EXTENDED EXPERIMENTAL PROCEDURES

Genome assemblies

We sequenced the diploid genome of an outbred Chinese lancelet (*Branchiostoma belcheri*) using next-generation technology. To overcome the high allelic polymorphism (5% on average), a novel assembly pipeline was developed to separate and reconstruct two haploid assemblies: the reference assembly (426 Mb), and the alternative assembly (416 Mb) that hold alternative alleles (Huang et al., 2012; Huang et al., 2014). The draft genome of another lancelet species (the Florida lancelet *Branchiostoma floridae*) was downloaded from the JGI website: http://genome.jgi-psf.org/Brafl1/Brafl1.home.html/.

Direct homology-based searches for transposon sequences and motifs

We used NCBI-BLAST to search the lancelet genomes for homologs of the RAG1/2-like protein-coding DNA fragments. Two methods were used to search for target site duplication (TSD) and terminal invert repeat (TIR) sequences in the vicinity of the RAG1/2-like homologs. The upstream and downstream 20 kb of sequence flanking the RAG1/2-like sequences were extracted and separated into a set of small fragments (using a window size of 80 bp and a step size of 1 bp). In the first method, each upstream fragment was compared with each downstream fragment for 3-7 bp TSDs and possible TIRs using a custom Perl script. We required 45% identity for potential TIR pairs, and allowed no mismatches for 3-4 bp TSD pairs and only one mismatch for 5-7 bp TSD pairs. In the second method, all upstream fragments were compared against all downstream fragments using BLAST, and all fragments were compared against vertebrate recombination signal sequences (RSSs) and *Transib* TIRs using BLAST. We required a minimum e-value of 100 and sequence identity of 45% in the BLAST search. The results of the two methods were combined using a custom Perl script. To identify potential non-TE recombination sites, we used a custom Perl script to search for motifs like these: "heptamer-spacer (10-50 bp)-nonamer". The consensus RSS heptamer and nonamer sequences used are shown in Figure 2B.

Identification of polymorphic (allele-specific) TE insertions

The LASTZ-ChainNet method was used to create a whole-genome alignment between the reference and the alternative haploid genome assemblies of the Chinese lancelet (Kent et al., 2003; Schwartz et al., 2003). LASTZ was tuned to high sensitivity mode to guarantee accurate alignment at the gap borders with the following special parameter settings: --masking=0, --hspthresh=3000, --ydrop=3400, --gappedthresh=3000, --gap=400,30, --step=1, --seed=12of19, --identity=75, and the score matrix "100 -225 -225 -225; -225 100 -225 -225; -225 -225 100 -225; -225 -225 -225 100". The LASTZ alignments were processed into reciprocally-best single-coverage chain-net alignments according to UCSC's documentation. Special parameters for axtChain and chainNet include --linearGap=medium, --minScore=2000 and –minSpace=50. Non-repeat-masked assemblies were used to allow TE sequences for alignments. The large polymorphic insertions or alignment gaps (300-10,000 bp) were identified from the output net file. These inserted sequences were searched for RAG-like DNA fragments using BLAST. The insert sequences containing RAG-like fragments were further searched for possible TSDs and TIRs by manual examination of their terminal sequences.

PCR screening of BAC libraries for transposon copies

The BAC libraries of *B. belcheri* were pooled together for PCR screening. Verified primer pairs targeting lancelet RAG1- and RAG2-like DNA fragments were used: BbRAG1_U3: TTCAATAGCATGGTAGAT, BbRAG1_L3: TCATTAAGGTTGGGTTTA, BbRAG2_U4: ATCGGTAGTGTTGACATA, BbRAG2_L4: GACTTCTCGCACATATTC. Bacteria harboring BAC clones were pooled in groups of 96 and cultured

overnight in 384-well plates with Luria Broth and chloromycetin (12.5 μg/mL). PCR was performed in 10 ul reactions with the addition of SYBR dye and raw bacteria culture medium (0.1 l) on a Roche 480 LightCycler for forty cycles (2 min at 95°C followed by 40 cycles of 30 s at 95°C, 15 s at 60°C, and 30 s at 72°C.). Melting curves were recorded and used to identify positive clone batches. Positive batches containing both RAG1- and RAG2-like DNA fragments were selected from a new round of PCR screening until the single positive clone was identified.

BAC plasmid sequencing and assembly

BAC plasmid clone BAC73 was sequenced to high quality using Sanger sequencing and primer walking. Traditional Sanger sequencing was performed on an ABI 3730 sequencer, with each step using a specific primer designed based on newly-obtained reads. Finally, all reads were assembled into a gap-free scaffold based on the overlaps between reads. Thirteen other BAC plasmid clones were sequenced to >300x coverage each using a shotgun strategy and the Illumina MiSeq 2×300bp system. Shotgun reads of each clone were assembled using CABOG v8.1 (http://wgs-assembler.sourceforge.net/wiki/index.php?title=Main_Page) and Arachne v3.2 (ftp://ftp.broadinstitute.org/pub/crd/ARACHNE/) using default settings.

Alignment and phylogenetic analysis

Alignments were produced using CLUSTALW (for proteins) and MAFFT (for DNA) and were edited and annotated using GeneDoc. Phylogenetic analysis and molecular dating analysis were performed using MEGA5 (Tamura et al., 2011). The molecular trees were computed with these settings: the 3'-terminal 700 bp sequences from sixteen *ProtoRAG* copies (see Table S1), the Neighbor-Joining method, 1000 bootstrap replicates, pairwise deletion, uniform rates among sites and Maximum Composite Likelihood.

Plasmid transfection and cDNA cloning

Human embryonic kidney (HEK) 293T cells were grown in Dulbecco's modified Eagle's medium and transfected with the purified BAC plasmid (clone ID: 73) one day after passage. Total RNA from transfected 293T cells and *B. belcheri* tissues were purified using a Qiagen RNeasy Plus Mini kit and treated with Promega DNaseI. The 5'RACE and 3'RACE for *B. belcheri* RAG1-like and RAG2-like full-length cDNA were performed using the GeneRACE Kit (Invitrogen) according to the manufacturer's protocol. Once the full-length cDNA sequences were obtained, specific primers were designed to clone different alternatively-spliced mRNA isoforms from the total RNA of both lancelet tissues and the 293T cells transfected with the BAC plasmid clone 73.

Quantitative RT-PCR

Total RNA was purified from lancelet (*B. belcheri*) tissues using a Qiagen RNeasy Plus Mini kit and then treated with Promega DNaseI. Double-stranded cDNA was synthesized from total RNA using the SYBR Perfect real-time series kits (Takara Bio). Real-time quantitative RT-PCR was performed on Roche's LightCycler 480 real-time PCR system. Quantitative PCR (qPCR) testing for each sample was performed in six replicates using a 10- l reaction system containing 50 ng initial RNA and annealing and extension temperatures of 60°C. Forty PCR cycles were run and melting curves were recorded. All other parameters for the reaction system and the PCR program were set according to the manufacturer's protocol for the SYBR PrimeScript RT-PCR kit (Takara Bio). GAPDH was used as internal control to normalize expression levels. Expression level was calculated using the comparative $2^{-\Delta\Delta}$Ct method. All samples were analyzed in three replicates with results expressed as the mean ± SD. Primer pairs used include:

RAG1F: GTAAGGATGGAGCAGACGGGATG,
RAG1R: GCGGTTAGAGCGGACGGAAT,
RAG2F: GACATCACCATCACACCAA,
RAG2R: CAGCCCTCCAAACAGAAT,
GAPDHF: TTCACCACCATCGCCAAG,
GAPDHR: CTTCTCATACTTCTCCTCGTTCAC.

BbRAG1L/2L expression plasmids used in *ex vivo* analysis

The full-length cDNAs encoding the longest proteins of bbRAG1L, bbRAG2L and mouse RAG1/2 were used for all sub-cloning and plasmid construction. Full-length *bbRAG1L* and *bbRAG2L* used for subcellular localization analysis were cloned into the pEGFP plasmid. The pCDNA3.1-based eukaryotic expression constructs bbRAG1-pcDNA3.1-Flag and bbRAG2-pcDNA3.1-HA were used for full-length bbRAG1L and bbRAG2L protein expression in 293T cells. The core region of mouse RAG1 (aa 384-1008) was subcloned into the PCDNA3.0 to create eukaryotic expression construct mRAG1C-pcDNA3.0-Flag. The core region of mouse RAG2 (aa 1-387) was subcloned into pCDNA3.0 to create eukaryotic expression construct mRAG2C-pcDNA3.0-HA. All DNA substrates for RAG-like proteins were inserted into pEGFP-N1 at suitable restriction enzyme sites.

Substrate plasmids used in *ex vivo* analysis

A series of plasmids containing substrate for bbRAG1L/2L have been constructed. The plasmid pTIRG1 contains an insert (i.e., an artificial transposon) defined by the first 240 bp of the *ProtoRAG* 5'-TIR, 838 bp spacer (containing multiple polyA signal motifs), and the first 170 bp inverted sequence of the *ProtoRAG* 3'-TIR. pTIRG8 contains an insert defined by the first 43 bp sequence of the *ProtoRAG* 5'-TIR, the same 838 bp spacer, and the first 47 bp inverted sequence of the *ProtoRAG* 3'-TIR. The pCJ-GFP construct contains the consensus 12-RSS sequence, the same 838 bp spacer, and the consensus 23-RSS sequence. The pTIRG2-pTIRG7 plasmid contain inserts defined by a series of truncated TIR sequences and the 838 bp spacer, while pTIRG9-pTIRG12 contain inserts defined by a series of mutated minimum TIR sequences. The plasmid pTIR104 was constructed by inserting an artificial transposon element which contains a bacterial transcription termination sequence flanked two *ProtoRAG* TIRs between the lac promoter and the chloramphenicol (Chl) resistance gene. The plasmid pTIR204 was similar to pTIR104 except with both TIRs inverted. The intermolecular transposition assay in human cells was performed using as donor the ampicillin resistance plasmid pTIR203 containing a lac promoter and a Chl resistance gene flanked by the short TIR sequences, and as the target, the pEGFP-N1 plasmid containing a kanamycin resistance gene.

Confocal immunofluorescence microscopy

HeLa cells on coverslips (10 mm × 10 mm) in a 24-well plate were transfected with 400 ng of the indicated expression plasmids by jetPEI (PolyPlus-transfection) according to the manufacturer's instructions. After 20-24 h, cells were fixed for 15 min in a 4% formaldehyde solution, washed, stained for 5 min with DAPI (4', 6-diamidino-2-phenylindole•2HCl) (Sigma), washed, and imaged using a LEICA TCS-SP5 confocal microscope.

Fluorescence microscopy and flow cytometry to detect GFP positive cells

293T cells in 12-well plates were transfected with 0.8 μg of RAG-like protein expression plasmids (either bbRAG1L/bbRAG2L or mouse RAG1C/RAG2C) and 0.4 μg of substrate plasmids (pHDJL-GFP, pHDJS-GFP,

or pCJ-GFP) as indicated by jetPrime (PolyPlus-transfection) according to the manufacturer's instructions. After 48 h, cells were examined using fluorescent microscopy then treated with trypsin, resuspended in DMEM medium and centrifuged for 5 min at 800g. Cell pellets were washed twice with PBS and resuspended in PBS at a density of 10^6 cell/ml. Then GFP expression was analyzed using a BD FACS Calibur. All samples in each independent experiment are triplicated. Data were analyzed by Student's t-test. Values of $p<0.05$ were considered significant.

PCR and sequence analysis of recombination products

Following fluorescent microscopy and flow cytometry, plasmid DNA was recovered from transfected 293T cells by alkaline lysis, treated with DpnI (NEB), purified with Gel extraction kit (Qiagen) and subject to PCR (40 cycles). PCR products were separated on 2% agarose gels. The specific PCR primers were: pTIRG-F: GTCGTAACAACTCCGC; pTIRG-R: ACGCTGAACTTGTGGC. The suspected recombination product were purified using Gel extraction kit (Qiagen), ligated into pGEM-T easy (Promega), and transformed into *E.coli* DH5α for Sanger sequencing.

Bacterial colony assays

The plasmid substrates pTIR104 or pTIR204 (0.8μg) and bbRAG1L (1.6μg) and bbRAG2L (1.6ug) expression vectors were co-transfected into 293T cells on 6 cm plates. After 48 hours, plasmid was recovered from the cells by alkaline lysis, transformed into *E. coli* DH5α, and spread on 2×YT agar plates containing 100 μg/ml Amp and 50 μg/ml Cam. Plasmids were recovered and used for sequence analysis.

Protein expression and purification for *in vitro* analysis

Codon optimized *bbRAG1L* and *bbRAG2L* open reading frames were cloned into pTT5 with an N terminal maltose binding protein (MBP) tag and co-transfected or transfected separately into expi293F™ cells using the ExpiFectamine™ 293 Transfection Kit (Gibco) (Durocher et al., 2002). Cells were harvested 60 h after transfection, frozen at -80°C, and then re-suspended in lysis buffer (25 mM Tris, pH7.5, 1 M KCl, 1 mM EDTA, 1 mM DTT). Resuspended cells were disrupted by three cycles of freeze/thaw, centrifuged at 45,000 r.p.m. for 1 h, and the cleared lysate was mixed with pre-equilibrated amylose resin (NEB) for one hour at 4°C with shaking. Resin was washed with 60 mL of lysis buffer and proteins were eluted with elution buffer (25 mM Tris, pH7.5, 0.5 M KCl, 1 mM DTT, 40 mM maltose). The eluate was concentrated and dialyzed in dialysis buffer (25 mM Tris, pH7.5, 150 mM KCl, 2 mM DTT) at 4°C for 3 h. Samples were frozen in small aliquots and stored at -80°C. Protein concentration was determined by SDS-PAGE followed by SYPRO Orange (Invitrogen) stain comparing to a BSA standard curve. Human HMGB1 was purified as described (Bergeron et al., 2006).

I*n vitro* cleavage and transposition assays

Cleavage substrate DNA was generated by PCR, using plasmid templates containing two TIRs or two RSSs arranged as depicted in Figure 5B, and purified from agarose gels. 16 l cleavage reactions contained 25 nM coexpressed MBP-bbRAG1L/2L protein (monomer bbRAG1L concentration), 175 ng HMGB1, and 10 nM substrate DNA in reaction buffer (25 mM HOPS, pH 7.0, 50 mM KCl, 2 mM DTT, 1.5 mM $MgCl_2$) and were incubated at 37°C for the indicated times. Reactions were stopped by adding 1.25 l 2.5% SDS, 5 l proteinase K (150 g/ml), 2 l 0.5 M EDTA and incubated at 55°C for at least 3 hr and were mixed with 80% glycerol before loading on 6% native TBE (tris-borate-EDTA) acrylamide gels. After electrophoresis, gels were stained with SYBR GOLD (Invitrogen) and imaged using a PharosFX™ Plus (Bio-Rad).

I*n vitro* intermolecular transposition reactions were performed as described previously (Agrawal et al.,

1998). The donor fragment was generated by PCR from a plasmid made by replacing the RSSs in pTetRSS* (Chatterji et al., 2006) with 5'- and 3'-TIRs. PCR primers were 5' phosphorylated and had the sequence: 5'-P-CACTATGAAAACTTACGTGTGCA and 5'-P-CACTATGATACTTACGCTATACCCAG). Reactions contained 25 nM co-expressed bbRAG1L/2L, 625 ng HMGB1, and 0.10 pmol of donor fragment in reaction buffer and were pre-incubated at 37°C for 15 minutes before addition of 0.25 pmol of pECPF-1 target plasmid. Reactions were incubated for another 3 hours and 45 minutes and stopped as described for cleavage reactions. DNA was ethanol precipitated, transformed into electrocompetent MC1061 bacterial cells, and plated on plates containing kanamycin and tetracycline (Kan-Tet) or, after extensive dilution, on kanacycin alone (Kan). Plasmid DNA from randomly selected colonies from Kan-Tet plates were sequenced using primers Seq5'TIR: 5'-GAATTCTCATGTTTGACAGCTTATCATCG and Seq3'TIR: 5'-CTCTTACCAGCCTAACTTCGATCACTGG. Transposition efficiency was calculated by dividing the number of colonies obtained on Kan-Tet by the number of colonies obtained on Kan (after correction for dilution).

Nick-hairpin assays

5'-Alexa 488 labeled top strand 3'-TIR oligo was mixed and boiled with unlabeled bottom strand oligo and annealed at room temperature to generate the intact fluorophore labeled 3'-TIRs substrate. Pre-nicked substrate was generated similarly but using a 5'-Alexa 488 labeled 16-mer together with an unlabeled second oligo to complete the top strand. 16 l reactions contained 50 nM coexpressed MBP-bbRAG1L/2L protein (monomer bbRAG1L concentration), 175 ng HMGB1, and 100 nM substrate DNA in reaction buffer and were incubated at 37°C for 1 hr. Reactions were stopped by adding 32 l loading buffer (95% formamide, 10 mM EDTA, 0.09% xylene cyanol, 0.09% bromophenol blue) on ice. Samples were boiled for 5 min, placed on ice immediately, and 30 l of sample was loaded on 16% 7M urea denaturing acrylamide TBE gels. After electrophoresis at 65°C, gels were imaged using a PharosFX^{TM} Plus. Oligonucleotides used were (top strand sequences are shown):

31-TIR:

Alexa488-5'-CCTTGGCAGCGCGCTGCACTATGATACTTACGCTATACCCAGCAGTGTCTGGTCGCCATC TTGAGCGCGAGATCTAAAC

Scrambled oligo:

Alexa488-5'-CCTTGGCAGCGCGCTGATAGCGTATACTTACGCTATACCCAGCAGTGTCTGGTCGCCATC TTGAGCGCGAGACTAAA

Pre-Nicked oligo: Alexa488-5'-CCTTGGCAGCGCGCTG

Statistical analyses

Results are expressed as means (+/-SEM). Data were analyzed by Student's t-test for paired or unpaired variables when appropriate. Values of $p<0.05$ (two tailed) were considered significant.

SUPPLEMENTAL REFERENCES

Agrawal, A., Eastman, Q.M., and Schatz, D.G. (1998). Transposition mediated by RAG1 and RAG2 and its implications for the evolution of the immune system. Nature *394*, 744-751.

Bergeron, S., Anderson, D.K., and Swanson, P.C. (2006). RAG and HMGB1 proteins: purification and biochemical analysis of recombination signal complexes. Methods Enzymol *408*, 511-528.

Chatterji, M., Tsai, C.L., and Schatz, D.G. (2006). Mobilization of RAG-generated signal ends by transposition and insertion in vivo. Mol Cell Biol *26*, 1558-1568.

Durocher, Y., Perret, S., and Kamen, A. (2002). High-level and high-throughput recombinant protein production by transient transfection of suspension-growing human 293-EBNA1 cells. Nucleic Acids Res *30*, E9.

Huang, S., Chen, Z., Huang, G., Yu, T., Yang, P., Li, J., Fu, Y., Yuan, S., Chen, S., and Xu, A. (2012). HaploMerger: reconstructing allelic relationships for polymorphic diploid genome assemblies. Genome Res *22*, 1581-1588.

Huang, S., Chen, Z., Yan, X., Yu, T., Huang, G., Yan, Q., Pontarotti, P.A., Zhao, H., Li, J., Yang, P., *et al.* (2014). Decelerated genome evolution in modern vertebrates revealed by analysis of multiple lancelet genomes. Nat Commun *5*, 5896.

Kent, W.J., Baertsch, R., Hinrichs, A., Miller, W., and Haussler, D. (2003). Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. Proc Natl Acad Sci U S A *100*, 11484-11489.

Lieber, M.R., Hesse, J.E., Lewis, S., Bosma, G.C., Rosenberg, N., Mizuuchi, K., Bosma, M.J., and Gellert, M. (1988). The defect in murine severe combined immune deficiency: joining of signal sequences but not coding segments in V(D)J recombination. Cell *55*, 7-16.

Schwartz, S., Kent, W.J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R.C., Haussler, D., and Miller, W. (2003). Human-mouse alignments with BLASTZ. Genome Res *13*, 103-107.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol *28*, 2731-2739.