Current Biology, Volume *22*

Supplemental Information

Genetic Consequences of Programmed

Genome Rearrangement

Jeramiah J. Smith, Carl Baker, Evan E. Eichler, and Chris T. Amemiya

Supplemental Inventory

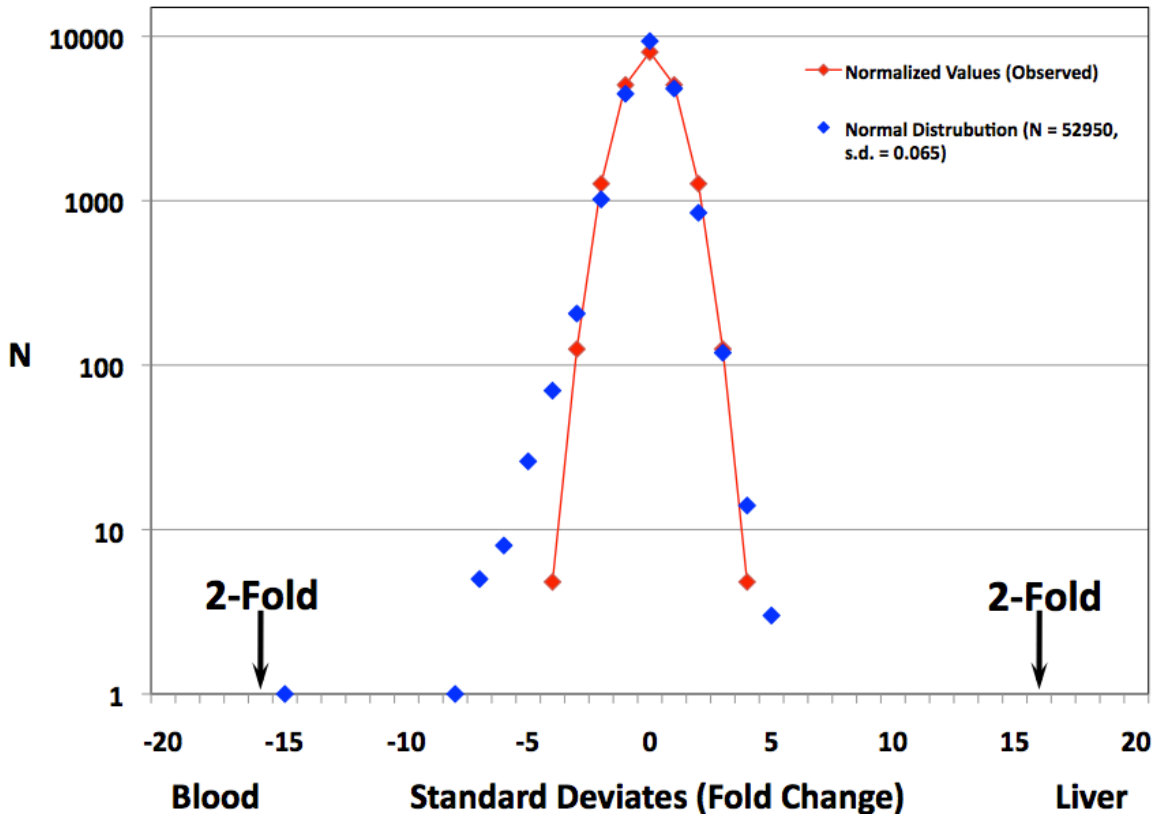

Figure S1. Analysis of Relative arrayCGH Hybridization Intensities for Blood Versus Liver Comparisons (DNA from Same Individual), Related to Figure 1

Identification of tissue-enriched candidate regions was performed by comparing observed relative hybridization intensities to a normal distribution with the same number of sampled sites (N) and standard deviation (SD). The y-axis is plotted on log scale in order to magnify differences at the tails of the distribution. s.d. = standard deviation. Arrows labeled "2-Fold" mark the values that correspond to a 2-fold ($\log_2$) difference in relative fluorescence intensity for the liver vs. blood comparison. (The relative location of 2-fold changes in Figure 1 differs substantially.)

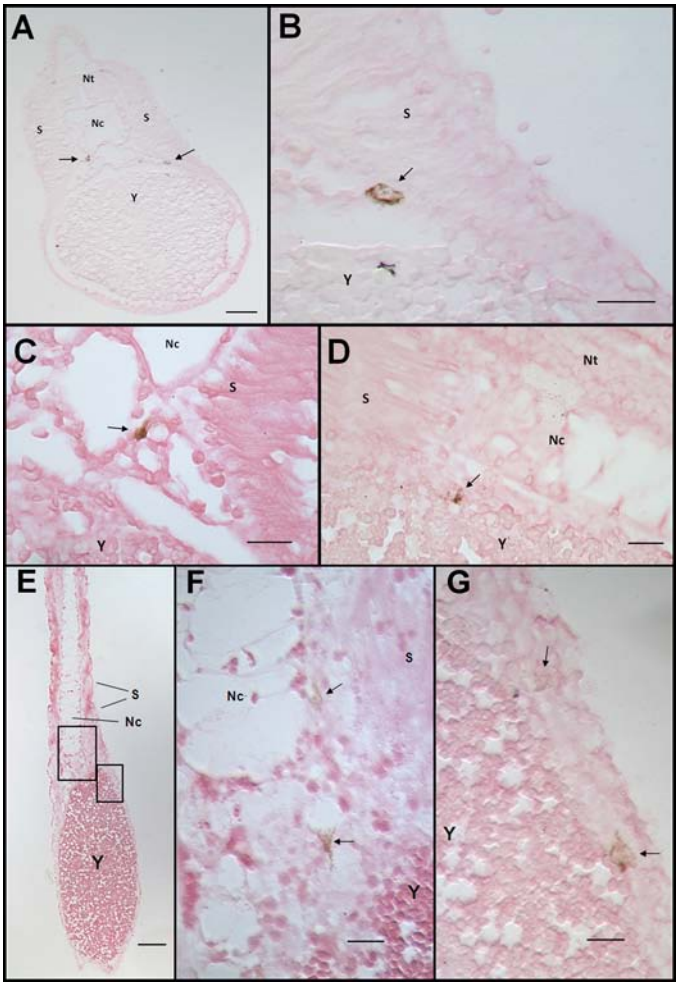

Figure S2. Assays for Stem-Cell-Specific Aldehyde Dehydrogenase Activity Provide Verification of the Developmental Origin and Migratory Path of PGCs, Related to Figure 2

Cryosections (10 μM) from day 14 embryos were stained with StemTag (Cell Biolabs, brown staining) per manufacturer's protocol and counterstained with eosin (pink). Arrows mark the position of stained cells. A-D) Cryosections of a day 14 embryo.

A) Transverse cross section (scale bar = 50 μM).

(B & C) Magnification of StemTag stained cells from transverse sections (scale bars = 20 μM). The right-most StemTag stained cell in (A) is shown in the image in panel (B) whereas the image in (C) is from a neighboring serial section. D) Sagittal section from a different day 14 embryo showing StemTag stained cell (scale bar = 20 μM).

(E-F) Oblique coronal cryosections from a day 18 embryo. F) Magnification of StemTag stained cells from the region marked by the upper left box in panel E (scale bar = 20 μM).

(G) Magnification of StemTag stained cells from the region marked by the lower right box in panel E; this image is from a neighboring serial section from the same embryo (scale bars = 20 μM). Nc = notochord, Y = yolk, S = somatic mesoderm.

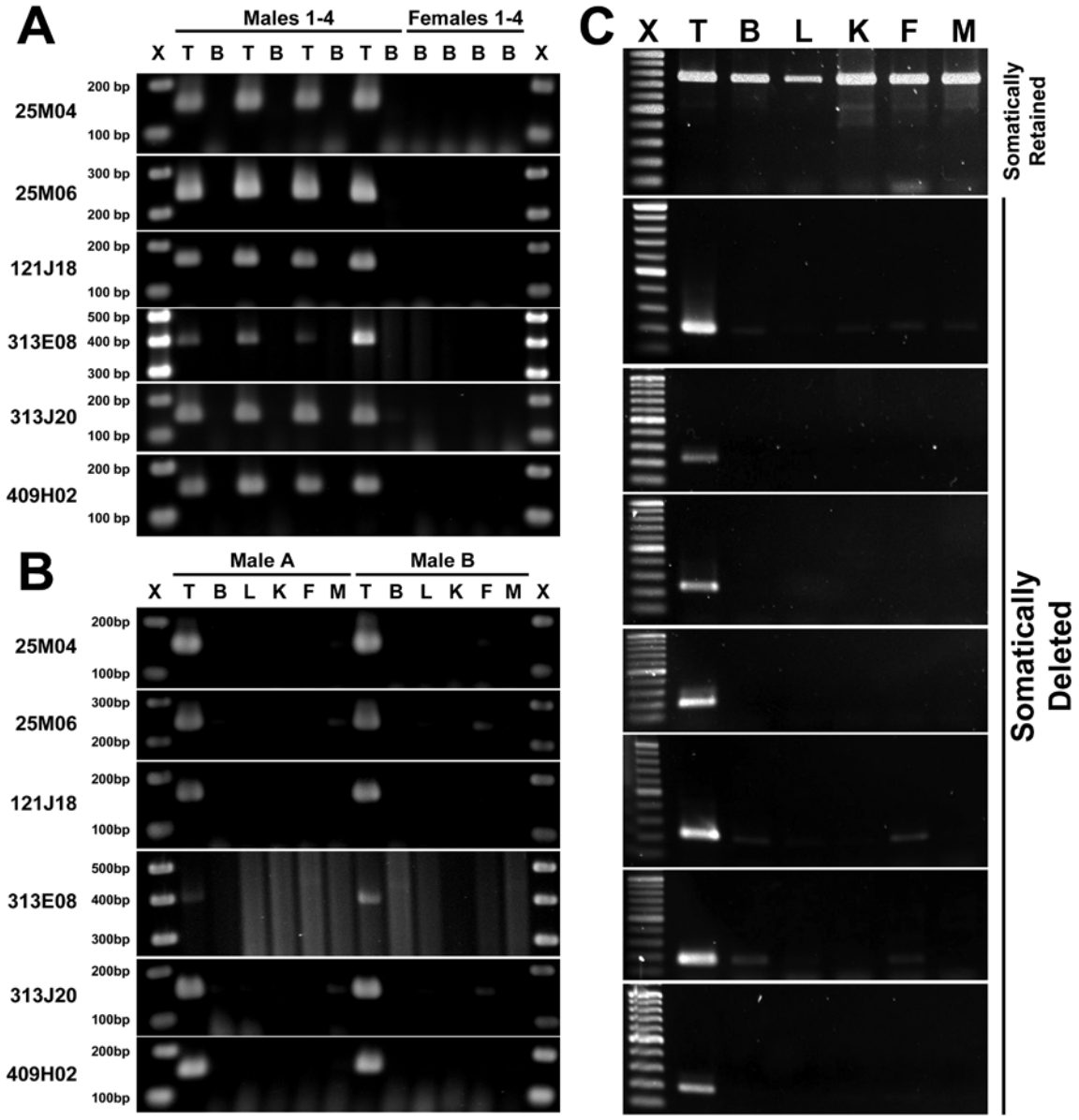

Figure S3. PCR Assays Used to Validate Predicted Germline-Specific Regions, Related to Figure 3

(A) PCR amplifications were performed using primers designed for the target region and 12 different genomic DNA samples isolated from eight different adult individuals: testes and blood from four males, and blood from four females. Results for validated germline-specific regions from arrayCGH experiments are shown here.

(B) PCR amplifications were performed using primers designed for the target region and 12 genomic DNA samples (isolated from six tissues of two different adult individuals: testes, blood, liver, kidney, fin and muscle). Results for validated germline-specific regions from arrayCGH experiments are shown here.

(C) PCR amplifications were performed using primers designed for the target region and genomic DNA isolated from testes, blood, liver, kidney, fin and muscle. Results for validated germline-specific regions from 454 sequence surveys are shown here. Somatically deleted regions are presented in the same order as shown in Figure 3B. Tissue sources of genomic DNA templates are designated as follows: T – testes; B – blood; L – liver; K – kidney; F – fin; M – muscle; X = 100 bp DNA ladder. The same quantity of DNA was used in each PCR.

Table S3. Summary of arrayCGH Experiments Performed for This Study, Related to Figure 1

Label			
Cy3[a]	**Cy5[a]**	**Hybridization ID**	**GEO Accession Number**
sperm (Pm_A)	blood (Pm_A)	305545CB05091	GSM586211
blood (Pm_A)	sperm (Pm_A)	305556CB05092	GSM586212
liver (Pm_21)	blood (Pm_21)	305545CB15091	GSM586213
blood (Pm_21)	liver (Pm_21)	305556CB15092	GSM586214
liver (Pm_4)	blood (Pm_4)	306998CB15093	GSM586215
blood (Pm_4)	liver (Pm_4)	307005CB15094	GSM586216

[a] – labeled samples were genomic DNA isolated from the respective tissues. The source individual for these DNA isolates is provided in parentheses.

Table S4. PCR Oligonucleotides Used in This Study, Related to Figure 3

Primer Name	Sequence	Purpose
25M04.b.F	TGTGACTTTCCTGGCCTGA	PCR amplification and RACE
25M04.b.R	TTTGTCTAACTGGGACTAGAAATGC	PCR amplification and RACE
25M04_nest.F	TTATTGGTGCACAGCGTCTC	Nested primer for RACE
25M04_nest.R	GAGACGCTGTGCACCAATAA	Nested primer for RACE
25M04_insitu.F	TTATTGGTGCACAGCGTCTC	PCR amplification of *in situ* probe
25M04_insitu.R	CGCTGCAGTAATGCAATGTT	PCR amplification of *in situ* probe
25M06.b.F	CAGAAACCCACGGACTCATC	PCR amplification
25M06.b.R	TCACGTGACGTCATTTAAGG	PCR amplification
121J18.g.F	GATGGAGTGGGAGACCAAGA	PCR amplification and RACE
121J18.g.R	CTCAACAGAGCGTACGACCA	PCR amplification and RACE
121J18_nest.F	AGAACGCAAGTAAGTTGTGG	Nested primer for RACE
121J18_nest.R	CCCCACAACTTACTTGCGTT	Nested primer for RACE
313E08b.g1.F	TCACCTACCCACCCAACCTA	PCR amplification
313E08b.g1.R	TATGAGTGGGTTTGCGAGTG	PCR amplification
313J20.g.F	TCTCTGCAACCTGCAACAAC	PCR amplification and RACE
313J20.g.R	AGGCACCTTGTCACCAAATC	PCR amplification and RACE
313J20_nest.F	AGAATTTGGTTTGCGTGGTC	Nested primer for RACE
313J20_nest.R	GACCACGCAAACCAAATTCT	Nested primer for RACE
409H02.b.F	AGCCGTATTGCGAATGAGAT	PCR amplification and RACE
409H02.b.R	AGTGCGGCCACTCTTGTAGT	PCR amplification and RACE
409H02_nest.F	AAGGAATATCTGAAGATGTTTTTCACCAA	Nested primer for RACE
409H02_nest.R	TTGGTGAAAAACATCTTCAGATATTCCTT	Nested primer for RACE

454_validation_1.F	ACGCTGAAGCAGACTCCAAT	PCR amplification
454_validation_1.R	CGACCTCTACGAGGATGAGC	PCR amplification
454_validation_2.F	ATGAAGAGGCGGTGATTGTC	PCR amplification
454_validation_2.R	GCGACCTCTACGAGGATGAG	PCR amplification
454_validation_3.F	GGATCTTCCGATGCTTTTGT	PCR amplification
454_validation_3.R	GGGAAGGCATTTTCATGGTT	PCR amplification
454_validation_4.F	CGTTTCCACGTTCATCTCTT	PCR amplification
454_validation_4.R	TGAAGTTGTTGTTTGCTGTGC	PCR amplification
454_validation_5.F	GCGATTGTCCACGCTAAAGT	PCR amplification
454_validation_5.R	GCGACCTCTACGAGGATGAG	PCR amplification
454_validation_6.F	TGGACCTGGAGGTATTCTGC	PCR amplification
454_validation_6.R	ACTGCATCTGCAACATCAGC	PCR amplification
454_validation_7.F	CTCCGTGTCGAGAGGCTTAC	PCR amplification
454_validation_7.R	GTTCGTGCACCACCAAGAC	PCR amplification
454_validation_8.F	GCGCAGGGCTTTAGAATACA	PCR amplification
454_validation_8.R	TCCTTGGAAAAAGCGGTATG	PCR amplification
454_validation_9.F	GACGATGAGTTGCTGCAGAG	PCR amplification
454_validation_9.R	GTCACGCTTGTGCGTGTGTA	PCR amplification
454_validation_10.F	TGTGGTTATGGTGGTGGAAA	PCR amplification
454_validation_10.R	GGTTACCGCGTGAGTAGAGC	PCR amplification
454_validation_11.F	CTCGGGGACACTCCAAATAA	PCR amplification
454_validation_11.R	CCAGACTCTGCACGACAAAA	PCR amplification
454_validation_12.F	CCCCGACTGGTTGAAAACTA	PCR amplification
454_validation_12.R	ACCTTGTCGCAAAGGACAAT	PCR amplification
454_validation_13.F	TTCCACACAACCGATGAAAA	PCR amplification
454_validation_13.R	GCGGTCGTTGTTGTTGTATG	PCR amplification
454_validation_14.F	GCTAATGCCCACGCTAACTC	PCR amplification
454_validation_14.R	ATTTGACTCGCTGCCACTTT	PCR amplification

Supplemental Experimental Procedures

Microarray Design, Processing and Analysis

The lamprey V1 NimbleGen array contains targets to a large number of non-redundant and repeat-masked sequences, including 752 sperm BAC ends each targeted by an average of 26 probes per read, the 57 kb Variable Lymphocyte Receptor genomic interval (2666 total probes) [38] and the 1000 largest lamprey contigs (V5 assembly) at an average inter-probe distance of 50 bp. The microarray also contains an extensive group of control probe sets that are targeted at computationally predicted "single copy" WGS reads with both high density (1091: avg 42 probes/read) and low density (4107: avg 8 probes/read) designs. Contigs and end-reads were repeat masked using all available vertebrate repetitive elements and an additional collection of repeats, which were identified among the 1000 largest contigs and a sample of 500,000 paired-end reads (WGS) using RepeatScout [39] and Tandem Repeat Finder [40]. Oligonucleotide arrays were designed by NimbleGen (Roche) using proprietary methods. Genomic target regions and array design details have been deposited in the NCBI gene expression omnibus (Table S3, http://www.ncbi.nlm.nih.gov/geo/).

Each experiment was performed in replicate with Cy3 and Cy5 dyes reversed between test and reference genomes and normalized relative fluorescence values (NimbleScan) were averaged between dye-swap replicates. Relative intensities for each ~500 bp interval were calculated using normalized and dye-swap-averaged intensities. The distribution of these values was compared to a normal distribution with the same sample size and dispersion as the observed data. Genomic regions yielding $\log_2$ ratios of hybridization signal intensity in excess of five standard deviations from the mean were considered as candidate sites of programmed deletion and were subjected to PCR validation. For this study, the cutoff of five standard deviations represented the point at which observed counts (of regions with excess hybridization to sperm DNA) began to exceed expected counts (Figure 1).

Tissue-Specific (Somatic) Variation

The possibility of genomic variation among somatic tissues was suggested by the observation of DNA breaks in nearly every cell type throughout lamprey embryonic development and slight differences in measured DNA content between different somatic cell types [1]. In an attempt to identify tissue-specific markers of programmed rearrangement, we employed the same arrayCGH platform described above to compare blood DNA from two animals with their respective liver DNA. These tissues were chosen because they present the largest measured difference in estimated DNA content (blood: 1C = 1.82 pg, liver: 1C = 1.96 pg) and they derive from distinct cell lineages (blood: mesoderm, liver: endoderm). Analysis of relative hybridization intensities revealed that most regions fell within the expected distribution for normalized data (Figure S1). A few regions showed some evidence of differentiation between blood and liver. However, the degree of differentiation was much smaller than that observed for germline versus soma comparisons, and validation assays (PCR and real-time PCR) yielded essentially no evidence for tissue-specific enrichment of candidate fragments (Figure S1). This last result may not be especially surprising given the small relative difference in measured DNA content and the fact that only 1% of the somatic genome and a far smaller percentage of potential somatically deleted sequences (i.e. sperm-derived sequences) are represented on this customized microarray. Our failure to identify tissue-specific markers does not rule out the possibility of sequence differentiation among tissues; rather, it underscores the need for more comprehensive sequence resources targeted at the lamprey germline and other specific tissues.

Validation and Characterization of Deleted Genes Detected by ArrayCGH and 454 Sequencing

Oligonucleotides were designed using Primer3 (Table S4) and used to prime PCR reactions that used germline and somatic DNA from several individuals as template. For changes involving candidate protein coding genes, we further tested the inference that these regions are transcribed by performing PCR amplifications of cDNA derived from germline (testes) RNA. cDNA were synthesized from testes RNA

using the SuperScript III First-Strand Synthesis System (Invitrogen) and manufacturer-specified reaction conditions. PCR assays were performed as above using either 1 ng DNA template, 1 μl cDNA or an equivalent amount of source RNA.

In order to test whether candidates were indeed germline-specific, we performed PCR validation assays using two different DNA panels. One panel consisted of genomic DNA from the germline (testes) and somatic tissue (blood) of four males and blood of four females. The second panel consisted of genomic DNA from several different tissues: testes, blood, liver, kidney, fin, and muscle, which were independently sampled from two males. Genomic DNA was isolated using standard phenol/chloroform extraction [33]. PCR experiments were performed using standard reaction conditions [1 ng DNA template, 50 ng each primer, 1.2 mM $MgCl_2$, 0.3 U Taq polymerase, 1x PCR buffer, 200 μM each of dATP, dCTP, dGTP, dTTP; thermal cycling at 94° C for 4 min; 33 cycles of 94° C for 15 s, 55-65° C annealing temperature for 15 s, 72° C for 15 s; and 72° C for 7 min]. For the six validated arrayCGH candidates, fragments of the expected size were amplified from the germline DNA of all six males that were used in this survey, but no amplification was observed from female blood DNA or DNA from any male somatic tissue. Primers designed from the remaining two arrayCGH candidates yielded either no or poor amplification despite two rounds of primer design. For sequences that showed a moderate difference in arrayCGH hybridization intensity between liver and blood, we performed additional validation via real-time PCR. Real-time PCR was performed on a 7900HT Fast Real-Time PCR System (Applied Biosystems), using the SYBR GreenER qPCR SuperMix, ~1 ng DNA and 50 ng of each primer. Thermal cycling conditions were 10 min initial denaturation at 95°C, followed by 99 cycles of 95°C for 15 s and 65°C for 20 s.

To test whether any of the validated losses of low-copy DNA might result in the loss of functional RNA or protein products, we assayed for transcription of all validated candidates identified by arrayCGH. Amplifications from juvenile and adult testes cDNA yielded fragments of the expected size for five regions. This is taken as evidence that the amplified portions of these sequences are transcribed into mRNA (Figure 1B, main text). Extension of these short fragments by RACE (rapid amplification of cDNA ends) revealed gene homologies for two (Heterogeneous Nuclear Ribonucleoprotein L and Rho GTPase Activating Protein 5) and indicated that the other three may represent long (>900 bp) transcripts that are potentially unique to the lamprey (there were no identifiable homologs in GenBank). Whole-embryo *in situ* hybridization was employed to assess the expression pattern of one of these, *25M04*, and showed conclusively that this unique germline-specific gene is likely to be important in the early development of lamprey germline cells. This gene is expressed specifically in primordial germline cells (PGCs) during the second to third week of development (Figure 2) and was subsequently found to be similar in sequence to putative KRAB domain zinc finger proteins that are also deleted via PGR (below). During this time, the PGCs are migrating dorsally from their early developmental origin (distributed within the yolky endoderm) and ultimately take up residence in the emerging gonadal anlagen [3-5] (Figures 2, S2). Thus, expression of somatically deleted genes is observed in the male mitotic and meiotic germline, and, for at least one gene (*25M04*), in the developing PGCs.

Supplemental References

38. Pancer, Z., Amemiya, C.T., Ehrhardt, G.R., Ceitlin, J., Gartland, G.L., and Cooper, M.D. (2004). Somatic diversification of variable lymphocyte receptors in the agnathan sea lamprey. Nature *430*, 174-180.

39. Price, A.L., Jones, N.C., and Pevzner, P.A. (2005). De novo identification of repeat families in large genomes. Bioinformatics *21 Suppl 1*, i351-358.

40. Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. *27*, 573-580.