4me3 stimulates the V(D)J RAG complex for both nicking and hairpinning in trans in addition to tethering in cis: implications for translocations. Mol Cell. 34:535–544.
Teng G, et al. 2015. RAG represents a widespread threat to the lymphocyte genome. Cell 162(2014):751–765.
Trancoso I, et al. 2013. A novel quantitative fluorescent reporter assay for RAG targets and RAG activity. Front Immunol. 4:110.

Witten IH, Frank E. 2005. Data mining?: practical machine learning tools and techniques. Morgan Kaufman, Burlington, MA.
Yancopoulos GD, Alt FW. 1986. Regulation of the assembly and expression of variable-region genes. Annu Rev Immunol. 4:339–368.
Yanisch-Perron C, Vieira J, Messing J. 1985. Improved M13 phage cloning vectors and host strains: nucleotide sequences of the M13mp18 and pUC19 vectors. Gene 33:103–119.
Zemach A, McDaniel IE, Silva P, Zilberman D. 2010. Genome-wide evolutionary analysis of eukaryotic DNA methylation. Science 328(80): 916–919.
Zetterblad J, et al. 2010. Genomics based analysis of interactions between developing B-lymphocytes and stromal cells reveal complex interactions and two-way communication. BMC Genomics 11:108.
Zhang M, Swanson PC. 2008. V(D)J recombinase binding and cleavage of cryptic recombination signal sequences identified from lymphoid malignancies. J Biol Chem. 283:6717–6727.
Zhu J, He F, Hu S, Yu J. 2008. On the nature of human housekeeping genes. Trends Genet. 24:481–484.

Associate editor: Michelle Meyer