# Transposon molecular domestication and the evolution of the RAG recombinase

Yuhang Zhang<sup>1,8</sup>, Tat Cheung Cheng<sup>2,8</sup>, Guangrui Huang<sup>3</sup>, Qingyi Lu<sup>3</sup>, Marius D. Surleac<sup>4</sup>, Jeffrey D. Mandell<sup>1</sup>, Pierre Pontarotti<sup>5,6</sup>, Andrei J. Petrescu<sup>4</sup>, Anlong Xu<sup>3,7</sup>\*, Yong Xiong<sup>2</sup>\* & David G. Schatz<sup>1</sup>\*

Domestication of a transposon (a DNA sequence that can change its position in a genome) to give rise to the RAG1-RAG2 recombinase (RAG) and V(D)J recombination, which produces the diverse repertoire of antibodies and T cell receptors, was a pivotal event in the evolution of the adaptive immune system of jawed vertebrates. The evolutionary adaptations that transformed the ancestral RAG transposase into a RAG recombinase with appropriately regulated DNA cleavage and transposition activities are not understood. Here, beginning with cryo-electron microscopy structures of the amphioxus ProtoRAG transposase (an evolutionary relative of RAG), we identify amino acid residues and domains the acquisition or loss of which underpins the propensity of RAG for coupled cleavage, its preference for asymmetric DNA substrates and its inability to perform transposition in cells. In particular, we identify two adaptations specific to jawed-vertebrates— arginine 848 in RAG1 and an acidic region in RAG2—that together suppress RAG-mediated transposition more than 1,000-fold. Our findings reveal a two-tiered mechanism for the suppression of RAG-mediated transposition, illuminate the evolution of V(D)J recombination and provide insight into the principles that govern the molecular domestication of transposons.

Jawed vertebrates have evolved a sophisticated adaptive immune system that relies on the assembly of immunoglobulin and T-cell receptor genes from arrays of V, D and J gene segments in developing B and T lymphocytes. The assembly reaction, known as V(D)J recombination, is initiated when RAG cleaves adjacent to the gene segments at recombination signal sequences (RSSs) that comprise conserved heptamer and nonamer elements separated by a 12- or 23-base pair (bp) spacer (12RSS and 23RSS, respectively)<sup>1</sup> (Fig. 1a). DNA cleavage by RAG occurs by a nick–hairpin mechanism in which hairpin formation occurs in a coordinated (coupled) manner in a synaptic complex that contains one 12RSS and one 23RSS, a restriction known as the 12/23 rule (Fig. 1b). The 12/23 rule and coupled cleavage are fundamental features of RAG that are thought to contribute to the proper orchestration of V(D)J recombination and protection of genome integrity<sup>1–3</sup>.

The 'molecular domestication' of transposons has contributed broadly to the evolution of new proteins and activities<sup>4–6</sup>, with RAG and V(D)J recombination representing a paradigmatic example of this process. The current evidence supports a model in which *RAG1* and *RAG2* evolved from the transposase genes of an ancient 'RAG transposon', while disassembled ('split') immunoglobulin and T cell receptor genes arose from the insertion of a transposon into a receptor gene, with the inserted terminal inverted repeats (TIRs) of the transposon becoming the RSSs<sup>7–9</sup>. This model was strongly supported by the discovery, in the cephalochordate amphioxus, of ProtoRAG—a transposon with numerous features that implicate it as a descendent of the RAG transposon<sup>10</sup>.

The RAG transposon domestication model predicts a critical divergence during chordate evolution in which, in jawed vertebrates, the RAG transposase acquired properties of a recombinase, whereas in amphioxus (and probably other invertebrate chordate lineages<sup>11</sup>), transposase functions were retained. A divergence in post-cleavage reaction steps (Fig. 1b) would have been particularly pivotal, with RAG-generated DNA ends preferentially undergoing end joining (recombination) instead of transposition, and ProtoRAG retaining a strong preference for transposition over end-joining<sup>10</sup>. Indeed, RAG is notably poor at performing transposition in living cells<sup>12–14</sup>, with only a single bone fide transposition event thus far identified in mice or humans<sup>15,16</sup>. How the ancestral RAG transposon was domesticated to yield a RAG recombinase with minimal in vivo transposition activity and a strong propensity for coupled cleavage of asymmetric substrates is a key question in the evolution of V(D)J recombination and adaptive immunity in jawed vertebrates. Here, we use the structure of ProtoRAG transposase as a lens through which to view this evolutionary transformation.

#### Uncoupled DNA cleavage by ProtoRAG

*ProtoRAG* from *Branchiostoma belcheri* (Extended Data Fig. 1a) is composed of convergently transcribed *RAG1*-like (Bb*RAG1L*) and *RAG2*-like (Bb*RAG2L*) genes flanked by 5' and 3' TIRs that comprise a heptamer similar to the RSS heptamer, an adjacent, conserved 9–10bp element referred to as TIR region 2 (TR2) and additional flanking sequences<sup>10</sup> (Fig. 1a). The *Bb*RAG1L protein contains a 'core' region (*cBb*RAG1L; amino acids (aa) 468–1136) with sequence similarity (33% aa identity) to the core region of RAG1 (cRAG1; aa 384–1008 in mouse; Fig. 1c). Within cRAG1 and *cBb*RAG1L, we define catalytic cores (CC and CC\*, respectively) that lack one or more DNA binding elements (Fig. 1c). *Bb*RAG2L resembles only core RAG2 (aa 1–350 in mouse; 22% aa identity) and lacks all RAG2 C-terminal elements, including an acidic hinge and plant homeodomain finger (Fig. 1c).

With a cleavage substrate containing a 5'/3'TIR pair, both core *Bb*RAGL (*cBb*RAG1L with *Bb*RAG2L; Fig. 1d) and full-length *Bb*RAG1L–*Bb*RAG2L (Extended Data Fig. 1b) generate a strong band that corresponds to single cleavage at the 3'TIR (black asterisk) and

<sup>&</sup>lt;sup>1</sup>Department of Immunobiology, Yale School of Medicine, New Haven, CT, USA. <sup>2</sup>Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT, USA. <sup>3</sup>Beijing University of Chinese Medicine, Beijing, China. <sup>4</sup>Department of Bioinformatics and Structural Biochemistry, Institute of Biochemistry of the Romanian Academy, Bucharest, Romania. <sup>5</sup>Aix Marseille Univ IRD, APHM, MEPHI, IHU Méditerranée Infection, Marseille, France. <sup>6</sup>Centre National de la Recherche Scientifique, Marseille, France. <sup>7</sup>State Key Laboratory of Biocontrol, Guangdong Province Key Laboratory of Pharmaceutical Functional Genes, Department of Biochemistry, School of Life Sciences, Sun Yat-sen University, Higher Education Mega Center, Guangzhou, China. <sup>8</sup>These authors contributed equally: Yuhang Zhang, Tat Cheung Cheng. \*e-mail: xuanlong@bucm.edu.cn; yong.xiong@yale.edu; david.schatz@yale.edu



Fig. 1 | Uncoupled DNA cleavage by BbRAGL. a, RSS and TIR substrates. Underlining indicates sequence identity in TR2; arrows indicate sites of cleavage. The TIR heptamer sequence can be found in endogenous human RSS sequences<sup>38</sup>. **b**, Schematic of DNA cleavage, recombination and transposition by RAG or BbRAGL. Inset, nick-hairpin mechanism of DNA cleavage. Triangle, RSS or TIR (here and in other Figs.; wide side of triangle is heptamer end). c, Domain diagrams of the RAG and BbRAGL proteins. PHD, plant homeodomain finger. Numbers indicate amino acid domain boundaries (mouse RAG is depicted and used in all experiments except where indicated). NBD\* is named for consistency and not to imply function. d, Time course of DNA cleavage by cRAG and cBbRAGL with substrates containing a pair of TIRs or RSSs. Red asterisk, double cleavage band; black asterisk, single 3'TIR cleavage band. Cleavage products were resolved on acrylamide gels and are indicated schematically (circles indicate hairpin ends). For gel source data, see Supplementary Fig. 1. e, f, Cleavage of substrates containing two or one TIRs or RSSs as indicated above the lanes, by cBbRAGL, cRAG or complexes in which cBbRAG1L or cRAG1 lack the indicated domain. Black asterisks mark the two single cleavage products. A reaction time of 60 min was used here and in other cleavage reactions unless otherwise specified. g, Transposition frequency measured in HEK293T cells using the assay shown in Extended Data Fig. 8a.

is comparable in intensity to the 5'/3'TIR double cleavage band (red asterisk). By contrast, core RAG (cRAG) predominantly generates the 12/23RSS double cleavage product (Fig. 1d). Furthermore, both core and full-length *Bb*RAGL robustly cleave substrates containing either a single 5'TIR or a single 3'TIR (Fig. 1e, Extended Data Fig. 1b) whereas cRAG cleaves single RSS substrates poorly (Fig. 1f). These results indicate that DNA cleavage by *Bb*RAGL is less tightly coupled than cleavage by RAG.

Deletion of the nonamer-binding domain (NBD) from cRAG1 eliminates RAG activity (Fig. 1f), whereas *cBb*RAGL that lacks its corresponding NBD\* domain (which has limited sequence similarity to NBD but, similar to the NBD<sup>17</sup>, forms a dimer in solution; Extended Data Fig. 1c, d) retains substantial activity (Fig. 1e). In addition, although the C-terminal tail (CTT) of RAG1 is dispensable for activity<sup>2</sup>, the C-terminal tail of *Bb*RAG1L (CTT\*; Fig. 1c) is important for *Bb*RAGL cleavage activity<sup>10</sup> (Fig. 1e). Therefore, RAG1 and *Bb*RAG1L have evolved different dependencies on the N- and C-terminal portions of their core regions.

RAG and *Bb*RAGL show a third important difference: while both are active transposases in vitro<sup>10,18,19</sup>, only *Bb*RAGL exhibits substantial

transposition activity in cells<sup>10,12–16</sup> (Fig. 1g). We note that *Bb*RAGL activity is assessed here in a heterologous (mammalian) cell context.

#### Structure of the ProtoRAG transposase

To better understand these functional differences, we determined the structure of *cBb*RAGL together with HMGB1 (a DNA-bending cofactor that stimulates cleavage by RAG<sup>1</sup> and *Bb*RAGL<sup>10</sup>) bound to the 3'TIR (which is bound more efficiently than the 5'TIR by the *cBb*RAGL tetramer; Extended Data Fig. 2a–c). Single-particle cryo-electron microscopy (cryo-EM) analysis yielded structures for *cBb*RAGL–HMGB1 bound to intact and nicked 3'TIRs (designed to mimic the first step of cleavage as in Fig. 1b, inset), with a resolution of 4.3 Å for the nicked 3'TIR structure after application of two-fold symmetry (Fig. 2a, Extended Data Fig. 2d–f, Extended Data Table 1).

cBbRAGL-3'TIR complexes contain a central cBbRAG1L dimer capped by two monomers of BbRAG2L and two DNA duplexes (Fig. 2a, Extended Data Fig. 2g, h). Rather than the Y shape adopted by cRAG complexes<sup>20-23</sup> (Extended Data Fig. 3a), cBbRAGL complexes were roughly V-shaped because no density was discernable for NBD\*, HMGB1 or the heptamer-distal 25 bp of the TIRs. Despite an estimated 700 million years of evolutionary divergence, ProtoRAG and RAG exhibit a marked degree of structural similarity. cBbRAG1L recapitulates the structural domains of the RAG1 catalytic core, and BbRAG2L-similar to RAG2—adopts a structure consistent with a six-bladed β-propeller fold (Fig. 2b, Extended Data Fig. 3b-d). Structural similarity is clear in the vicinity of the active site and heptamer, with the nicked 3'TIR exhibiting two flipped (extrahelical) bases similar to those of nicked RSSs bound by RAG<sup>21,22</sup> (Fig. 2c, Extended Data Fig. 3e). Similar to RAG<sup>21-23</sup>, BbRAGL switches from an 'open' to a 'closed' conformation upon TIR nicking and is particularly flexible in the BbRAGL-intact TIR complex (Extended Data Fig. 3f-h), with both molecules of BbRAG1L making extensive contacts with both DNA molecules (Extended Data Fig. 4). These notable structural parallels support the hypothesis that RAG and BbRAGL evolved from a common RAG transposon ancestor.

#### A novel ProtoRAG DNA-binding domain

The cryo-EM map of the complex of *cBb*RAGL and nicked 3'TIR contained an unaccounted-for density at the C terminus of the BbRAG1L catalytic core that could readily accommodate the C $\alpha$  backbone of BbRAG1L CTT\* (Extended Data Fig. 5a) and was in close proximity to the TR2 element of the 3'TIR (Fig. 2d). This suggested that CTT\* is a DNA-binding domain that-together with residues from the opposite subunit of *Bb*RAG1L—forms a clamp to bind TR2 (Fig. 2e). CTT\* exhibits sequence conservation among RAG1-like proteins from invertebrates (Fig. 2f) but not with vertebrate RAG1 CTT (Extended Data Fig. 5b). Mutation of residues Cys1114, His1222 or Cys1227 in the highly conserved CX<sub>2</sub>CX<sub>3</sub>GHX<sub>4</sub>C motif of CTT\* reduced activity, whereas mutation of two less well-conserved residues had no discernable effect on activity (Fig. 2g), which indicates that integrity of the conserved CCHC motif is important for CTT\* function. Consistent with the hypothesis that *Bb*RAGL makes important contacts with TR2, the heptamer and TR2 are the only portions of the TIR that are essential for cleavage (Extended Data Fig. 5c-g).

#### Modular domain function and the 12/23 rule

To investigate how the distinct functional properties of RAG and *Bb*RAGL relate to structural domains, we generated chimeric RAG1– *Bb*RAG1L proteins in which the *Bb*RAG1L catalytic core—with or without CTT\*—was fused with the RAG1 NBD (Fig. 3a, b) and, reciprocally, the RAG1 catalytic core was fused to NBD\* and/or CTT\* of *Bb*RAG1L (Fig. 3c). Corresponding hybrid RSS–TIR DNA targets were used as cleavage substrates (Fig. 3b, c).

When supplied with the RAG1 NBD, the *Bb*RAG1L catalytic core no longer required CTT\* (Fig. 3d) and became dependent on the RSS nonamer for activity (Fig. 3e), with spacer length requirements  $(12 \pm 1 \text{ bp or } 23 \pm 1 \text{ bp})$  identical to those of RAG (Extended Data Fig. 6a, b). Reciprocally, when deprived of its NBD, the RAG1



**Fig. 2** | **Cryo-EM structure of cBbRAGL-nicked 3'TIR complex. a**, Symmetrized cryo-EM structure of the cBbRAGL-HMGB1-nicked 3'TIR complex at 4.3 Å resolution. BbRAG1L-a and BbRAG1L-b are defined as the subunits that interact with BbRAG2L-a and BbRAG2L-b, and which have active sites that engage 3'TIR-a and 3'TIR-b, respectively. **b**, Superimposition of one subunit of cBbRAG1L (nicked-3'TIR complex) with cRAG1 (nicked RSS complex; RCSB Protein Data Bank code (PDB) 5ZDZ). **c**, Superimposition of nicked RSS (PDB 5ZDZ) with nicked 3'TIR showing flipped bases and three catalytic residues and calcium ions

catalytic core became dependent on CTT\* and TR2 for activity and was active without NBD\* or any portion of the DNA substrate except the heptamer and TR2 (Fig. 3f-h). Thus, CTT\* renders the RAG1 catalytic core independent of an NBD, the RSS nonamer, substrate asymmetry and-hence-the 12/23 rule. Notably, proteins that contain the BbRAG1L catalytic core exhibit uncoupled cleavage (Fig. 3d) whereas those that contain the RAG1 catalytic core display coupled cleavage (Extended Data Fig. 6c, d). We conclude that the catalytic cores of RAG1 and BbRAG1L dictate the propensities of these enzymes for coupled versus uncoupled cleavage, and that the functional organization of ProtoRAG TIRs is different from that of RSSs because of a dependency on different DNA-binding domains (Extended Data Fig. 6e). Furthermore, our findings argue that the choice of dominant DNA-binding domain was pivotal for the evolution of the 12/23 rule, as CTT\* would need to have been eliminated to allow dependency on the rule.



**Fig. 3** | **DNA cleavage properties of chimeric RAG1**–*Bb***RAG1L proteins. a**–**c**, Schematics of *cBb*RAG1L and cRAG1 (**a**), and chimeric proteins containing the catalytic core of *Bb*RAG1L (**b**) or RAG1 (**c**) with matching chimeric cleavage targets. In targets T1 and T2, the heptamer derives from the TIR and the remainder from the 12/23RSS; in T3 and T4, the heptamer derives from the RSS and the remainder from the 5'/3'TIR. **d**–**h**, Cleavage reactions using chimeric proteins and substrates containing one or two targets as indicated above and below the lanes. ΔNon, nonamer region deleted; T1\*, T1 with a C-to-A mutation of heptamer position 1, which renders the target uncleavable; mTR2, scrambling of TR2 in both target sites. Asterisks as in Fig. 2g.

(spheres) in the active site of RAG or *Bb*RAGL. **d**, **e**, The additional density at the C terminus of *Bb*RAG1L is in close proximity to TR2 (orange; **d**) and, with the opposite subunit of *Bb*RAG1L, largely encircles the DNA (**e**). **f**, Sequence alignment of CTT\* from deuterostome invertebrate RAG1L and cnidarian (hydra) Transib proteins. Species name abbreviations are defined in the legend of Extended Data Fig. 5b. Asterisks, conserved residues with Zn<sup>2+</sup> coordination potential. **g**, Cleavage reactions using CTT\* mutants of *Bb*RAG1L. Red asterisk, double cleavage band; black asterisks, single TIR cleavage bands.

#### Residues that control coupled cleavage

While searching for features that might explain the intrinsic functional differences between the catalytic cores of RAG and BbRAGL, we observed that Ser963—which flanks the RAG1 catalytic glutamate Glu962—is positioned to form a hydrogen bond with Glu649 in apo RAG (Fig. 4a) and RAG bound to intact RSSs (Fig. 4b), but not when RAG is bound to nicked RSSs and poised for hairpin formation (Fig. 4c). *Bb*RAG1L cannot form this hydrogen bond because Glu649 and Ser963 have been replaced by Val751 and Ala1064 (Fig. 4d). Whether bound to intact or nicked TIRs, *Bb*RAG1L adopts a structure similar to that of RAG1 bound to nicked DNA (Fig. 4e, f, Extended Data Fig. 6f) and therefore appears to be constitutively poised for hairpin formation. Notably, the Glu649–Ser963 amino acid pair, which is strictly conserved in jawed vertebrate RAG1, is absent from known invertebrate RAG1-like proteins (Fig. 4g).

Incorporating residues of *Bb*RAG1L into RAG1 revealed that the RAG mutants RAG1(E649V), RAG1(S963A) and RAG1(E649V/ S963A) showed increased uncoupled cleavage activity when compared to the wild type (Fig. 4h, Extended Data Fig. 6g). By contrast, the RAG1(Y994F) mutation had no effect and the RAG1(N961A) mutant showed lower uncoupled cleavage than wild-type RAG1 (Fig. 4h, Extended Data Fig. 6g, h). Reciprocal mutations in BbRAG1L revealed that the BbRAG1L(V751E) mutant—but not the BbRAG1L(A1064S) mutant-showed decreased uncoupled cleavage, whereas the BbRAG1L(V751E/A1064S) double mutation almost abolished cleavage (Extended Data Fig. 6i). We propose that Glu649 in RAG1 helps to dictate coupled cleavage by mechanisms that are partially dependent on formation of a hydrogen bond with Ser963, and that because BbRAG1L lacks Glu649, it is more likely than RAG to adopt an active site configuration that is 'hairpin-competent'. Notably, RAG1(E649A) has previously been shown to exhibit increased uncoupled cleavage activity in vitro and in vivo<sup>24</sup>.

#### Two-tiered control of RAG transposition

We reasoned that structural comparisons of RAG and *Bb*RAGL might shed light on their markedly different capacities to perform transposition in cells. In the RAG post-cleavage complex, RAG1 Arg848 is near the RSS 3'-OH that attacks target DNA during transposition (Fig. 5a). Arg848 is strictly conserved in jawed vertebrate RAG1 but is replaced by methionine in *Bb*RAG1L and other invertebrate RAG1like proteins (Fig. 5b, c). RAG1(R848M) cleaves DNA at wild-type levels and exhibits a notable (approximately eightfold) increase in transposition activity in vitro relative to the wild type; this is manifest as efficient generation of a slow-mobility band that represents



Fig. 4 | Residues that control coupled cleavage. a-c, Structure of region surrounding RAG1 Glu649 and Ser963 before RSS binding (PDB 4WWX; a), bound to intact RSS (PDB 6CIK; b) and bound to two nicked RSSs with base flipping (PDB 5ZE1; c). In c, the potential for a hydrogen bond between Glu649 and Ser963 is disrupted by a change in the relative orientation of the residues and acquisition of a potassium ion (purple sphere). d, Structure of region surrounding BbRAG1L Val751 and Ala1064 bound to nicked TIR. e, f, Superimposition of protein structural elements containing RAG1 Glu649 and Ser963 (e) or BbRAG1L Val751 and Ala1064 (f) bound to intact or nicked DNA. Glu649, Ser963, Val751 and Ala1064 are highlighted with dark colours. In e, the intact DNA structure was obtained with a RAG1(E962Q) mutant<sup>22</sup>. g, Sequence alignments of RAG1 and RAG1-like proteins in the vicinity of RAG1 Glu649 and Ser963. Species name abbreviations are defined in the legend of Extended Data Fig. 5b. h, Cleavage reactions using cRAG with RAG1 mutations and DNA substrates containing one or two RSSs, as indicated above the lanes. Asterisks as in Fig. 2g.

inversion-circle intramolecular transposition products<sup>18</sup> (Fig. 5d, Extended Data Fig. 7a, b) and enhanced transposition of an RSSflanked antibiotic-resistance gene into a target plasmid (Fig. 5e, Extended Data Fig. 7c). Therefore, methionine at RAG1 position 848 stimulates RAG-mediated transposition at a post-cleavage step. Several amino acids at position 848 can support cleavage, with alanine stimulating and glutamate suppressing transposition relative to the wild type (Extended Data Fig. 7d, e).

Notably, in an in vivo plasmid-to-plasmid transposition assay (Extended Data Fig. 8a), the R848M mutation in RAG1 increased activity to detectable levels (Fig. 5f) whereas a reciprocal M949R mutation in *Bb*RAG1L decreased activity relative to the wild type (Extended Data Fig. 8b, c). RAG1(R848M) was, however, still about 100-fold less active than *Bb*RAGL (Fig. 5f), which raises the possibility that additional mechanisms suppress RAG-mediated transposition in vivo.

The RAG2 protein used in the assays of Fig. 5f (aa 1–383) contains part of the RAG2 acidic hinge (Fig. 1c), a domain that is present in jawed vertebrate RAG2 but absent from *Bb*RAG2L<sup>10</sup> and other known invertebrate RAG2-like proteins<sup>11</sup>. Notably, complete removal of the acidic hinge (RAG2 aa 1–350) increased in vivo transposition activity about 100-fold (Fig. 5g); this result was reproduced in a second cell line and with human RAG proteins (Extended Data Fig. 8d, e). Stimulation of transposition depended strongly on the RAG1(R848M) mutation, as wild-type RAG1 lacked detectable transposition activity when paired with RAG2(1–350) (Fig. 5g). Together, RAG1 Arg848 and the RAG2 acidic hinge suppress RAG-mediated transposition in vivo more than 1,000-fold. Transposition products generated in vitro and in vivo showed predominantly 5-bp target site duplications, as expected<sup>18,19</sup> (Extended Data Fig. 7f). Deletion of the RAG2 acidic hinge did not increase RAG-mediated transposition or DNA cleavage in vitro (Extended Data Fig. 7g, h) or substantially alter protein expression or V(D)J recombination activity in vivo (Extended Data Fig. 8f, g). Therefore, the RAG2 acidic hinge suppresses transposition specifically at a post-cleavage step and only in cells. Mapping experiments revealed that aa 362–383 have a critical role in suppressing in vivo transposition by RAG2(1–383) (Extended Data Fig. 8h, i). In assays using RAG2(1–350), a RAG1(E649V) mutation boosted transposition whereas the S963A mutation had little effect (Fig. 5g). We conclude that evolutionary adaptations arose early during jawed vertebrate evolution in RAG1 and RAG2 to provide two-tiered protection against RAGmediated transposition.

To test whether this conclusion extends to RAG-mediated transposition into the genome, we used a plasmid-to-genome transposition assay, with transposition target sites identified by high-throughput sequencing (Extended Data Fig. 9a-c). When paired with RAG2(1-350), RAG1(E649V/R848M), wild-type RAG1 and no RAG1 yielded 930, 16 and zero independent transposition events, respectively (Fig. 5h, Extended Data Fig. 9d). Insertion sites were found on all chromosomes (Extended Data Fig. 9e) and were strongly biased to active genes, particularly in the vicinity of the transcription start site (Extended Data Fig. 9f-h). Of the 930 RAG1(E649V/R848M)-mediated insertions, 180 (19%) occurred in protein-coding exons ( $P = 4 \times 10^{-82}$ ), which is noteworthy given that the primordial split antigen receptor gene of jawed vertebrates is believed to have been generated by insertion of the RAG transposon into an exon<sup>7,25,26</sup>. These data show that reversal of the protective adaptations acquired by jawed vertebrate RAG1 and RAG2 'reawakens' the RAG transposase and enables widespread transposition into genes and exons in the human genome.

#### Molecular domestication of the RAG transposon

The evolutionary adaptations that protect jawed-vertebrate lymphocytes from insertional mutagenesis caused by RAG-mediated transposition have been a long-standing target of investigation<sup>27,28</sup> and, a priori, could have involved changes in the RAG proteins, changes in the host cellular milieu or both. Efficient RSS ligation was unlikely to suffice as a protective mechanism because signal joints can be re-cleaved and transposed by RAG<sup>29</sup>. Our findings reveal two critical adaptations, intrinsic to the RAG proteins and found only in jawed vertebrates, that each potently suppress RAG-mediated transposition in vivo and together render the reaction almost undetectable. Similar to RAG1 Arg848, the RAG2 acidic hinge suppresses transposition at a post-cleavage step of the reaction, but-unlike Arg848-its suppressive effects are detectable only in the context of living cells. The RAG2 acidic hinge has previously been implicated in the regulation of RAG catalytic activity<sup>30</sup>, chromatin targeting<sup>31</sup>, repair pathway choice<sup>32,33</sup> and stability of the RAG-signal end complex<sup>33</sup>. It remains to be determined whether these activities are relevant to the suppression of RAGmediated transposition in vivo and whether other proteins contribute to this suppression.

Accumulating evidence supports a model for RAG evolution (Extended Data Fig. 10) in which a Transib transposon<sup>34</sup> captured a RAG2-like open reading frame in an early deuterostome to give rise to the original RAG transposon, which in turn gave rise to RAG1, RAG2 and RSSs in jawed vertebrates and RAG1L and RAG2L transposable elements and gene pairs in invertebrates<sup>9</sup>. We propose that the modular design of the RAG complex-with largely autonomous catalytic cores, swappable DNA binding modules and a RAG2 accessory subunit-facilitated the adaptation of RAG family enzymes to changing host environments and functional demands, including the adaptations in jawed vertebrates that led to a 'tamed' RAG recombinase that possesses coupled cleavage activity, adherence to the 12/23 rule and suppressed transposition activity (Extended Data Fig. 10). Our findings contribute to the paradigm of transposon molecular domestication<sup>4,6</sup>, which is now recognized to encompass elements in almost all branches of life ranging from CRISPR in bacteria<sup>35</sup> to active transposases encoded in the human genome, the function and process of domestication of which remain unknown<sup>36,37</sup>.



**Fig. 5** | **Reawakening the RAG transposon in vivo. a**, Structure of region surrounding RAG1 Arg848 after hairpin formation (PDB 5ZE2). **b**, Structure of region surrounding RAG1 Arg848 (PDB 5ZDZ) or *Bb*RAG1L Met949 after nicking. **c**, Sequence alignments of RAG1 and RAG1-like proteins in the vicinity of RAG1 Arg848. Red-shaded residues, highly conserved binding surface for adenine base of heptamer adjacent to flipped C +1. d, Cleavage reactions comparing intramolecular transposition by wild-type RAG1 and RAG1(R848M). The intramolecular transposition product was confirmed to contain inversion circles by inverse PCR DNA sequencing<sup>18</sup>. **e**, Results of in vitro transposition

#### **Online content**

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-019-1093-7.

Received: 10 August 2018; Accepted: 7 March 2019; Published online 10 April 2019.

- 1. Gellert, M. V(D)J recombination: RAG proteins, repair factors, and regulation. Annu. Rev. Biochem. **71**, 101–132 (2002).
- Schatz, D. G. & Swanson, P. C. V. V(D) Jrecombination: mechanisms of initiation. Annu. Rev. Genet. 45, 167–202 (2011).
- Lewis, S. M. The mechanism of V(D)J joining: lessons from molecular, immunological, and comparative analyses. *Adv. Immunol.* 56, 27–150 (1994).
- Sinzelle, L., Izsvák, Z. & Ivics, Z. Molecular domestication of transposable elements: from detrimental parasites to useful host genes. *Cell. Mol. Life Sci.* 66, 1073–1093 (2009).
- Levin, H. L. & Moran, J. V. Dynamic interactions between transposable elements and their hosts. *Nat. Rev. Genet.* 12, 615–627 (2011).
- Jangam, D., Feschotte, C. & Betrán, E. Transposable element domestication as an adaptation to evolutionary conflicts. *Trends Genet.* 33, 817–831 (2017).
- Thompson, C. B. New insights into V(D)J recombination and its role in the evolution of the immune system. *Immunity* 3, 531–539 (1995).
- Fugmann, S. D. The origins of the Rag genes—from transposition to V(D)J recombination. Semin. Immunol. 22, 10–16 (2010).
- Carmona, L. M. & Schatz, D. G. New insights into the evolutionary origins of the recombination-activating gene proteins and V(D)J recombination. *FEBS J.* 284, 1590–1605 (2017).
- Huang, S. et al. Discovery of an active RAG transposon illuminates the origins of V(D)J recombination. Cell 166, 102–114 (2016).
- 11. Morales Poole, J. R., Huang, S. F., Xu, A., Bayet, J. & Pontarotti, P. The RAG transposon is active through the deuterostome evolution and domesticated in jawed vertebrates. *Immunogenetics* **69**, 391–400 (2017).
- Chatterji, M., Tsai, C. L. & Schatz, D. G. Mobilization of RAG-generated signal ends by transposition and insertion *in vivo. Mol. Cell. Biol.* 26, 1558–1568 (2006).
- Reddy, Y. V., Perkins, E. J. & Ramsden, D. A. Genomic instability due to V(D)J recombination-associated transposition. *Genes Dev.* 20, 1575–1582 (2006)
- Curry, J. D. et al. Chromosomal reinsertion of broken RSS ends during T cell development. J. Exp. Med. 204, 2293–2303 (2007).
- Messier, T. L., O'Neill, J. P., Hou, S. M., Nicklas, J. A. & Finette, B. A. *In vivo* transposition mediated by V(D)J recombinase in human T lymphocytes. *EMBO* J. 22, 1381–1388 (2003).
- Little, A. J., Matthews, A. G., Oettinger, M. A., Roth, D. B. & Schatz, D. G. in Molecular Biology of B Cells (eds Alt, F. W. et al.) Ch. 2, 13–34 (Academic/Elsevier, 2015).

reactions with wild-type RAG1 or RAG1(R848M) (mean  $\pm$  s.e.m.). Two-tailed *t*-test: \*\**P* < 0.01. **f**, **g**, Results of in vivo plasmid-to-plasmid transposition assays with RAG2(1–383) (**f**) or RAG2(1–350) (**g**) and the indicated full-length wild-type or mutant RAG1 proteins, and with full-length *Bb*RAGL (mean  $\pm$  s.e.m.). Total antibiotic-resistant colony numbers (grey bars) were corrected (black bars) for the fraction of colonies found to contain plasmids with bone fide transposition events. Two-tailed *t*-test: \**P* < 0.05; \*\**P* < 0.01; \*\*\**P* < 0.005 compared to wild-type RAG1. **h**, Number of bone fide transposition events (3–7-bp target site duplications) identified in plasmid-to-genome transposition experiment.

- Yin, F. F. et al. Structure of the RAG1 nonamer binding domain with DNA reveals a dimer that mediates DNA synapsis. *Nat. Struct. Mol. Biol.* 16, 499–508 (2009).
- Agrawal, A., Eastman, Q. M. & Schatz, D. G. Transposition mediated by RAG1 and RAG2 and its implications for the evolution of the immune system. *Nature* 394, 744–751 (1998).
- Hiom, K., Melek, M. & Gellert, M. DNA transposition by the RAG1 and RAG2 proteins: a possible source of oncogenic translocations. *Cell* 94, 463–470 (1998).
- Kim, M. S., Lapkouski, M., Yang, W. & Gellert, M. Crystal structure of the V(D)J recombinase RAG1–RAG2. Nature 518, 507–511 (2015).
- 21. Ru, H. et al. Molecular mechanism of V(D)J recombination from synaptic RAG1-RAG2 complex structures. *Cell* **163**, 1138–1152 (2015).
- Kim, M. S. et al. Cracking the DNA code for V(D)J recombination. *Mol. Cell* 70, 358–370.e354, (2018).
- Ru, H. et al. DNA melting initiates the RAG catalytic pathway. Nat. Struct. Mol. Biol. 25, 732–742 (2018).
- Kriatchko, A. N., Anderson, D. K. & Swanson, P. C. Identification and characterization of a gain-of-function RAG-1 mutant. *Mol. Cell. Biol.* 26, 4712–4728 (2006).
- Sakano, H., Hüppi, K., Heinrich, G. & Tonegawa, S. Sequences at the somatic recombination sites of immunoglobulin light-chain genes. *Nature* 280, 288–294 (1979).
- Hsu, E. & Lewis, S. M. in *Molecular Biology of B Cells* (eds Alt, F. W. et al.) Ch. 9, 133–149 (Academic, 2015).
- Jones, J. M. & Gellert, M. The taming of a transposon: V(D)J recombination and the immune system. *Immunol. Rev.* 200, 233–248 (2004).
- Chatterji, M., Tsai, C. L. & Schatz, D. G. New concepts in the regulation of an ancient reaction: transposition by RAG1/RAG2. *Immunol. Rev.* 200, 261–271 (2004).
- Neiditch, M. B., Lee, G. S., Huye, L. E., Brandt, V. L. & Roth, D. B. The V(D)J recombinase efficiently cleaves and transposes signal joints. *Mol. Cell* 9, 871–878 (2002).
- Lu, C., Ward, A., Bettridge, J., Liu, Y. & Desiderio, S. An autoregulatory mechanism imposes allosteric control on the V(D)J recombinase by histone H3 methylation. *Cell Rep.* **10**, 29–38 (2015).
- Ward, A., Kumari, G., Sen, R. & Desiderio, S. The RAG-2 inhibitory domain gates accessibility of the V(D)J recombinase to chromatin. *Mol. Cell. Biol.* 38, e00159 (2018).
- 32. Corneo, B. et al. Rag mutations reveal robust alternative end joining. *Nature* **449**, 483–486 (2007).
- Coussens, M. A. et al. RAG2's acidic hinge restricts repair-pathway choice and promotes genomic stability. *Cell Rep.* 4, 870–878 (2013).
- Kapitonov, V. V. & Jurka, J. RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons. *PLoS Biol.* 3, e181 (2005).
- Krupovic, M., Béguin, P. & Koonin, E. V. Casposons: mobile genetic elements that gave rise to the CRISPR-Cas adaptation machinery. *Curr. Opin. Microbiol.* 38, 36–43 (2017).

- Majumdar, S., Singh, A. & Rio, D. C. The human THAP9 gene encodes an active P-element DNA transposase. Science 339, 446–448 (2013).
- 37. Henssen, A. G. et al. Genomic DNA transposition induced by human PGBD5. *eLife* **4**, e10565 (2015).
- 38. Lefranc, M. P. & Lefranc, G. The T Cell Receptor FactsBook (Academic, 2001).

Acknowledgements We thank Y. Kong for computational analysis of genome transposition data, W. Eliason for assistance with size-exclusion chromatography-multiple angle light scattering, C. Akey for help with freezing of the grids for *Bb*RAGL-intact 3/TIR, R. Huang and H. Chuan for help with cryo-EM data collection, J. Wang for advice and insight in structural analysis, E. Hendrickson for HCT116 cells, XEGEN for NBD phylogenetic sequence analysis, Z. Chou, X. Liu and H. Zhang for an insertion-site mapping script, M. Ciubotaru for advice and members of the Schatz laboratory for comments on the manuscript. This work was supported in part by R01 Al32524 and R01 Al137079 (D.G.S.), R01 Al102778 (Y.X.), 2013CB917800 from the Ministry of Science and Technology of China and 91231206 from the National Natural Science Foundation of China (A.X.), UEFISCDI grant PN-III-ID-PCE-2016-0650 and Romanian Academy programs 1 & 3 of IBAR (M.D.S. and A.J.P.) and grants from the Centre National de la Recherche Scientifique and Aix-Marseille Université (P.P.).

**Reviewer information** *Nature* thanks Frederick Alt, Thomas Boehm and the other anonymous reviewer(s) for their contribution to the peer review of this work.

**Author contributions** Y.Z. and D.G.S. designed the experiments. Y.Z. purified the proteins and performed the biochemical and cell-based experiments. T.C.C. performed freezing of the cryo-EM grids, data collection and processing, and model building with input from Y.X. Y.Z. and T.C.C. performed structural analyses. M.D.S. and A.J.P. created the computational model of *Bb*RAGL. G.H. and Q.L. helped to establish the in vivo transposition assays, and Q.L. performed the human RAG transposition assays. J.D.M. performed computational analysis of genome transposition data. A.X. provided the BbRAG1L and BbRAG2L codon-optimized cDNAs and information about *Bb*RAGL function. P.P. performed phylogenetic analyses of *Bb*RAG1L sequences. D.G.S. wrote the paper with input from other authors.

Competing interests The authors declare no competing interests.

#### Additional information

**Extended data** is available for this paper at https://doi.org/10.1038/s41586-019-1093-7.

**Supplementary information** is available for this paper at https://doi.org/ 10.1038/s41586-019-1093-7.

**Reprints and permissions information** is available at http://www.nature.com/ reprints.

**Correspondence and requests for materials** should be addressed to A.X. or Y.X. or D.G.S.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

#### **METHODS**

Statistical methods were not used to predetermine sample size and experiments were not randomized. Investigators were not blinded to allocation during experiments and outcome assessment.

**Plasmid generation**. pTT5M, a derivative of pTT5 containing a maltose-binding protein (MBP) open reading frame (ORF), previously described<sup>10</sup>, was modified by inserting an in frame PreScission Protease cleavage site at the C terminus of MBP, to create pTT5MP. Codon-optimized *Bb*RAG1L core (aa 468–1136) and full-length *Bb*RAG2L were cloned into pTT5MP at a NotI restriction site that lies downstream of the protease cleavage site by In-Fusion cloning. Truncated *Bb*RAG1L ORFs (aa 468–1136, aa 484–1136 and ΔNBD (aa 547–1136)) were cloned into pTT5M, as were mouse RAG1 core (aa 384–1008), RAG1 core ΔNBD (aa 462–1008) and RAG2 core (aa 1–383) ORFs. Chimaera protein ORFs and point mutants thereof were cloned into pTT5MP. No difference in expression levels was noted between pTT5M and pTT5MP vectors.

A 5'TIR and a 3'TIR, each with 3' flanking *ProtoRAG* sequences, were inserted together into the BamHI site of pUC19 by In-Fusion, creating a substrate with 402 bp between the tips of the TIRs. This vector was further modified to eliminate all instances of 5'-CAC in the DNA between the TIRs and in the ~130 bp of pUC19 flanking the 5'TIR and ~280 bp of pUC19 flanking the 3'TIR. This CAC-free region containing the TIRs was then subcloned into the EcoRV/NruI sites of pBR322 to create pB-5'/3'TIR. pB-5'/3'TIR was modified by deletion of the 5'TIR or the 3'TIR using PCR and In-Fusion cloning to create pB-5'/3'TIR, pB-3'TIR. Other alterations to replace or modify the TIRs of pB-5'/3'TIR, pB-5'TIR or pB-3'TIR, using In-Fusion cloning, resulted in plasmids containing the needed combinations of RSS, chimeric TIR/RSS and scrambled TIR mutant sequences. The mutations that scrambled portions of the TIR were made by changing A to C, T to G, C to A and G to T.

Protein expression and purification. pTT5MP-BbRAG1L core and pTT-5MP-BbRAG2L plasmids were cotransfected into expi293F cells using the ExpiFectamine 293 Transfection Kit. Cells (30-200 ml culture) containing co-expressed proteins were collected 5 days after transfection by centrifugation (500g) and frozen at -80 °C. Cells were re-suspended in lysis buffer (25 mM Tris, pH7.5, 1M KCl, 1 mM DTT) and disrupted by three cycles of freezing in liquid nitrogen and thawing in a room temperature water bath. Cell lysates were further disrupted by dounce homogenization and centrifuged at 45,000 r.p.m. (Beckman Coulter Optima LE-80K Ultracentrifuge, Type 50.2 Ti rotor) for 1 h at 4°C (all subsequent steps at 4 °C), and the supernatant was mixed with pre-equilibrated amylose resin and incubated for 2 h with continual rotation. The beads were loaded onto a gravity flow column and washed with 50-100 ml of lysis buffer and protein eluted with 5-10 ml of elution buffer (25 mM Tris, pH7.5, 0.5M KCl, 1 mM DTT, 40 mM maltose) depending on the initial cell culture volume. The eluate was further purified by size-exclusion chromatography (SEC) on a Superdex 200 Increase 10/300 GL column in 20 mM HEPES pH7.6, 0.5 mM TCEP, 150 mM KCl and 5 mM MgCl<sub>2</sub> or 5 mM Ca(OAc)<sub>2</sub> (Ca<sup>2+</sup> buffer used when protein was purified for assembling protein-DNA complexes for cryo-EM because Ca<sup>2+</sup> supports DNA binding but not cleavage by RAG<sup>1</sup> and BbRAGL<sup>10</sup>). SEC peak fractions were collected and pooled and protein concentrated to 4–10  $\mu M$  using an Amicon centrifugal concentrator and frozen at -80 °C. Other proteins were expressed and purified following a similar procedure. In all cases, the RAG1 core, BbRAG1L, or chimeric protein was co-expressed with the appropriate RAG2 core or BbRAG2L protein.

Full length (FL) His<sub>6</sub>-hHMGB1 and His<sub>6</sub>-hHMGB1 $\Delta$ C (aa 1–165 lacking the acidic C-terminal region) were expressed and purified as previously described<sup>10,39</sup>.

HEK293T cells were obtained from ATCC, expi293F cells were obtained from Thermo Fisher Scientific, and HCT116 cells were obtained from E. Hendrickson, University of Minnesota. Cell lines used were not authenticated or tested for mycoplasma contamination.

In vitro DNA cleavage. Wild-type *Bb*RAGL or RAG proteins (25 nM final concentration of each protein), substrate DNA (final concentration 10 nM) and 175 ng His<sub>6</sub>–hHMGB1 were incubated in reaction buffer (25 mM HOPS, pH7.0, 50 mM KCl, 2 mM DTT, 1.5 mM MgCl<sub>2</sub>; 16  $\mu$ l final reaction volume) at 37 °C for

1 h or for the indicated time period. For reactions with chimeric proteins containing the RAG1 catalytic core, the final concentration of each protein was 50 nM. For reactions with chimeric proteins containing the *Bb*RAG1L catalytic core, the final concentration of each protein was 50 nM, the Mg<sup>2+</sup> concentration was 5 mM and reaction time was 2 h. In these experiments, control reactions for each experiment used the same conditions as the reactions with the chimeric proteins. Reactions were stopped by adding 1.25  $\mu$ l 2.5% SDS, 5  $\mu$ l proteinase K (150  $\mu$ g/ml) and 2  $\mu$ l 0.5 M EDTA followed by incubation at 55 °C for at least 3 h. Samples were briefly centrifuged and the supernatant mixed with 1.7  $\mu$ l 80% glycerol and loaded on a non-denaturing 1 × TBE (Tris-borate-EDTA buffer) 6% polyacrylamide gel. After 1 h electrophoresis at 100 V, gels were stained with SYBR GOLD in 1 × TBE buffer for 20 min and imaged using a PharosFX Plus (Bio-Rad).

**Confirmation of intramolecular transposition band.** The slow mobility band (as in Fig. 5d, arrow) was excised and DNA purified and subject to inverse PCR using primers F and R. The major PCR product band (Extended Data Fig. 7a, arrow) was excised and the DNA purified, cloned and sequenced. Inversion circle transposition products were identified as previously described<sup>18</sup>. F: TATTATGAGGCCCTTTCGTCTTC; R: CGCCTATTTTTATAGGTTAATGTCATG.

**BbRAGL-3'TIR synaptic complex assembly for cryo-EM.** Purified MBP-BbRAGL complex was mixed with 3'TIR DNA substrate and His<sub>6</sub>-hHMGB1 $\Delta$ C at a ratio of 1:2.5:2.5 in 20 mM HEPES pH7.6, 0.5 mM TCEP, 10 mM CaCl<sub>2</sub> and 150 mM KCl, and incubated at room temperature for 30 min. After incubation, 5% (v/v) PreScission Protease was added and the sample was incubated at room temperature for 1 h to cleave off the MBP tags. The mixture was loaded on a Superdex 200 Increase 10/300 GL column and purified by SEC in 20 mM HEPES pH7.6, 0.5 mM TCEP, 150 mM KCl and 5 mM Ca(OAc)<sub>2</sub>. The peak column fractions were collected and concentrated (if necessary) to a protein concentration of ~0.4 mg/ml. The sample was immediately used to prepare cryo-EM grids.

**Cryo-EM data acquisition.** Purified complex (3 µl) was applied to freshly glow-discharged C-flat 400 mesh, R2/1 and R1.2/1.3 holey grids for intact DNA and nicked DNA complexes, respectively. Grids were blotted for 4 s in 100% humidity and plunge frozen in liquid ethane using a Vitrobot Mark 3 (FEI). A Titan Krios electron microscope (Janelia Research Campus, HHMI) operated at 300 kV, with a spherical aberration corrector and a Gatan Image Filter (slit width of ~20 eV), was used to acquire images with a K2 Summit direct electron detector (Gatan) in super-resolution mode. The image stacks were collected at a nominal magnification of 81,000×, corresponding to 0.675 Å per super-resolution pixel, at a dose rate of ~10.2 electrons per physical pixel per second. The total exposure was 80 and 54 electrons per Å<sup>2</sup>, fractionated into 50 and 40 frames, for intact DNA complex and nicked DNA complex, respectively. All images were acquired in a defocus range from –1.0 to –2.5 µm. The statistics of data acquisition are summarized in Extended Data Table 1.

Image processing. A total of 5,164 and 4,429 LZW-compressed TIFF image stacks were collected for intact DNA complex and nicked DNA complex, respectively. MotionCor2 1.140 was used for beam-induced motion correction and dose weighting. The first two frames were discarded, and the output aligned images were binned  $2 \times$  in Fourier space, resulting in a pixel size of 1.35 Å for further processing. The non-dose weighted aligned images were used for contrast transfer function estimation by Gctf 1.06<sup>41</sup>. The dose-weighted images were used for autopicking, classification and reconstruction. Roughly 3,000 particles were manually picked, followed by a round of 2D classification to generate templates for RELION 1.4 autopicking. The autopicked particles were subjected to 2D classification in RELION-2.142,43 to remove junk particles. Particle coordinates in good classes were extracted for further manual inspection such that bad particles and images were discarded. A previously published cryo-EM map (Electron Microscopy Data Bank code EMD-6488)<sup>21</sup> was low-pass filtered to 60 Å to serve as a starting reference for multiple rounds of 3D classification in RELION-2.1 without imposing symmetry. Good 3D classes were combined and used for gold standard refinement in RELION-2.1 with either C1 or C2 symmetry. Resolution estimation was based on the Fourier shell correlation cutoff at 0.143 (FSC<sub>0.143</sub>) between the two half-maps, after a soft mask was applied to mask out solvent region (Extended Data Table 1). The final maps were corrected for K2 detector modulation and sharpened by their corresponding negative B-factors within RELION-2.1. Local resolution variation was estimated by the local resolution module in RELION-2.1.

**Modelling and refinement.** An initial model was obtained by structural profiling of *cBb*RAG1L and *Bb*RAG2L sequence propensities as previously described<sup>44–46</sup>. In brief, separate models of *cBb*RAG1L (aa 473–1110) and *Bb*RAG2L were built, accounting for accessibility, charge, hydropathy, consensus secondary structure, consensus intrinsic disorder profiles and fold recognition assessment using Discovery Studio software suite 3.0 (Accelrys). The models were refined by remote homology techniques starting from mouse RAG1 structural templates in PDB 3GNA<sup>17</sup> for NBD\* of *Bb*RAG1L and PDB 4WWX<sup>20</sup> for *cBb*RAG1L and *Bb*RAG2L. To eliminate steric conflicts and further minimize energy, these models were iter-

atively refined until convergence by repeated cycles of Generalized Born simulated annealing molecular dynamics for implicit solvent using NAMD 2.12 with CHARMM36 force field<sup>47</sup> followed by model assessment of the global distance test total score (GDT\_TS) with QA-RecombineIT<sup>48</sup> and local loop remodelling in regions showing the highest divergence. Annealing simulations were performed with harmonic restraints on the backbone protein atom positions in regions of regular secondary structure, while irregular loop regions were left to move freely. This brought the *cBb*RAG1L and *Bb*RAG2L models to GDT\_TS 60 and >67, and root mean square deviations of 2.9 Å and 2.3 Å, respectively. Finally, the assembled *cBb*RAG1L or *Bb*RAG2L structure was subjected to molecular dynamics simulation in explicit solvent to confirm robustness and stability and to assess configuration dynamics of *cBb*RAG1L and *Bb*RAG2L domains relative to one another.

The BbRAG1L (aa 545-1104) and BbRAG2L (aa 1-366) model thus derived was flexibly fitted into the C2 symmetrized map of the nicked 3'TIR complex (4.3 Å) by molecular dynamics flexible fitting<sup>49</sup>. The flexibly fitted model was able to account for most density, except for BbRAG1L loops 640-650, 704-720, 732-740 and 1046-1053, and BbRAG2L loops 11-22, 34-49, 67-74, 85-108, 179-190 and 300-314, which were adjusted and rebuilt in COOT 0.8<sup>50</sup>. The density for loop 603-630 in BbRAG1L was insufficient for model building. An all-alanine chain was built to fit the density for the C-terminal tail (CTT\*) of BbRAG1L (1105-1125). The DNA chains from the previously published model (PDB 3JBY)<sup>21</sup> were fit into the map and then changed to the correct DNA sequence in COOT 0.8. The model was adjusted in COOT 0.8 manually with iterative cycles of automatic rebuilding using the Rosetta FastRelax protocol<sup>51</sup>. The model was further refined using the phenix.real\_space\_refine module in PHENIX with secondary structure restraints and Ramachandran restraints<sup>52</sup>. The final model was validated using MolProbity<sup>53</sup> and EMringer<sup>54</sup> (Extended Data Table 1). All molecular representations were generated in PyMol (https://www.pymol.org) and Chimera<sup>55</sup>

In vitro transposition assay. The in vitro intermolecular transposition reaction (Extended Data Fig. 7b) was performed as described in 'In vitro DNA cleavage'. The 12/23RSS substrate was replaced by 10 nM linear donor fragment with tetracycline-resistant marker and 10 nM pECFP-1 target plasmid. The final concentrations of RAG protein,  $Mg^{2+}$  and DTT were 50 nM, 1.5 mM and 2 mM, respectively. After proteinase K digestion, DNA was ethanol-precipitated. DNA (50 ng) was transformed into electrocompetent MC1061 bacterial cells that were spread onto plates containing kanamycin or tetracycline-kanamycin-streptomycin. Transposition efficiency was calculated by dividing the number of colonies obtained on the kanamycin-alone plate<sup>10,18</sup>.

In vivo plasmid-to-plasmid transposition assay. The in vivo plasmid-to-plasmid transposition assay (Extended Data Fig. 8a) was performed as previously described<sup>12</sup>. In brief, 293T cells were transfected with 4  $\mu$ g each of the pEBB-RAG1 or mutant and pTT5M-RAG2 truncations or pEBB-FL RAG2, 6  $\mu$ g donor plasmid (pTetRSS), and 10  $\mu$ g target plasmid (pECFP-1) using polyethyl-enimine. The medium was changed 24 h after transfection and cells were collected after 48 h. Plasmid DNA was precipitated and 300 ng DNA was transformed into electrocompetent MC1061 bacterial cells, which were plated onto kanamycin or kanamycin-tetracycline-streptomycin (KTS) plates. For each protein combination assayed in Fig. 5f, g, plasmids from 30 colonies (except for very low-efficiency reactions) from KTS plates were sequenced to determine whether they contained a bone fide transposition event (3–7 bp target site duplication (TSD)). Total transposition efficiency was calculated from the fraction of plasmids that contained a transposition event.

Western blotting. Cells were collected 48 h after transfection of protein expression vectors, resuspended in lysis buffer (50 mM Tris, pH 7.5; 150 mM NaCl; 1% N-P40; cocktail protease inhibitor) on ice, and further disrupted by sonication. After centrifugation to remove insoluble debris, samples were mixed with loading buffer, subjected to sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS–PAGE) and transferred to a PVDF membrane, then incubated separately with anti-RAG1<sup>56</sup>, anti-RAG2<sup>56</sup> and mouse monoclonal anti- $\beta$ -actin (Sigma #A1978) antibodies.

In vivo recombination assay. RAG or *Bb*RAGL expression vectors (1 µg) were co-transfected with 2 µg of pCJGFP<sup>32</sup> or pTIRG8<sup>10</sup>, respectively, into expi293F cells using polyethylenimine (DNA:PEI ratio of 1:3). Cells were collected 72 h post-transfection, washed twice with PBS containing 1% FBS and stained with DAPI (4',6-diamidino-2-phenylindole), and the percentage of live cells expressing GFP was determined by flow cytometry as shown in Supplementary Data 2.

In vivo plasmid-to-genome transposition assay. The in vivo plasmid-to-genome transposition assay (Extended Data Fig. 9a) was performed by transfecting 293T cells with 4  $\mu$ g each of pEBB-RAG1, pEBB-RAG1(E649V/R848M) or empty vector and pTT5M-RAG2(1–350), and 6  $\mu$ g of donor pBSK-12puro23<sup>12</sup>. Forty-eight hours after transfection, 5 × 10<sup>6</sup> cells were split into medium containing 0.8  $\mu$ g/ml puromycin. After 2–3 weeks of culture, colonies (many hundreds from each

experiment; colony formation was dependent on inclusion of the donor plasmid) were digested with trypsin, pooled and re-seeded into new medium containing puromycin and cultured further to obtain sufficient cells. For each experiment, 10<sup>7</sup> cells were collected and the genomic DNA was precipitated. Transposition insertion targets from three independent experiments (no RAG1, wild-type RAG1 or RAG1(E649V/R848M)) were amplified using 12RSS and 23RSS linear amplification (LAM)-PCR primers with six different barcodes (12v, 23v, 12wt, 23wt, 12m, 23m) as previously described<sup>57</sup>. Equal amounts of LAM-PCR product from the six groups were mixed and diluted as the library for high-throughput sequencing.

LAM-PCR primers: 12RSS, 5'-biotin-ctttattgaggcttaagcagtgggtte; 23RSS, 5'-biotin-actgacactcgacctgacaggattg; nested-PCR primers: 12RSS, 5'-acactct-ttccctacacgacgctcttccgatctXXXXXgcaaaaagcagatcttatttcgtt; 23RSS: 5'-acactct-ttccctacacgacgctcttccgatctXXXXXcttatcatgtctggatcgctttatatacg, where XXXXX represents the barcode.

High-throughput sequencing and data analysis. High-throughput sequencing was performed on an Illumina NextSeq 500. Cutadapt version 1.16 was used to identify barcodes with the adaptor-matching error rate set to allow one mismatch in a 7-8-bp barcode. We successfully matched 95.9% of reads to a barcode and unmatched reads were discarded. Next, cutadapt was used to trim the barcode sequences, primer sequences, and 12RSS and 23RSS sequences. The resulting trimmed sequencing data contained only vector sequence or genomic DNA sequence from transposition events and other random integration events. Overall, 60% of barcoded sequencing reads contained identifiable RSS sequences and other reads were discarded. Trimmed sequences were aligned to human genome GRCh38 using Bowtie2 (version 2.2.9) using 'very sensitive' end-to-end alignment mode. High-quality alignments (MAPQ  $\geq$  30; identified with Samtools 1.5) were converted to bed intervals using the bedtools bamToBed utility (bedtools version 2.27.1). Overlapping same-stranded events were merged for each of the six libraries. Bone fide transposition events give rise to 12RSS- and 23RSS-flanking genomic sequences that mapped to the same site in the genome but in opposite directions and with short overlaps (the target site duplication), a signature that was readily distinguished from random integration of the donor plasmid or excised RSS fragments (Extended Data Fig. 9b). To accomplish this, Bedtools intersect was used to identify loci at which corresponding 12RSS and 23RSS libraries showed evidence of transposition events on opposite strands. All intersecting intervals with a 3-7-bp overlap were judged to be transposition events. Gene, exon and transcription start site (TSS) definitions were downloaded from Ensembl gene v93, dataset Human genes (GRCh38.p12). Active TSSs and active genes or exons were defined based on H3K4me3 (experiment ENCSR000DTU) and H3K36me3 (experiment ENCSR910LIE) chromatin immunoprecipitation with sequencing datasets, respectively, from HEK293 cells from ENCODE (https://www.encodeproject.org). Statistics and reproducibility. DNA cleavage experiments were typically performed three or more times. Exceptions are Figs. 3i, 4i, Extended Data Fig. 6a, c, f (lanes 10 and 11), all n = 2. Statistical analyses were performed using a two-tailed t-test (for example, Fig. 5e-g) or a one-tailed Fisher's exact test (Fig. 5i).

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

#### Data availability

The model of the *cBb*RAGL-nicked 3'TIR synaptic complex has been deposited in the PDB with accession code 6B40. The cryo-EM maps of *cBb*RAGL in complex with intact or nicked 3'TIRs have been deposited in Electron Microscopy DataBank with accession codes EMD-7043, -7044, -7045 and -7046. High-throughput DNA sequence data to identify transposition events in the human genome have been deposited in the NCBI Sequence Read Archive with accession codes SRR8430227– SRR8430233 (project PRJNA514369).

- Bergeron, S., Anderson, D. K. & Swanson, P. C. RAG and HMGB1 proteins: purification and biochemical analysis of recombination signal complexes. *Methods Enzymol.* 408, 511–528 (2006).
- Zheng, S. Q. et al. MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat. Methods* 14, 331–332 (2017).
- Żhang, K. Gctf: Real-time CTF determination and correction. J. Struct. Biol. 193, 1–12 (2016).
- Scheres, S. H. RELION: implementation of a Bayesian approach to cryo-EM structure determination. J. Struct. Biol. 180, 519–530 (2012).
- Kimanius, D., Forsberg, B. O., Scheres, S. H. & Lindahl, E. Accelerated cryo-EM structure determination with parallelisation using GPUs in RELION-2. *eLife* 5, e18722 (2016).
- 44. Slootweg, E. J. et al. Structural determinants at the interface of the ARC2 and leucine-rich repeat domains control the activation of the plant immune receptors Rx1 and Gpa2. *Plant Physiol.* **162**, 1510–1528 (2013).
- Zhang, Y. H., Shetty, K., Surleac, M. D., Petrescu, A. J. & Schatz, D. G. Mapping and quantitation of the interaction between the recombination activating gene proteins RAG1 and RAG2. J. Biol. Chem. 290, 11802–11817 (2015).



- Kozuki, T. et al. Roles of the C-terminal domains of topoisomerase IIα and topoisomerase IIβ in regulation of the decatenation checkpoint. *Nucleic Acids Res.* 45, 5995–6010 (2017).
- Phillips, J. C. et al. Scalable molecular dynamics with NAMD. J. Comput. Chem. 26, 1781–1802 (2005).
- Pawlowski, M., Bogdanowicz, A. & Bujnicki, J. M. QA-Recombinelt: a server for quality assessment and recombination of protein models. *Nucleic Acids Res.* 41, W389–W397 (2013).
- Trabuco, L. G., Villa, E., Mitra, K., Frank, J. & Schulten, K. Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics. *Structure* 16, 673–683 (2008).
- Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. Acta Crystallogr. D 66, 486–501 (2010).
- DiMaio, F. et al. Atomic-accuracy models from 4.5-Å cryo-electron microscopy data with density-guided iterative local refinement. *Nat. Methods* 12, 361–365 (2015).
  Adams, P. D. et al. PHENIX: a comprehensive Python-based system for
- macromolecular structure solution. Acta Crystallogr. D 66, 213–221 (2010).

- Chen, V. B. et al. MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr. D 66, 12–21 (2010).
- Barad, B. A. et al. EMRinger: side chain-directed model and map validation for 3D cryo-electron microscopy. *Nat. Methods* 12, 943–946 (2015).
- Pettersen, E. F. et al. UCSF Chimera—a visualization system for exploratory research and analysis. J. Comput. Chem. 25, 1605–1612 (2004).
- Coster, G., Gold, A., Chen, D., Schatz, D. G. & Goldberg, M. A dual interaction between the DNA damage response protein MDC1 and the RAG1 subunit of the V(D)J recombinase. J. Biol. Chem. 287, 36488–36498 (2012).
- Hu, J. et al. Detecting DNA double-stranded breaks in mammalian genomes by linear amplification-mediated high-throughput genome-wide translocation sequencing. *Nat. Protocols* **11**, 853–871 (2016).
- Lapkouski, M., Chuenchor, W., Kim, M. S., Gellert, M. & Yang, W. Assembly pathway and characterization of the RAG1/2-DNA paired and signal-end complexes. J. Biol. Chem. 290, 14618–14625 (2015).



**Extended Data Fig. 1** | *ProtoRAG* transposon and analysis of the *Bb*RAG1L NBD\* domain. a, Schematic of the *ProtoRAG* transposon and, below it, the jawed-vertebrate RAG locus and prototypical antigen receptor gene (*IGH*). b, Schematic of full-length and truncated *Bb*RAG1L proteins (top), and cleavage reactions performed with those proteins (plus *Bb*RAG2L) and TIR substrates, as indicated above and below the lanes. Core *Bb*RAG1L (aa 468–1136) retains the cleavage pattern of full-length *Bb*RAG1L, whereas full-length *Bb*RAG1L exhibits strong single

TIR cleavage (lanes 6, 7). Closed and open arrowheads, single 5'TIR and single 3'TIR cleavage products, respectively. For gel source data, see Supplementary Fig. 1. **c**, Sequence alignment of *Bb*RAG1L NBD\* with RAG1 NBD showing divergent sequences with similar predicted secondary structure elements ( $\alpha$ -helices 1, 2 and 3). **d**, SEC-multiple angle light scattering (SEC-MALS) analysis of the purified NBD\* protein, indicating that the protein is a dimer in solution.



Extended Data Fig. 2 | Biochemical properties and cryo-EM structure of cBbRAGL-3'TIR synaptic complexes. a, SEC-MALS of MBPcBbRAGL, which indicates that the complex is a heterotetramer with two subunits each of cBbRAG1L and BbRAG2L. b, c, SEC profiles of cBbRAGL incubated with intact (b) or nicked (c) 3'TIR, 5'TIR or 3'/5'TIRs showing resolution of protein-DNA complex from free DNA. Gels display the components of pooled column fractions containing the protein-DNA complex. d, Representative 2D class averages of cryo-EM particles of cBbRAGL bound to intact or nicked 3'TIRs. e, Left, FSC curves of the half maps from gold standard refinements of the cBbRAGL-nicked 3'TIR complex with no symmetry applied (blue) and the cBbRAGL-intact 3'TIR complex with no symmetry applied (red) and with C2 symmetry applied (green). Right, FSC curves of the gold standard refinement of cBbRAGL-nicked 3'TIR complex with C2 symmetry applied (blue) and of the C2 symmetrized map and model (green). Resolutions of the maps are read by the cutoff values at FSC = 0.143. f, Colour-coded local-resolution estimation of the C2 symmetrized map of cBbRAGL in

complex with nicked 3'TIR, viewed from a perspective similar (with a 30° rotation) to that of g. Resolution is, in general, better for cBbRAG1L than BbRAG2L. g, h, Cryo-EM maps of cBbRAGL bound to intact 3'TIRs (5.3 Å overall resolution; g) or nicked 3'TIRs (5.0 Å overall resolution; h). One BbRAG1L subunit (grey) has been rendered partially transparent to allow visualization of DNAs inside the protein. Continuous DNA density running through the protein core is visible with nicked but not intact TIRs, which suggests that the DNA in the vicinity of the active site becomes more rigidly constrained upon nicking. This is notable in light of the recent finding that DNA in the RAG active site melts and swivels in preparation for nicking<sup>23</sup>. Clear differences between the two DNAs are visible in the bottom half of the structures, with 3'TIR-a (orange) protruding below the protein and density for 3'TIR-b (red) dissipating before the DNA emerges from the protein core. This argues that the two identical DNA molecules are engaged differently by cBbRAGL, with one (3'TIR-b) less rigidly constrained by its interactions with protein.

#### **RESEARCH** ARTICLE



Extended Data Fig. 3 | Structural features of cBbRAGL. a, Comparison of the models of cBbRAGL and cRAG (PDB 5ZDZ) bound to nicked DNA but with DNA removed, illustrating the absence of NBD\* from the *cBb*RAGL structure. NBD is a dimer that can pivot on a flexible hinge to accommodate the different spacer lengths of 12RSS and 23RSS, providing a structural explanation for the 12/23 rule<sup>20–22,58</sup> We speculate that NBD\*, HMGB1 and distal TIR sequences constitute a flexible domain located below the main complex, by analogy with RAG-RSS complexes. b, Superimposition of cBbRAGL-nicked 3'TIR synaptic complex with RAG-nicked RSS synaptic complex (PDB 5ZDZ). c, BbRAG2L adopts a doughnut-shaped structure consistent with that of a six-bladed  $\beta$ -propeller. Because of low resolution, some elements cannot be unambiguously modelled as  $\beta$ -strands. Putative  $\beta$ -propellers I–VI are labelled, as are the N and C termini of the protein, showing that-as with RAG2-propeller I is composed of both N- and C-terminal sequences. d, Colour-coded linear diagram of cBbRAG1L subdomains (top) and cartoon of the BbRAG1L dimer (bottom) with the subdomains of one subunit colour-coded as in the linear diagram. The other subunit is grey

except for the preR subdomain. Stars indicate a gap in the BbRAG1L model that spans aa 603-630. Nomenclature and figure layout as in ref. 20. DDBD, dimerization and DNA binding domain; PreR, pre-RNase H domain; RNH, RNase H domain; ZnC2 and ZnH2, domains that contribute two cysteines and two histidines, respectively, for zinc coordination; CTD, C-terminal domain; CTT\*, C-terminal tail. e, Superimposition of cryo-EM map on the model of the nicked 3'TIR in the vicinity of the flipped bases near the site of nicking. f, g, Threedimensional classes of cryo-EM maps of cBbRAGL bound to intact (f) or nicked (g) 3'TIRs (DNA omitted). One class is enlarged and shown from two vantage points below. The arrow points to the cleft that narrows in the open-to-closed transition. With intact DNA, three distinct 3D classes are distinguishable that vary in the degree of closure of the two arms of the V. h, Superimposition of three forms of cBbRAGL illustrating the movement of a 3'TIR and *Bb*RAG2L subunit (colour-coded as in **e**, **f**) that takes place during the open-to-closed transition. One cBbRAG1L-2L dimer has been aligned and movement is visualized in the other dimer.



**Extended Data Fig. 4** | **Protein–DNA interactions in the** *cBb***RAGL–nicked 3'TIR synaptic complex. a**, Schematic diagram of the detailed interactions between *Bb*RAG1L and nicked 3'TIR DNA. Bold underlined text, main-chain interactions; regular text, side-chain interactions; purple text, interactions involving *Bb*RAG1L subunit a (defined as the subunit with an active site that engages the TIR depicted); cyan text, interactions involving *symmetric Bb*RAG1L subunit b. *Bb*RAG2L–DNA interactions

could not be unambiguously assigned and are not depicted. **b**, **c**, Orthogonal views of the nicked 3'TIR–*Bb*RAG1L subunit a interaction (**b**) and the nicked RSS–RAG1 subunit a interaction (**c**). Protein electrostatic surface potential is indicated with blue (positive charge) and red (negative charge) using the scale (KT/e) below **d**, **e**. **d**, *Bb*RAG1L subunit b–nicked 3'TIR interaction. **e**, RAG1 subunit b–nicked RSS interaction.

#### RESEARCH ARTICLE

а



Extended Data Fig. 5 | CTT, CTT\* and mutational analysis of ProtoRAG TIRs. a, Superimposition showing CTT\* extending from a structurally conserved region at the C terminus of the catalytic core regions of mouse RAG1 (mRAG1), zebrafish RAG1 (zRAG1) and BbRAG1L. b, Sequence alignment of CTT from six vertebrate RAG1 proteins. Species name abbreviations used in this paper: Mm, Mus musculus (mouse); Hs, Homo sapiens (human); Gg, Gallus gallus (chicken); Xl, Xenopus tropicalis (frog); Dr, Danio rerio (zebrafish); Bb, Branchiostoma belcheri (amphioxus); Pf, Ptychodera flava (acorn worm); Sp, Strongylocentrotus purpuratus (purple sea urchin); Af, Asterias forbesi (sea star); Hv, Hydra vulgaris (hydra); Rn, Rattus norvegicus (rat). c, Schematic indicating sub-regions of TIRs. Region 1 contains the heptamer and one additional bp, which in Fig. 1a

and throughout the paper is defined as part of TR2. Otherwise, region 2 (broken up into 2a and 2b for the 5'TIR) corresponds to TR2. Poorly conserved regions 3 and 4 separate TR2 from a distal conserved 9-bp element (region 5). **d**-**g**, Cleavage of substrates containing a single 5'TIR (d, e) or a single 3'TIR (f, g), either intact (WT) or with the indicated region scrambled, by  $cBbRAGL(\mathbf{d}, \mathbf{f})$  or the  $\Delta NBD^* cBbRAGL$  complex (e, g). Closed and open arrowheads, 5'TIR and 3'TIR cleavage products, respectively. Region 5 is completely dispensable for cleavage, and regions 3 and 4 contribute modestly to 3'TIR-but not 5'TIR-cleavage. Upon deletion of NBD\* from cBbRAG1L, 3'TIR cleavage loses all dependency on regions 3 and 4, consistent with the possibility that NBD\* engages in functionally important interactions with regions 3 and 4 of the 3'TIR.

#### ARTICLE RESEARCH



Extended Data Fig. 6 | Activities of chimeric RAG1-BbRAG1L proteins and residues that influence coupled cleavage. a, b, Cleavage by NBD-CC\* is dependent on the length of the spacer between the TIR heptamer and the RSS nonamer. Substrates depicted schematically above the gel images. In a, the substrates contain a single target based on T1 (Fig. 3b), the spacer of which ranges in length from 10 to 14 bp. In **b**, the substrate contains target T1 and a partner target based on T2 (Fig. 3b), the spacer of which ranges in length from 20 to 25 bp. Dark arrowheads, T1 cleavage products; open arrowheads, T2 cleavage products. c, d, Cleavage reactions using the NBD\*-CC-CTT\* and CC-CTT\* proteins and T3 and T4 substrates (all depicted schematically in Fig. 3c), as indicated above the lanes. T3\* and T4\*, T3 and T4 targets with a C-to-A mutation of heptamer position 1 that renders the target uncleavable; [T4/T4]H+TR2 and [T3/ T3]H+TR2, substrates in which both targets have had all substrate sequences except the heptamer and TR2 deleted. Asterisks as in Fig. 2g. e, Cartoon depicting differences in the major protein-DNA interactions of BbRAGL and RAG. f, Superposition of RAG1 and BbRAG1L in the region containing Glu649 and Ser963 in complexes bound to nicked DNA substrates, illustrating the similarity of positioning of the active site

residues Glu962 and Glu1063 and flanking residues Asn961 and Asn1062. h, RAG1 Asn961 and BbRAG1L Asn1062 have the potential to participate in hydrogen-bond networks after nicking and could thereby stabilize the hairpin-competent configuration of the enzyme. This is notable in light of the fact that mutant RAG1(N961A) displays enhanced coupled cleavage compared to wild-type RAG1. i, Cleavage reactions using wildtype and mutant cBbRAG1L proteins (with BbRAG2L) and substrates containing one or two TIRs as indicated above and below the lanes (left). cBbRAG1L(V751E), but not cBbRAG1L(A1064S), reduces uncoupled single 3'TIR cleavage (lower black asterisk, lane 2; reduction also seen in lane 8) and single 5'TIR cleavage (seen most clearly in lane 5). The strong reduction in cleavage seen with the BbRAG1L(V751E/A1064S) double mutant suggests the possibility that hydrogen bonding between these two residues holds the active site in an inactive configuration. Right, quantification of uncoupled cleavage as the ratio of the intensity of the 3'TIR single cleavage band (lower black asterisk) to that of the double cleavage band (red asterisk) as in lanes 1–3. Mean  $\pm$  s.e.m. Two-tailed *t*-test: \*\**P* < 0.01, compared to wild-type *cBb*RAG1L. ns, not significant.

#### **RESEARCH** ARTICLE



Extended Data Fig. 7 | In vitro transposition by wild-type and mutant RAG proteins. a, Schematic of intramolecular transposition. If the 3' OH nucleophiles attack the strand on which they are located, the products are two deletion circles (top), but if they attack the opposite strand, a single inversion circle product is generated (bottom). Staggered attack on the target DNA backbone yields single-stranded gaps in the products, represented as five short vertical lines. **b**, Inverse PCR reaction to amplify inversion circles from purified intramolecular transposition product as in Fig. 5d, third lane. The band indicated with an arrow was excised, cloned and sequenced, yielding sites at which intramolecular transposition occurred to yield inversion circles, indicated in the map of the excised 12/23RSS central fragment (below). Half arrows indicate approximate locations of PCR primers. The location of deletion circle joints detected by sequencing are not indicated. c, Schematic of intermolecular in vitro transposition assay. An RSS-flanked Tet gene is mobilized from a linear donor by RAG-mediated DNA cleavage and can transpose into a target plasmid, which is detected after bacterial transformation by the appearance of colonies on Kan/Tet/Str (KTS) plates (streptomycin (str) is not relevant in this assay). d, In vitro DNA cleavage and intramolecular transposition by position 848-mutant cRAG1 (with RAG2(1-383)). Increased transposition compared to wild-type cRAG1 is revealed by diminished intensity of the double cleavage band and increased intensity of the slow-migrating intramolecular inversion circle transposition product band (red arrow). The intensity of the inversion circle band underestimates the efficiency of transposition because deletion circle

not visible as a discrete band-are also produced<sup>18</sup>. e, Quantification of intramolecular transposition efficiency from three independent experiments as in d, measured by ratio of double cleavage band to 23RSS cleavage band (the latter serving as an internal control for the total amount of cleavage). The ratio decreases as intramolecular transposition increases in efficiency, consuming the double cleavage band. Mean, with data range indicated by box. Two tailed *t*-test; *P* values are indicated. **f**, Distribution of transposition target site duplication lengths determined by sequencing of plasmid transposition products or from high-throughput sequencing of plasmid-to-genome transposition products (Extended Data Fig. 9d), as indicated above the bars. The RAG1 protein used is indicated below the bars. In vitro reactions as in Fig. 5e using RAG2(1-383); in vivo plasmid target reactions as in Fig. 5g using RAG2(1-350); genome transposition products generated using RAG2(1-350). In a small fraction of plasmids, sequencing revealed deletions at the site of insertion of the RSSs (red; deletion). g, In vitro cleavage and intramolecular transposition reactions using RAG2(1-352) and RAG2(1-383) (as indicated above the lanes) and wild-type or mutant cRAG1 (as indicated below the lanes). Transposition is readily detected with both forms of RAG2 and is increased by the RAG1(R848M) mutation. h, In vitro intermolecular transposition assays using RAG2(1-383) and RAG2(1-352) and wild-type or mutant cRAG1 (as indicated below the lanes). Deleting the RAG2 acidic hinge does not increase the efficiency of intermolecular transposition in vitro.

#### ARTICLE RESEARCH



Extended Data Fig. 8 | In vivo transposition by RAG and BbRAGL proteins. a, Schematic of plasmid-to-plasmid in vivo transposition assay. An RSS-flanked Tet gene is mobilized from a donor plasmid by RAGmediated DNA cleavage and can transpose into a target plasmid, which is detected after bacterial transformation by the appearance of colonies on KTS plates (streptomycin (str) reduces background in the assay by selecting against bacteria containing the *rpsL* gene, present in the donor plasmid). b, Schematic of in vivo GFP fluorescence recombination assay, used to generate data in c (right), e (right) and g. Excision of the polyadenylation sequence (poly-A) together with its flanking RSSs or TIRs (triangles) by RAG or BbRAGL and resealing of the plasmid allows expression of GFP. c, In vivo transposition (left) and recombination (right) activity in HEK293T cells of wild-type *Bb*RAG1L and *Bb*RAG1L(M949R) (together with *Bb*RAG2L). Mean  $\pm$  s.e.m. Two-tailed *t*-test: \*\*\**P* < 0.005, compared to wild-type BbRAG1L. d, In vivo transposition activity assayed in human colon cancer cell line HCT116 with full-length RAG1(R848M/ E649V) and either RAG2(1-350) or RAG2(1-383). As in HEK293T

cells, transposition is strongly inhibited by the RAG2 acidic hinge. Mean  $\pm$  s.e.m. **e**, In vivo transposition (left) and recombination (right) activity in HEK293T cells of human wild-type RAG1 and RAG1(R851M) with different forms of human RAG2, beginning at amino acid 1 and ending with the amino acid indicated below the bars. Mean  $\pm$  s.e.m. Two-tailed *t*-test: \*\*\**P* < 0.005; \*\*\*\**P* < 0.001 compared to wild-type human RAG1.  $\mathbf{f}$ ,  $\mathbf{g}$ , Protein expression ( $\mathbf{f}$ ) and recombination activity (g) in HEK293T cells of wild-type and mutant mouse RAG1 and RAG2 proteins used in the in vivo transposition assays. The data show that the large increases in transposition activity observed with some proteins (for example, RAG2(1-350), RAG2(1-352), and RAG1(R848M)) are not due to large increases in protein expression or cleavage/recombination activity. h, i, In vivo transposition activity assayed in HEK293T cells with full length RAG1(R848M) (h) or RAG1(R848M/E649V) (i) and various forms of RAG2, beginning at amino acid 1 and ending with the amino acid indicated below the bars. FL, full-length RAG2.

#### **RESEARCH** ARTICLE



Extended Data Fig. 9 | Transposition into the human genome by mutant RAG proteins. a, Schematic of plasmid-to-genome in vivo transposition assay. An RSS-flanked Puro expression cassette is mobilized from a plasmid donor by RAG-mediated DNA cleavage and can transpose into the genome, which is detected by selection with puromycin and highthroughput sequencing. b, Schematic illustrating detection of bone fide transposition events into the genome by LAM-PCR and high-throughput sequencing. LAM-PCR is performed on genomic DNA with biotinylated primers (half arrows) that extend into the DNA flanking either the 12RSS or 23RSS; thereafter, independent libraries are prepared and sequenced for the 12RSS and 23RSS flanks. If the donor plasmid randomly inserts into genome (i), then the RSS is flanked by donor plasmid sequences. If the RSS fragment is cleaved at one or both RSSs and randomly inserted into genome (ii), then a match with an appropriate sequence duplication (indicative of a TSD) will not be found between the 12RSS and 23RSS libraries. Finally, if the RSS fragment is inserted into the genome by transposition (iii), a match with a 3-7-bp TSD will be found in the 12RSS and 23RSS libraries. c, Tissue culture plates stained with crystal violet showing puromycin-resistant colonies for experiments using RAG2(1-350) and either wild-type RAG1 or RAG1(R848M/E649V). Colony numbers increase about twofold with the mutant RAG1 protein but many

colonies are seen with wild-type RAG1 owing to random integration of the donor plasmid. Essentially no colonies are seen if the donor plasmid is omitted (first column of plates) **d**, Summary of sequence data from the plasmid-to-genome transposition experiments. For each of the six libraries, column 1 shows the total number of reads with a barcode and RSS, columns 2 and 3 show a breakdown of number of reads in which RSS flanking sequences map to the human genome or the donor plasmid (a small fraction of reads do map to either genome or plasmid owing to poor read quality), column 4 shows the number of unique reads that map to the genome (after elimination of duplicates) and column 5 shows the number of bone fide transposition events detected. e, Rainfall circos plot of transposition events into chromosomes of HEK293T cells. f-h, Genome features of transposon integration sites mediated by RAG1(R848M/ E649V) and RAG2(1-350). f, Number (per cent) of transposition events into the genome features indicated. One-tailed Fisher's exact test was used to determine whether the frequency of transposition events was greater than that expected by chance: genes ( $P = 9 \times 10^{-30}$ ); protein-coding genes  $(P = 5 \times 10^{-35})$ ; exons  $(P = 6 \times 10^{-86})$ ; protein-coding exons  $(P = 4 \times 10^{-82})$ and within 2 kb of a TSS  $(P = 5 \times 10^{-180})$ . **g**, **h**, Meta-analysis of integration sites within gene bodies (g) and flanking TSSs (h).



**Extended Data Fig. 10** | **Model of RAG evolution in metazoans.** Steps leading from the ancestral *Transib* transposon, consisting of a *RAG1-like* ORF flanked by RSS-like TIRs<sup>34</sup>, to the RAG recombinase and split antigen receptor genes of jawed vertebrates. I, capture of a *RAG2-like* ORF by a *Transib* transposon to generate the ancestral RAG transposon in an early deuterostome; II, key events in the evolution of *RAG1*, *RAG2* and antigen receptor genes of jawed vertebrates: A, insertion of the RAG transposon into the exon of a gene encoding an immunoglobulin-domain receptor protein to generate the ancestral antigen receptor gene; B, loss of CTT\* and acquisition of Glu649 and Ser963 by RAG1 facilitated evolution of

the 12/23 rule and coupled cleavage, respectively, while acquisition of RAG1 Arg848 and the RAG2 acidic hinge powerfully suppressed RAG transposition activity. The order of events depicted in II is not known. RAG-related elements, if found in members of a given lineage, are indicated at right, as is the presence of the CTT\* domain. Protostome lineages have been collapsed into a single branch. While vertical transmission is consistent with the distribution of RAG1 and RAG2 transposon and recombinase elements in deuterostomes<sup>11</sup>, horizontal transmission might have contributed to the spread of *Transib* elements.

#### Extended Data Table 1 | Cryo-EM data collection, refinement and validation statistics

	BbRAGL-3'TIR	BbRAGL-3'TIR	BbRAG1L-	BbRAG1L-
	synaptic	synaptic	3'TIR synaptic	3'TIR synaptic
	complex with	complex with	complex with	complex with
	nicked DNA	nicked DNA	intact DNA	intact DNA
	refined with C2	refined with c1	refined with C2	refined with C1
	symmetry	symmetry	symmetry	symmetry
	(EMDB-7046)	(EMDB-7045)	(EMDB-7044)	(EMDB-7043)
	(PDB 6B40)	`````	· · · ·	× ,
Data collection and processing	· · ·			
Magnification	81,000	81,000	81,000	81,000
Voltage (kV)	300	300	300	300
Electron exposure $(e - / Å^2)$	54	54	80	80
Defocus range (µm)	-1.2 to -2.5	-1.2 to -2.5	-1.2 to -2.5	-1.2 to -2.5
Pixel size (Å)	1.35	1.35	1.35	1.35
Symmetry imposed	C2	C1	C2	C1
Initial particle images (no.)	496,221	496,221	414,309	414,309
Final particle images (no.)	350,143	205,845	94,922	94,922
Map resolution (Å)	4.3	5.0	4.6	5.3
FSC threshold	0.143	0.143	0.143	0.143
Map resolution range (Å)	4.1 to 5.3	4.8 to 6.5	4.5 to 6.6	5.0 to 7.1
Refinement				
Initial model used (PDB code)	<b>3</b> JBY			
Model resolution (Å)	4.5			
FSC threshold	0.5			
Model resolution range (Å)				
Map sharpening B factor ( $Å^2$ )	-258			
Model composition				
Non-hydrogen atoms	16204			
Protein residues	1838			
Ligands	4			
B factors (Å <sup>2</sup> )				
Protein	50			
Ligand	50			
R.m.s. deviations				
Bond lengths (Å)	0.0078			
Bond angles (°)	1.35			
Validation				
MolProbity score	2.14			
Clashscore	7.58			
Poor rotamers (%)	0			
Ramachandran plot				
Favored (%)	80.94			
Allowed (%)	18.84			
Disallowed (%)	0.22			

Summary of relevant parameters used during cryo-EM data collection and processing. Refinement and validation statistics are provided for the molecular model of the BbRAGL-3'TIR synaptic complex with nicked DNA with C2 symmetry.

## natureresearch

Corresponding author(s): Yong Xiong, Anlong Xu, David G. Schatz

## **Reporting Summary**

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

#### Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

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

Our web collection on statistics for biologists may be useful.

#### Software and code

#### Policy information about availability of computer code

Data collection	No software was used.
Data analysis	Image analysis was performed using MotionCor2 ver 1.1, Gctf 1.06, RELION1.4, and RELION-2.1. Initial BbRAGL model construction was performed using the Discovery Studio software suite from Accelrys 3.0 and refined using NAMD 2.12 and QA-RecombineIT. The structural model was refined using COOT 0.8, Rosetta FastRelax protocol, and PHENIX, and validated using MolProbity and EMringer. Molecular representations were generated in PyMol and Chimera. Cutadapt version 1.16 was used to divide multiplexed sequencing reads by barcoded conditions and to trim the barcode sequences, primer sequences, and 12RSS and 23RSS sequences. Bowtie2 (version 2.2.9) was used to align to GRCh38 with "very sensitive" presets in end-to-end-alignment mode. High-quality alignments (MAPQ >= 30; identified with Samtools 1.5) were converted to bed intervals using the bedtools bamToBed utility (bedtools version 2.27.1). Bedtools intersect was used to identify loci where corresponding 12RSS and 23RSS libraries showed evidence of transposition events on opposite strands.

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

Policy information about availability of data

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

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

The model of the cBbRAGL-nicked 3'TIR synaptic complex has been deposited in the Protein Data Bank with accession code PDB: 6B40. The cryo-EM maps of cBbRAGL in complex with intact or nicked 3'TIRs have been deposited in EMDataBank with accession codes EMD-7043, 7044, 7045, and 7046. High-throughput DNA sequence data to identify transposition events in the human genome have been deposited in the NCBI Sequence Read Archive with accession codes SRR8430227-SRR8430233 (Project PRJNA514369).

## Field-specific reporting

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

K Life sciences

Behavioural & social sciences E

Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see <u>nature.com/authors/policies/ReportingSummary-flat.pdf</u>

## Life sciences study design

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

Sample size	No sample-size calculations were performed. Such calculations were not relevant given the types of data reported in our study. The number of times that experiments were repeated is addressed under "Replication" below.
Data exclusions	No data were excluded from the analyses.
Replication	Reproducibility of biochemical experiments was verified by performing such experiments (e.g., DNA cleavage assays) at least 3 times, except as noted in Methods under "Statistics and Reproducibility". For transposition assays, reproducibility was verified by repeating experiments at least three times, and the data for each assay are shown as an individual symbol. Each replicate was an independent experiment and did not represent re-assay of the same material. All attempts at replication were successful.
Randomization	Randomization was not relevant to our study because our study did not involve the allocation of samples/organisms/participants into experimental groups.
Blinding	Investigators were not blinded to group allocation because group allocation was not involved in our study. Investigators were not blinded during data collection because the data being collected were quantitative in nature (gels or numbers of colonies on a plate) and were not prone to subjective interpretation.

## Reporting for specific materials, systems and methods

#### Materials & experimental systems Methods n/a Involved in the study n/a Involved in the study Unique biological materials $\mathbf{X}$ ChIP-sea |X|Antibodies K Flow cytometry Eukaryotic cell lines MRI-based neuroimaging Palaeontology Х Animals and other organisms Human research participants

Antibodies

Antibodies used

Validation

anti-RAG1 monoclonal antibody #23; anti-RAG2 monoclonal antibody #39; anti-beta-actin monoclonal antibody (Sigma #A1978)

The specificity of the anti-RAG antibodies has been verified using RAG-knockout cells and protein extracts from those cells in western blots, ChIP-seq, and other assays. The specificity of the anti-beta-actin antibody was verified as a single intense band of

the expected molecular weight by western blot, a finding also reported on the manufacturer's website in multiple mouse and human cell lines.

#### Eukaryotic cell lines

HEK293T cells were obtained from ATCC. Expi293 cells were obtained from Thermo Fisher Scientific. HCT116 cells were obtained from Eric Hendrickson, University of Minnesota.
Cell lines used were not authenticated.
Cell lines were not tested for mycoplasma contamination.
None

#### Flow Cytometry

#### Plots

Confirm that:

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

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

All plots are contour plots with outliers or pseudocolor plots.

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

#### Methodology

Sample preparation	Expi293F cells were harvested 72 h post-transfection, washed twice with PBS containing 1% FBS, stained with DAPI (4',6-diamidino-2-phenylindole).
Instrument	Stratedigm STD-13L
Software	FlowJo 10.3
Cell population abundance	Live cell sorting was not performed. Data are reported on live cells as determined by DAPI exclusion.
Gating strategy	Negative gate established in cells transfected with GFP reporter vector but no RAG or BbRAGL expression vectors; see Supplementary Data 2.

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