

advances.sciencemag.org/cgi/content/full/7/1/eabd8180/DC1

Supplementary Materials for

Antigen receptor repertoires of one of the smallest known vertebrates

Orlando B. Giorgetti, Prashant Shingate, Connor P. O'Meara, Vydianathan Ravi, Nisha E. Pillai, Boon-Hui Tay, Aravind Prasad, Norimasa Iwanami, Heok Hui Tan, Michael Schorpp, Byrappa Venkatesh*, Thomas Boehm*

*Corresponding author. Email: boehm@ie-freiburg.mpg.de (T.B.); mcbbv@imcb.a-star.edu.sg (B.V.)

Published 1 January 2021, *Sci. Adv.* **7**, eabd8180 (2021)
DOI: 10.1126/sciadv.abd8180

This PDF file includes:

Supplemental Methods
Figs. S1 to S8
Tables S1 to S5
References

Supplemental Methods

Estimation of genome size and heterozygosity level

All Illumina reads from the male minifish were trimmed to remove adapter and poor quality sequences. These trimmed reads were used to generate a k-mer copy number (KCN) profile using the Jellyfish v2.2.6 program (56). We used several k-mer values ranging from 15 to 31. At a k-mer value of 17, the k-mer frequency distribution showed two distinct peaks (fig. S8). Hence genome size was estimated at k-mer=17 using the Lander-Waterman method (57) and was found to be 420 Mb. The heterozygosity level was estimated using the KCN distribution at k-mer=17 and GenomeScope programme (58) and was found to be 0.5%.

Genome assembly

Whole-genome assembly was performed using DISCOVAR de novo v52488 (59). The male and female minifish genome assemblies were generated using quality-filtered Illumina paired-end reads totaling 69 Gb (164× genome coverage) and 87 Gb (208× genome coverage), respectively. The genome assemblies generated contained several highly identical copies of contigs representing two copies of the alleles. An in-house Perl script was developed for removing such allelic duplicates whereby all contig sequences were compared against each other using the MegaBLAST module (60) with the following threshold values: word size = 100, E-value = 10^{-8} , coverage = 80% and identity = 90%. In the resulting non-redundant contig set, many contigs pairs were found to overlap. Such overlapping contigs with more than 92% identity, minimum 20% overhang with an overlap length of above 100 bp were joined using CAP3 programme (61).

Scaffolding using Chicago library. High molecular weight genomic DNA from male and female minifish was used to prepare Dovetail Chicago[®] libraries (Dovetail Genomics, Santa Cruz, CA) and approximately 142 million read pairs of length 101 bp were generated for each minifish. The contig level assemblies were scaffolded using the Dovetail Chicago library reads with the HiRise scaffolding programme (62).

Gap filling using super-reads

The trimmed Illumina paired-end reads from each minifish were assembled using MaSuRCA v3.2.4 (63) to generate super-reads. This program generated ~7 million super-reads with N50

length of 870 bp and ~10 million super-reads with N50 length of 780 bp for the male and female minifish, respectively. These super-reads were used to fill gaps in the scaffold-level assembly using the PBJelly program from PBsuite v15.8.24 (64). The final genome assembly statistics for male and female minifish are shown in Table S1.

RNA-seq

Total RNA was extracted from a male and a female minifish using the TRIzol reagent (Invitrogen, Carlsbad, USA), treated with DNase I (TaKaRa Bio Inc, Shiga, Japan) followed by purification using the RNeasy Mini Kit (QIAGEN, Hilden, Germany). An RNA-seq library was constructed for each minifish using the Ribo-Zero Gold reagent and ScriptSeq v2 Library Preparation Kit (Epicentre, Madison, USA). The library quality and quantity were analyzed on an Agilent 2100 Bioanalyzer. Sequencing was performed on an Illumina NextSeq 500. A total of 167 and 169 million paired-end reads of 150 bp length were generated for the male and female minifish, respectively. Sequences were quality filtered for low-quality bases and adapter content using Trimmomatic v2.2.30 (65). The filtered reads were assembled *de novo* using Trinity v2-2-26 (66), which generated approximately 0.8 and 1 million transcripts for the male and female minifish, respectively. In order to reduce redundancy in the transcriptome assembly, transcripts were clustered using CD-HIT v4.6.1 at 98% identity (67). The clustered transcripts were subjected to BLASTX search against 25,638 zebrafish protein sequences obtained from the Ensembl database (68) with an E-value cut-off of $10e-7$. Those proteins showing at least 80% of the query coverage were considered as full-length proteins. All transcripts as well as full-length protein sequences (4,416 for male minifish and 4,325 for female minifish) were used for genome annotation.

Evaluation of assembly completeness

The Benchmarking Universal Single-Copy Orthologs (BUSCO) version 3.0 (69) and 4,584 Actinopterygii gene set from OrthoDB v9 were used to evaluate the completeness of the male and female genome assemblies. BUSCO results for the two assemblies are shown in Table S1. These results show that approximately 5.5% of BUSCO genes are missing from these assemblies.

Repeat content prediction

Repeat sequences were predicted *de novo* in the male minifish genome assembly using RepeatModeller v1.0.10 (ref. (70)). Approximately 1,370 specific repeat families were identified including some unknown repeat classes. These classes were further annotated using TEclass v2.1.3 program (71). This *de novo* repeat library was combined with a known repeat library for 'Cypriniformes' which was obtained from RepBase.v.22 (72). These repeats were screened against zebrafish proteins using Blastx (E-value $10e-20$) to identify and remove any potential protein-coding sequences. The final set of repeat sequences contained 2,868 repeat elements. This repeat library and RepeatMasker v4.0.7 (73) program were used to mask both genome assemblies and the results are reported in Table S2. The overall repeat content in the male and female minifish genome assemblies is approximately 27%.

Genome annotation

For both genomes, evidence-based gene prediction was performed followed by *ab initio* gene prediction using the MAKER pipeline v2.31.963 (74). For the evidence-based annotation step, we used a protein dataset containing proteins from *Lepisosteus oculatus*, *Ictalurus punctatus*, *Oreochromis niloticus*, *Xiphophorus maculatus*, *Danio rerio* and *Oryzias latipes* that were downloaded from NCBI. This dataset was filtered for proteins with keywords ‘low quality proteins’ and ‘partial’ in their descriptions. The final dataset contained around 124,000 proteins. We also used CD-HIT clustered transcripts and full-length protein datasets prepared for both the genomes during this step. The gene models obtained from the MAKER run were used to train AUGUSTUS v3.2.3 program (75). The hint files generated during evidence-based maker annotation were used as input to facilitate the gene prediction process and to calculate the Annotation Edit Distance (AED) score. Predicted protein sequences that had an AED score ≤ 0.6 or had no similarity to any protein in the NCBI-NR protein database (BLASTP; E-value: $10e-7$) or had similarity to “low-complexity” proteins were removed from the final set. We predicted 20,013 and 18,003 protein-coding genes in the male and female minifish genome assemblies, respectively.

Identification of immune gene elements

Homology searches were carried out using the BLAST suite of algorithms (<https://blast.ncbi.nlm.nih.gov/Blast.cgi>) using nucleotide and protein sequences of three gene families (immunoglobulin, T cell receptor, and MHC) of zebrafish, a closely related cyprinid species; the positions of individual elements on the two genome assemblies were recorded. Additional elements were identified using the sequences of minifish transcripts for these three gene families and incorporated into the final dictionaries of V, D, and J elements. Homologs of other immune-related genes were identified in transcriptomes and genome assemblies using the zebrafish sequences as baits. The deduced protein sequences of these genes are listed below.

```
>tdt
MFQALRRRQAEGTPTRVGSEVKFSHVTLYLVERKMGTSRRNFLSDLARSKGFYVDNTLSGRVTHVVSSEGNSDLELWKW
LEDKGFREKLGQKVLIDINWFTESMSAGRPVNVETRHILRNPCGETKPCMLSSKPNPEFFVSPYACKRRTPLQNHKNLTL
DALEVLQAQNSEFIGNTGPCLAFRLRAVSVLKSLPTALSRLEDAQYLPCLGKHSNAIEEIFFKFGASSKVEEILNDERYLT
LKLFNVSFVGVPKTAESWYCOGLRTFEHVLTEPRVKLRNMOMAGFMFYDDITEPVSRTEAGVVQMMVEEAVGQINSTAT
VAITGGFRRGKDCGHDVDFLIKTQAGQEDGLPAIIQTFRQNILLYSDYQPSLTLGGKQLPSRRFEAMDHFHKCFLLIV
KVKRVWTSPEGSDGAANRDWKAIRVDLVAPPLECFAYALLGWTGSTMFDRDLRRFARLERGKLLDNHALYDKATNSFIP
ALTEEDIFNHLGLEFIEPWQRNA
```

```
>cd4_1
MDRTRMFWIVAFQVKAEDPTVIYAQVGGTVILPRVTKLVEKNLYVNWKSNTSKLEISRNPQSDNIKKEHASHTSN
FSLQLSPVKESDFGKWWCEMHILNTKYKEEYILRRVTVPSSAVMVGNSVTLKCDVEKSSVSVAVTWKWPPVNSHCSDX
NLNTDAGKLEKXSVSRCHSGEWTCEVKYSQRKAEAKTTVSVIDLSPNPQDPMYVSESSPLVTIPCSLTSKTPWSVLEH
GLQGGWSFTPTDQPDQKAVTLLSLALGPAVSWNVSAGADADVQKLLKDQDLSMERAPSRKIRGTYTCSLTFTSKTL
SSLVNVEVLHVSSSSNRVQEGVNLTCSLGRPLPADVQLKWKCPSCSSSEAKPTGVKTIDLPTVHDNGKWTCELWKNEEM
LTSAYLQINIGKPPNVVGLWWSVGIGCGSVVIVLLLVIXVMVIRRRKQMMMYRRHKTKFCCENSQQKKGfyKT
```

```
>cd4_2
MLAVKLLVFLAVCVIPGESTVLVYQAGADASLVCNVPENSDMEWRFNVLIFKVNGKTGKIKGTSRLSNIKSKLAG
STLKIFKVNKDDSGEYSCRTNDKMILRVVSFVKPHNPVLKSSDAELHCEISGDPEAKVEWLKPPNGIKQSTTNQVLH
LKSVRAEDQGLWCKVNDLTMTIQLTLIDLTTKNATVPLGGDIQLPCSLTGSSHRVIGGKWSAHNSAIAFPTLDDTGG
LRWKGQSSSKVAFSSEQLNTDFSITLKNVQKQDAGIYVCSVEFDTVVVKAEMNLKVLRSSSSDYNEQSIAGTTKTPVWV
SNFFKDYSRNTLGLPLWAWVALGVGSVVLMLVLTGLIILVILRRNRVVKVGRRTMRKPLTDQDYCQCHKY>
```

```
>cd8a
```

MPHFFSFFLLFCAACVTFSLGSARVRAGGEASVRCDPSLSSSVFWFRITRAGPEYLLTVRPGKDPVPSDQONIQFISG
EKSVKILGFRADTDAGVYSCFSINNNQLKFGEATEVHAEAPSSAPVTQKKTTSLPPTTTPCQCKTQKAALLKCESWIF
YSLVSGCLFLFLLLIFTIITCNRLRTRRCPPHYQRRRPEKLPDGRF

>cd8b

<CEGVVQDIYPGINSTHTLTCDSDSTCQDVFWFRLPHSQSLQFLAFANSNGGRVQRAEVATRFTNTSSAGVRLTSFSLH
LMRLQEEDSALYACFLRASHALYGFRVSVGGQRSIARTGRGCGCRPGVGVVGCPRVLWWGGGAVLALTLVLLATLFYF
SRLPRKCRHQFLKTNQLR

>cd79a

MWTEQTSACENVSSGRKTEIILKADQPFRRVAVFREVITRCCYECPAKPRVTWIVNTITSNGTTQLELVPLSDELEASE
ITQGDVSCSQLVFKSVRLWHMGLYRCFLNHTQYVSTISHGTFLQVYEPLEKTLNLSENVKNSIITAEAVLLLLCLLPFG
TVLLCKSKRLNELQKRKEREENIYEGLNLDDCNSAYHPIRRSNMQGTQYQDVASCGERIQLEKP

>cd79b

MLHLLMRCSVLALLHLSAGVQLYQKPRFVGVPRGRSVTIYCVWQKSLPAYVEWSKARTNTDQKQTLINPRMTLLNKIT
NASITIRRVIMEDSGIYFCRMNGTEGPGTELQVSRHSDPQSVLKRVRKDVIIVFQGILLILCLVVLVRFKSLDKKEE
IVYEPEDDHTYEGLNVDQCGCDLYEDISAFAMSPDQACEVEYPNQE

>aicda

MISKLDSVLMTQKKFIFHYKNMRWARGRHETYLFCVVKRRTGPDSLSFDFGHLRNRTGCHVELLFLRYLGVLCPLGLGS
SDGERLCYSVTWFCWSWSPCFKCAQHLARFLEDTTNLRLRIFVSRLYFCDDDESDVEREGLRHLKSAGVTTITVMAYKDYFY
CWKTFVARRQRSFKAWEGLQENSURLGRKLNQILQASETEDLGESFALLGL

>foxp3_version_a

<PTLKEEPDSSSLESSEVVVEWFNTFALQLGRPASDSLSSRHLLHLVLQESGAGQFRPSVLRSTALRLDAKDSRSST
INRKDKGHNROIPIGQNDGINQSRKPFNPISHPNPVQFEGQSCLCVNGQCCWPGCDKLEGGQDFSSHLNRDHSPNDRTI
AQCRQLQRDLVRHMESQLSQEKQRLSAMQLHLQLFHHLASHAGASASCCRPLALNMPLWEEPDRAGRAREALAPHKHGH
MPRPQIMPDLISSVEYKHSNVRPPYTYAFLIRWSILDSPKQQTLLNDIYNWFTLMFYFRHSTPTWKNVRHNLHLK
CFVRVDGRTGAVWTVDEEEFQKRKGQKINRDTFKWMTFVAPSPFHMTSNESVKSDDSH

>foxp3_version_b

<RARTEMDKAQTRVPQLRPSVLRKGNQFPQDAQTDQGTGIVQLSSDQHKDHFVTHAAASCHQSNLEQTEKHQYGVPG
LCVEGQCRWPGCPKSKEVFREYIQFLRHLVTDHCHGDRSLAQLKMQRDQVQYLEFQLMVERQKFKAMQLHLTSIKPNT
PAYCMEEPEHLKDIVPPAAAFQKSRDAHDSEKVTADALAQRNWHISTSPVIPGIIPSFYKYTNIRPPFTYASMIWAI
LESTEQTLTNEIYHWFAMHFFYFRHNTATWKNVRHNLHLKCFVRVEGKGSVWTVDEEEYHKKRGQKQFQDQIGW
MTPFHVFPPALQGETLQM

>foxn1

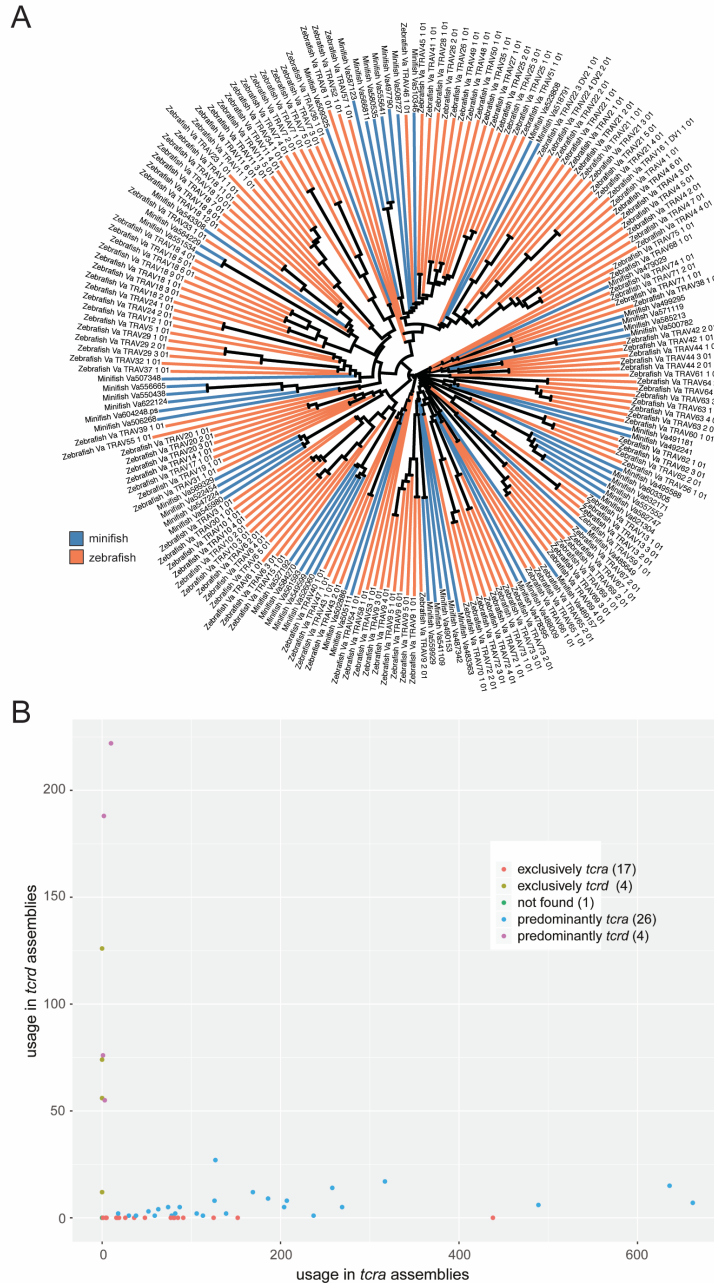
MSSEVAGLAFLSLSSGRSSTPSPEQPGICCFKMPESKQRMVLVGEALAPCSQKESIVGEKSGVCERFRHSVDGSLGKQ
DFSLPESSHYPYRRQYSEGSIPVEPFSCVTTAEGVEEQSPWTPLCNRETSSFMGSQQPYDELETESEASGYTTFN
HQTYNSPLQQQLFSSRGINNVSIFYNQSLSSQTSPPSSAQTLYPKPVYSYSILIFLALRNSKTGSLPVSEIYSFMTEHF
PYFKTAPDGWKNVSRHNLNLNCFEKIENKNGSSSRKGLWALNPAKVEKMQEELHKWRRKDPLTVRRSMARPEELERL
LGERPEKVKSFGAHFLSSSHNHSQSSRIGLHPIYGOQSVQDTSVQHOKPLFNCLPSQNVLQPPPYSDPSAFSYSPIS
QLPCTGHPSSPGPACLDSPPLAHTPPNYSSTLQAGHGTAGSIQLLMEGEISNDIDALNPSTDLQLNGLWEALRNDL
TPDSLIVLEMPAPPQQGLGVYSLSAGEMETDDGVQGSMEYLTGPRTHMFSNVNLAASLLSSSGNTPIPL

>foxn4

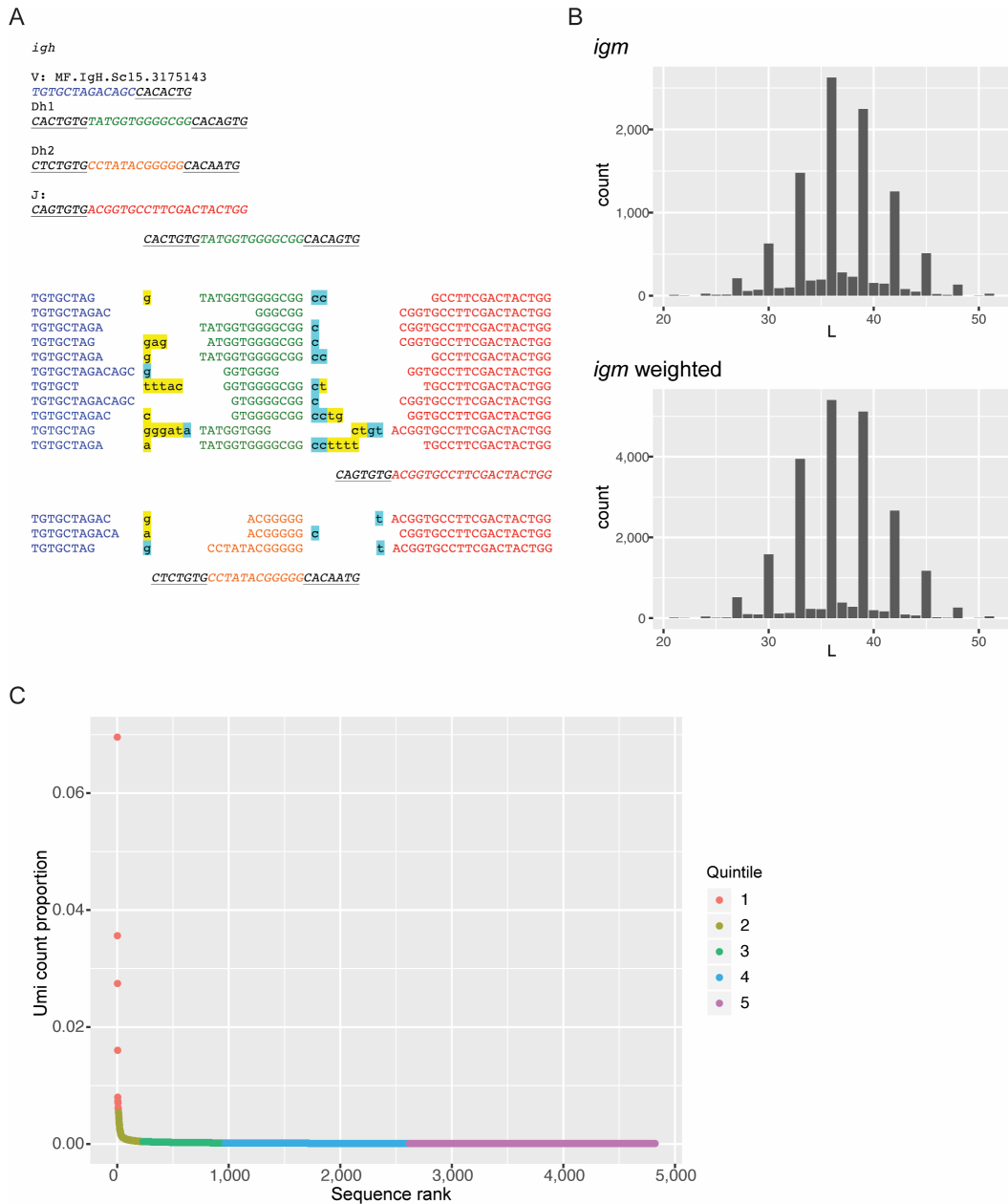
MIESGITSRMSGIQENPGQTQHGSALDYRLLTDPSQLKDDLPGDLQSLSWLTSVDVPRLQQMRAGQSDFSSSAQTSIM
ERQTGPMNNMAAVAGAASSLHLQSEMQHSLAISSMPQFSPGFPCAASMFQTTTPQQVLFTFTQTNQQCSQSGLYGNYTSP
NLFPQPRLTAHNQELQPKTFPKPIYSYSLIAMALKNSKSGSLPVSEIYSFMKEHFPYFKTAPDGWKNVSRHNLNLNKC
FEKVENKMSGSSRKGLWALNPAKIDKMEEMQKWKRDLPAIRRSMANPDELDKLITDRPESCROKAADVVISRLSS
CPAPQMOPVVTLSQLCLPMHQLMQIQSQSRPAPVSPAAAQTPPLHSALHHPAKQHGPDFYAMHSEAHSEVDALDPSI
MDFSWQGNLWEDMKDDSFNLEALGTLSNSPLRLSDCLDGTGSAASVPGAYPDLYSSYAAVEALTHPHYISTQGGTKPIIL
L

>tlx1

MDLMGAHLHQNHADTISFGIDQILSNAEQGSCMISNPRMQDLDYGLGCIVSTAYNTMTGNYNVNNSTGYNGNSCNVATL
GGSYNMNVGTNVNGNCLNSSGVIRVPAHRPLSNVHSSIPTNSATVPGMGSMGTINNLTGLTFPWMESNRRYTKDRFTVA
LSPFTVTRRIGHPYQNRTPPKKKKPRTSFTRLQICELEKRFHRQKYLASAERAALAKALKMTDAQVKTWFQNRRTKWRR
QTAEEREAEERQANRILMQLOQEAQKSMNQPVTPDPICLHNSSLFALQNLQPWTTAKISSVPNTD



Supplementary Fig. 1. Diversity and usage of Va/d elements. (A) Phylogenetic tree of variable gene elements at the nucleotide level of zebrafish (31) and minifish. V genes of zebrafish and minifish cluster together, indicating that – when compared to zebrafish – all V families are affected by gene loss in minifish. **(B)** Usage of individual Va/d elements in *tcra* and *tcrd* assemblies. The color code represents the categorization indicated by the key.



Supplementary Fig. 2. Structural characteristics of *igm* assemblies. (A) Examples of partial nucleotide sequences of CDR3 regions with individual elements indicated by font color; the sequences of the genomic precursors of the assemblies are shown at the top, with heptamer sequences of their corresponding recombination signal sequences are underlined. N nucleotides are highlighted with yellow shading, P-nucleotides in blue shading. (B) Length (L, in nucleotide) distribution of CDR3 regions; clonotypes (top panel) and adjusted for frequency (bottom panel). The CDR3 region is operationally defined as the sequences occurring between and including the characteristic C-terminal cysteine of V elements and the characteristic tryptophan residue in J region sequences. (C) Clonal distributions of clonotypes from a single individual (fish #5) represented in quintiles. Note that the skewed representation of clonotypes in the repertoire may be influenced by the presence of plasma cells, which are known to produce large quantities of mRNA.

tcra

V: MF_Vas.5\$Va561593
TGCGCTCTGAGGCCCACAGTG
J: MF.J.a.11
TAGTGTGTGACTACTACTGGATTCAATAAGATTATATT

TGCGCTCTGAGGCCCACAGTG

TGCGCTCTGAGGCC **ggc** GACTACTACTGGATTCAATAAGATTATATT
TGCGCTCTG **cagcc** GACTACTACTGGATTCAATAAGATTATATT
TGCGCTCTGAGGC CTACTACTGGATTCAATAAGATTATATT
TGCGCTC TGACTACTACTGGATTCAATAAGATTATATT
TGCGCTCTGAGGC CTGGATTCAATAAGATTATATT
TGCGCTCTGAG GACTACTACTGGATTCAATAAGATTATATT
TGCGCTCTGAG GACTACTACTGGATTCAATAAGATTATATT

TAGTGTGTGACTACTACTGGATTCAATAAGATTATATT

tcrg

V: MF.Vg_1201219
TGCGCGATGTGGAGCGGAGACACACATTGT
J: MF.J.g.1
GACTGTGGTAGACAAGAAAGTCTTC

TGCGCGATGTGGAGCGGAGACACACATTG

TGCGCGATGTGGAGCGG AGACAAGAAAGTCTTC
TGCGCGATGTGGAGCGGAGAC GTAGACAAGAAAGTCTTC
TGCGCGATGTGGAGCGGAGAC **gg** GAAAGTCTTC
TGCGCGATGTGGAGCGGAGAC **ttac** GTAGACAAGAAAGTCTTC
TGCGCGATGTGGAGCGGAG **tctac** GTAGACAAGAAAGTCTTC
TGCGCGATGTGGAGCGGAGACA **g** GAAAGTCTTC
TGCGCGATGTGGAGCGGAGACA **tgag** GAAAGTCTTC
TGCGCGATGTGGAGCGGAG **t** GTAGACAAGAAAGTCTTC
TGCGCGATGTGGAGCGGAGACA **t** GAAAGTCTTC
TGCGCGATGTGGAGCGGAGAC TAGACAAGAAAGTCTTC
TGCGCGATGTGGAGCGGAGAC GAAAGTCTTC
TGCGCGATGTGGAGCGG **ggt** CAAGAAAGTCTTC
TGCGCGATGTGGAGCGGAGAC **tg** GAAAGTCTTC
TGCGCGATGTGGAGCGGAGAC **g** AGAAAGTCTTC
TGCGCGATGTGGAGCGGAGACA **tagg** CAAGAAAGTCTTC
TGCGCGATGTGGAGCGGAGAC **gg** AGACAAGAAAGTCTTC
TGCGCGATGTGGAGCGGAGAC GTAGACAAGAAAGTCTTC

GACTGTGGTAGACAAGAAAGTCTTC

tcrd

V:MF_Vas.5\$Va555641
TGCGCTCTAAGTTTTGCAGCACAGTG
Dd1
CGTTGTGGATTGGGGTACCACAGTG
Dd2
TCGTGTGGATACGTTATTACCACAGTG
J: MF.J.d.1
TGAAGTGGAGTACCCCTAATCTTC

TGCGCTCTAAGTTTTGCAGCACAGTG

CGTTGTGGATTGGGGTACCACAGTG

TGCGCTCTAAGTTT **cagatacttggg** TGGGGTAC **ACGTTATTAC** **a** AGTACCCCTAATCTTC
TGCGCTCTAAGTTTGC **ttgggtc** GATTGGGG **a** GAGTACCCCTAATCTTC

TCGTGTGGATACGTTATTACCACAGTG
GTCATGTGACTGAATGAAGTGGAGTACCCCTAATCTTC

tcrb

V: MF.Vb_782821
TGTGCTGCTTATTACCACGCCG
Db
CAGTGTGGGGACAGGGGCCACGGTG
J: MF.J.b.793306
CAGTGTGAATAACTACAACCCCTGCTTATTTT

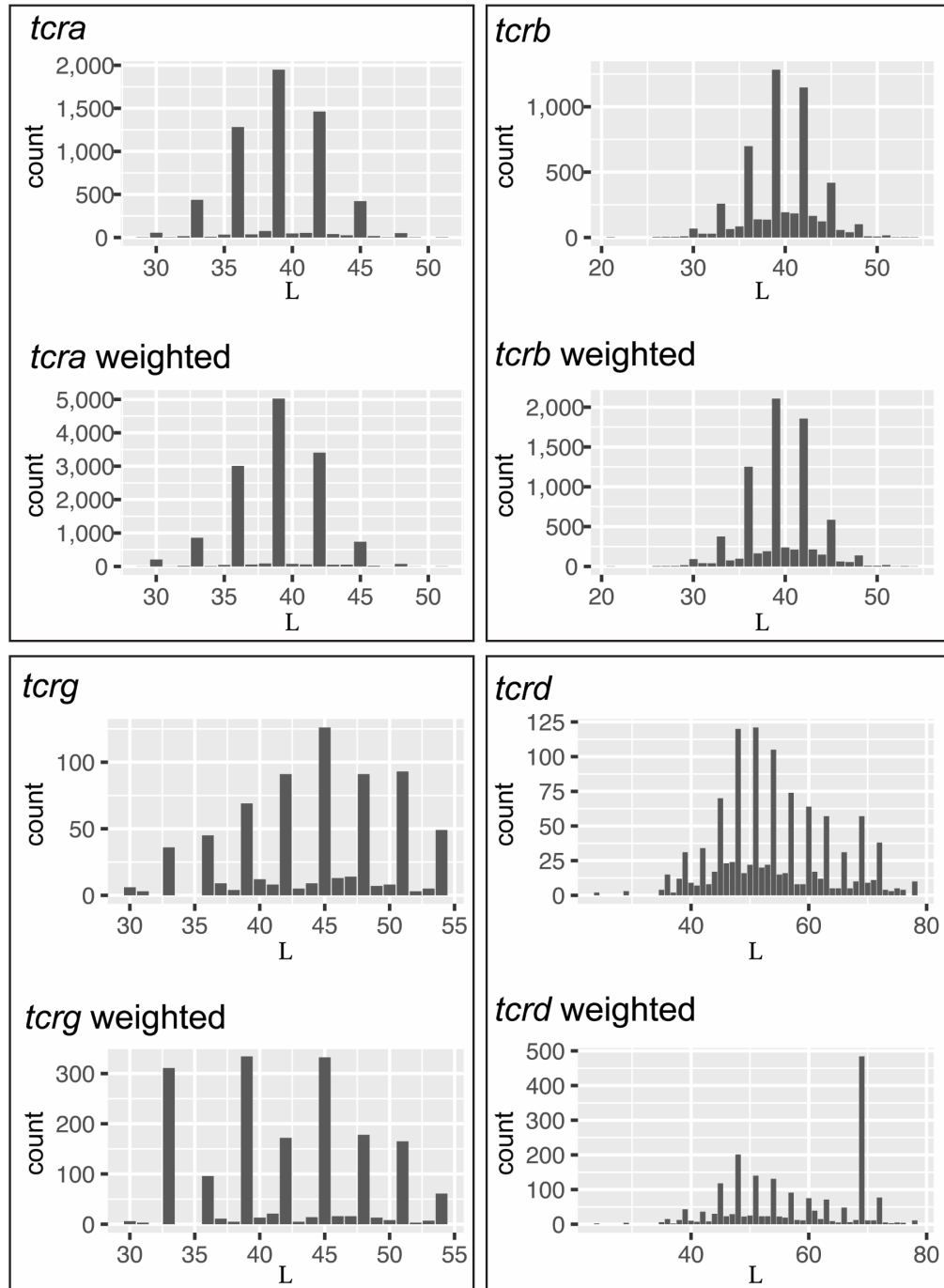
TGTGCTGCTTATTACCACGCCG

CAGTGTGAATAACTACAACCCCTGCTTATTTT

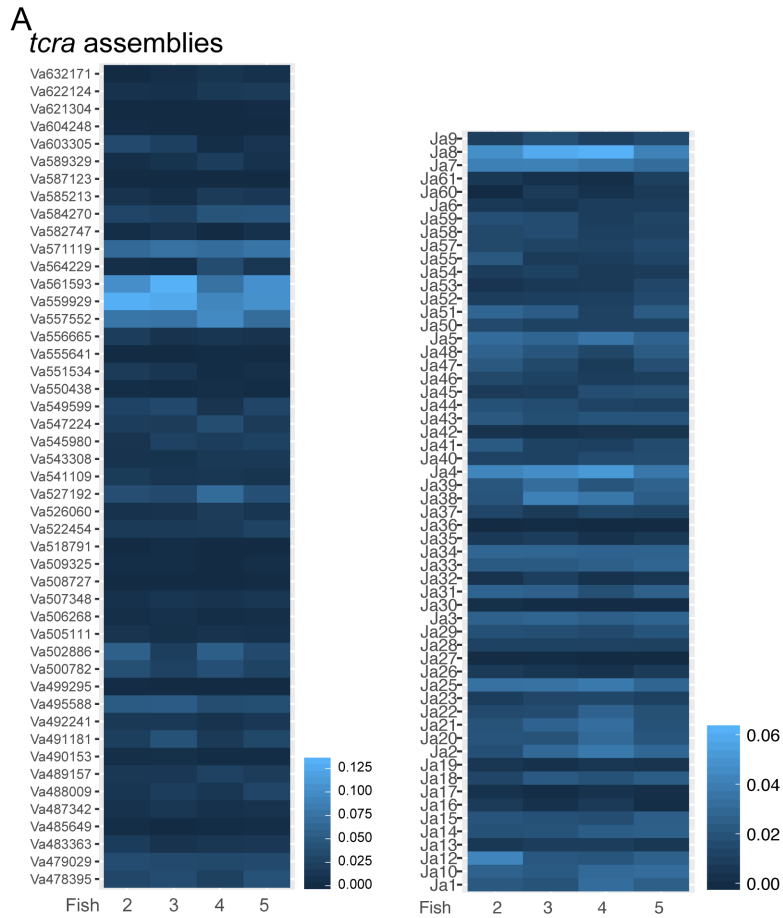
TGTGCTGCT GGACAGGGGGC TAATAACAACCCCTGCTTATTTT
TGTGCTGCTTAT **g** GGGGGC ACTACAACCCCTGCTTATTTT
TGTGCTGCTT **t** ACAGGG **tac** CTACAACCCCTGCTTATTTT
TGTGCTGCT **aa** GACAGGG ACTACAACCCCTGCTTATTTT
TGTGCTGCTTATTA **a** GACAGGGGGC **tg** ACTACAACCCCTGCTTATTTT
TGTGCTGCTTAT **g** AGGGGGC **gag** CTACAACCCCTGCTTATTTT
TGTGCTGCTTA GACAGGG AACTACAACCCCTGCTTATTTT
TGTGCTGCTTATTAC GGGGGC **tg** CAACCCCTGCTTATTTT
TGTGCTGCTTAT GACAGGG ACTACAACCCCTGCTTATTTT
TGTGCTGCTTATTAC **t** GGACAGGG TACAACCCCTGCTTATTTT
TGTGCTGCTTA GGACAGGG ACTACAACCCCTGCTTATTTT
TGTGCTGCTTATTAC GGGACAGGGGGC **g** ACTACAACCCCTGCTTATTTT
TGTGCTGCTTATAAC **a** GGA **gtcg** AACCCCTGCTTATTTT
TGTGCTGCTTATTAC **gtc** AGGGGGC **ctt** ACTACAACCCCTGCTTATTTT

CAGTGTGGGGACAGGGGCCACGGTG

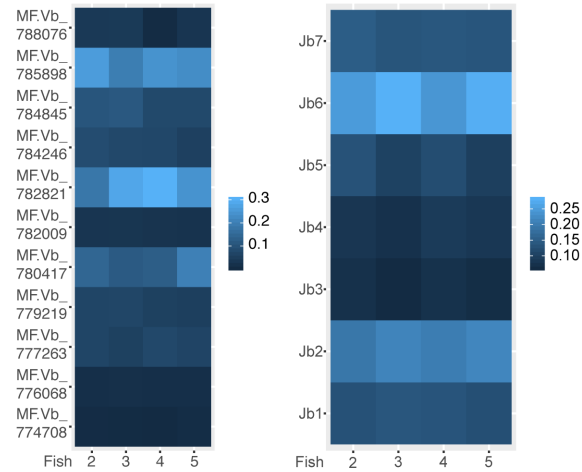
Supplementary Fig. 3. Structural characteristics of *tcr* assemblies. Examples of partial nucleotide sequences of CDR3 regions of the four types of *tcr* assemblies with individual elements indicated by font color; the sequences of the genomic precursors of the assemblies are shown at the top, with heptamer sequences of their corresponding recombination signal sequences are underlined. N nucleotides are highlighted with yellow shading, P-nucleotides in blue shading. Note that most *tcra* assemblies do not show evidence of N-region addition; as is the case for *tcrg*, the limited combinatorial diversity of the *tcrd* chain is greatly expanded by N region-mediated diversification of the CDR3 regions; this is particularly true for assemblies with two Dd elements and the corresponding three N regions.



Supplementary Fig. 4. Size distribution of CDR3 regions of *tcr* assemblies. Length (L, in nucleotide) distribution of CDR3 regions of the four types of *tcr* assemblies; clonotypes (top panels) and adjusted for frequency (bottom panels). Note that the weighted distributions of *tcrγ* and, in particular, of *tcrδ* are severely skewed. In *tcr* assemblies, the CDR3 region is operationally defined as the sequences occurring between and including the characteristic C-terminal cysteine of V elements and the characteristic phenylalanine residue in J region sequences.



B
tcrβ assemblies



C

	MI	JI
<i>tcrα</i>	0.3913	9.9341
<i>tcrβ</i>	0.0132	5.5368
<i>tcrγ</i>	0.0044	1.9534
<i>tcrδ</i>	0.1029	4.3151
<i>igh</i>	0.0051	4.8047

Supplementary Fig. 5. Usage of variable gene elements in *tcrα* and *tcrβ* assemblies. (A, B) Heatmaps illustrating the usage of different variable and joining elements (rows) for *tcrα* (A) and *tcrβ* (B) assemblies compared across the four individual fish (columns). The scale represents the fraction of UMI counts. (C) Mutual information (MI) and joint information (JI) for V-J combinations is given in bits.

A

Gene	Transcripts		Genomic loci		Comments
	male	female	male	female	
<i>mhc1_z</i>	1	1	1	1	identical sequences
<i>mhc1_u-like</i>	7	3	7	6	3-4 genes arranged in tandem; one <i>mhc2_b</i> - <i>mhc2_a</i> - <i>mhc1_u</i> array
<i>mhc2_a</i>	6	5	6	5	three (<i>mhc2_b</i> - <i>mhc2_a</i>) tandem arrays; one <i>mhc2_b</i> - <i>mhc2_a</i> - <i>mhc1_u</i> array
<i>mhc2_b</i>	5	5	4	4	three (<i>mhc2_b</i> - <i>mhc2_a</i>) tandem arrays; one <i>mhc2_b</i> - <i>mhc2_a</i> - <i>mhc1_u</i> array

B

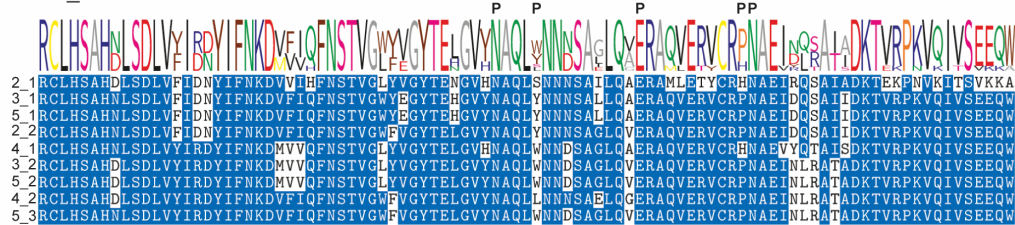
mhc1_u-like



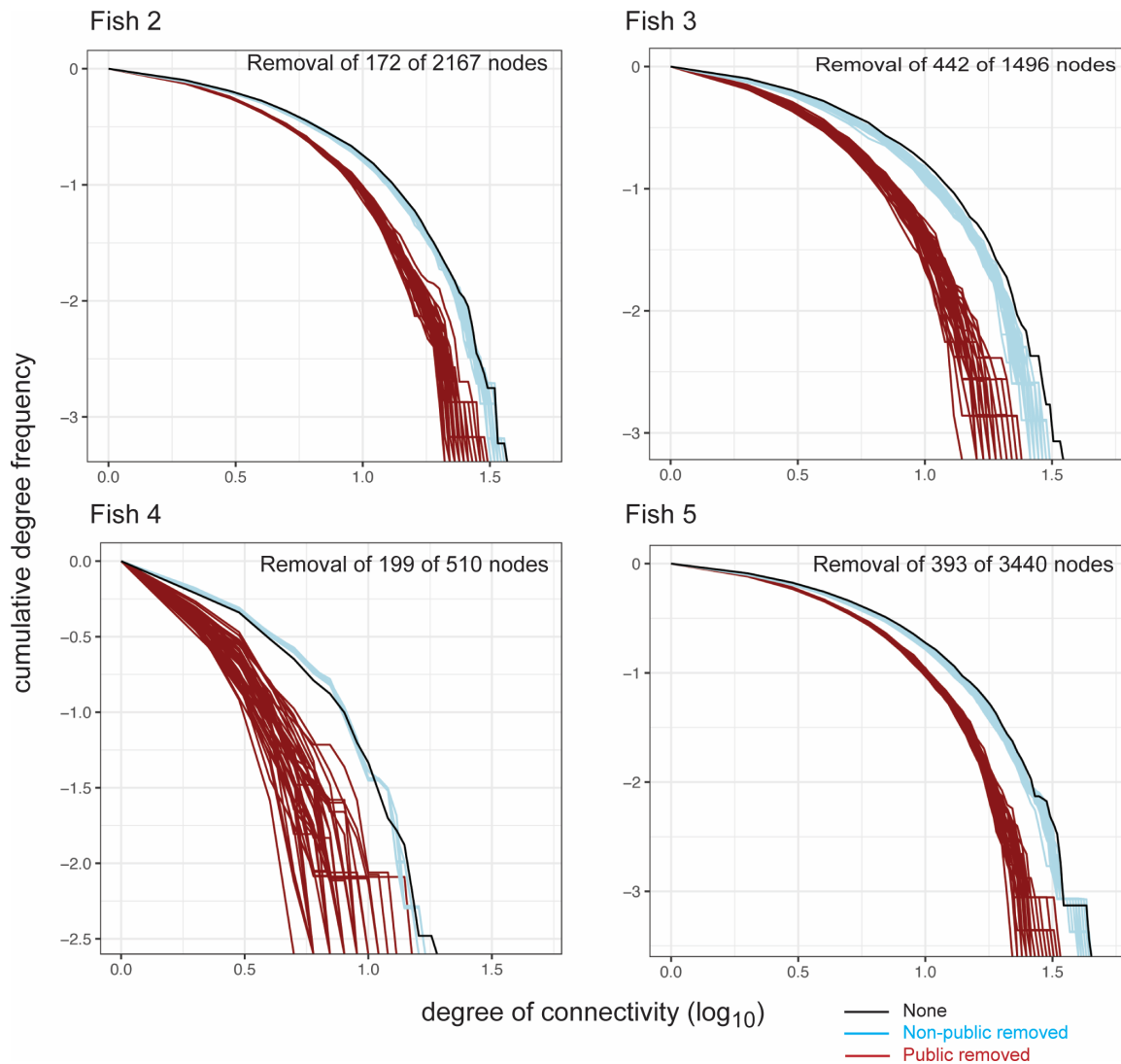
mhc2_a



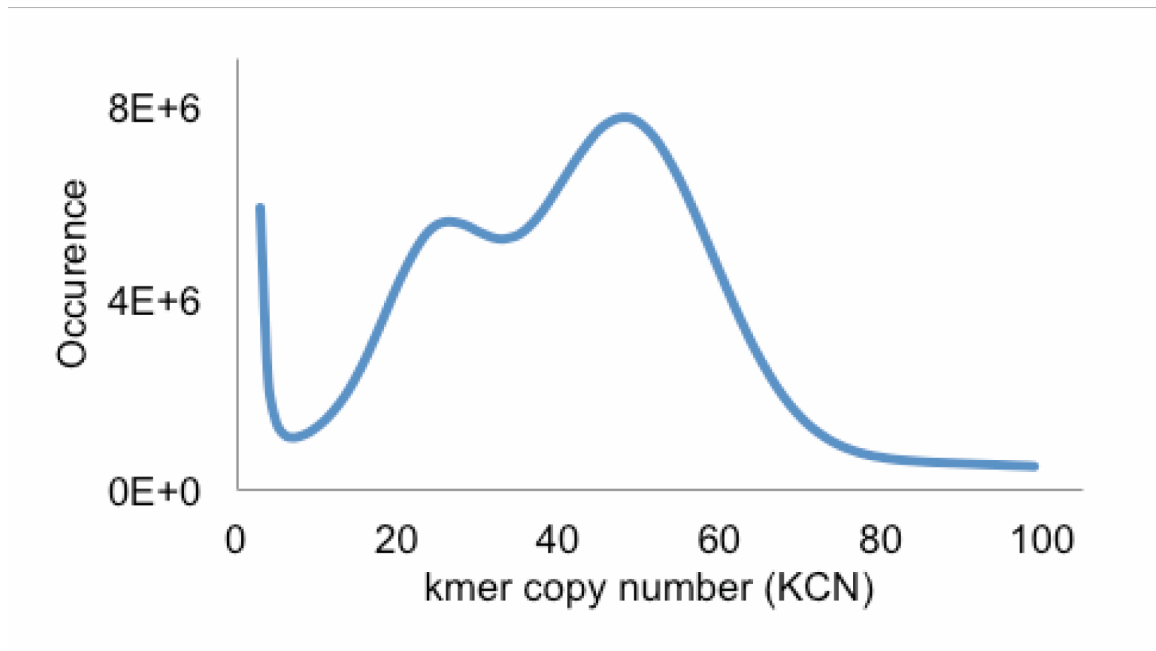
mhc2_b



Supplementary Fig. 6. Characterisation of *mhc* genes. (A) Number of different *mhc* sequences identified in transcriptomes and genome assemblies in four independent individuals. The numbers of all *mhc1*-like genes other than *mhc1z* are shown. (B) Examples of partial *mhc* sequences demonstrating that the four individuals examined with respect to antigen receptor assemblies in this study are genetically dissimilar. The residues predicted to contact antigenic peptides are indicated by "P" (76).



Supplementary Fig. 7. Stability of igh antigen receptor networks. The degree of network connectivity is a measure of network structure; the cumulative frequency distributions are shifted to the left, if removal of nodes reduces network connectivity. In this analysis, two thirds of public CDR3 clonotypes and the same number of non-public sequences were randomly removed (indicated above each panel) from the collection of nodes (40 iterations) of four fish. Removal of public sequences has a more drastic effect than removal of non-public sequences; the differences are statistically significant (Mann-Whitney U test) (fish 2, $P=8.8 \times 10^{-15}$; fish 3, $P=1.3 \times 10^{-14}$; fish 4, $P=1.9 \times 10^{-16}$; fish 5, $P=1.1 \times 10^{-14}$).



Supplementary Fig. 8. k-mer plot of the male minifish genome. This plot shows the distribution of k-mer copy number (KCN) at k=17. For better visualization, only KCN values between 3 and 100 are shown.

	Male minifish genome	Female minifish genome
Assembly statistics		
Number of scaffolds	2,827	2,969
Assembly size (Mb)	402.75	403.68
N50 scaffold length (Mb)	7.35	11.03
L50 scaffold count	13	13
Longest scaffold (Mb)	27.48	27.91
No. of contigs	25,040	25,951
N50 contig (Kb)	42.83	36.31
BUSCO analysis		
Complete and single copy genes (number; %)	4,058 (88.53)	4,056 (88.52)
Complete and duplicated genes (number; %)	155 (3.38)	154 (3.36)
Fragmented genes	122 (2.66)	117 (2.55)
Total BUSCOs present	4,335 (94.57)	4,333 (94.52)
Missing BUSCOs	249 (5.43)	251 (5.48)

Supplementary Table 1. Characteristics of minifish genome assemblies.

	Male			Female		
Repeat type	No. of elements	Length (Mb)	%	No. of elements	Length (Mb)	%
SINEs	53,092	7.5	1.87	53,267	5.7	1.41
LINEs	89,400	14.4	3.6	89,181	14.5	3.59
LTR elements	60,245	10.6	2.65	59,643	10.8	2.69
Transposable elements	294,767	49.1	12.2	293,167	49.6	12.28
Unclassified	16,941	2.5	0.63	33,137	45.1	1.12
Simple repeats	351,077	20.8	5.17	351,950	21.1	5.23
Satellites	4,319	0.5	0.13	4,298	0.6	0.14
Small RNA	131	0.02	0	128	0.01	0
Low complexity	33,026	2.6	0.64	32,894	2.6	0.65
Total	918,590	108.2	26.86	917,665	109.4	27.11

Supplementary Table 2. Repeat content in genome assemblies.

Read statistics													
<i>tcrα</i>	fish 2 exp 1	fish 3 exp 1	fish 4 exp 1	fish 5 exp 1	fish 2 exp 2	fish 3 exp 2	fish 4 exp 2	fish 5 exp 2	fish 2 exp 3	fish 3 exp 3	fish 4 exp 3	fish 5 exp 3	
Total reads	67444	45998	110872	64599	43139	41410	32617	71412	132877	181153	196474	306217	
Reads UMI present	51510	35481	88335	50829	31392	31425	20894	52010	105522	142956	158130	242267	
Reads Constant region	50795	35056	87120	50132	30898	31002	20643	51346	34282	40841	67981	75494	
Reads mapped to V at least 3x	27558	18637	34416	26381	17780	15380	6589	19152	18947	24418	26832	36023	
cDNA molecules	793	430	1003	2433	1556	866	1169	3302	358	246	640	1149	
<i>tcrβ</i>													
Total reads	45862	59647	66302	66808	34137	58037	45027	255543	132877	181153	196474	306217	
Reads UMI present	36278	46104	53486	53856	24768	46328	35367	205133	105522	142956	158130	242267	
Reads Constant region	36064	45548	52831	53293	24474	45973	34997	203329	23122	41741	46863	57075	
Reads mapped to V at least 3x	11732	18385	19164	27120	10143	13294	15307	107101	8408	18241	19267	31309	
cDNA molecules	285	348	476	1232	573	607	834	2421	133	180	328	615	
<i>tcrγ</i>													
Total reads	26362	38114	42875	56527	45254	77927	54430	66070	132877	181153	196474	306217	
Reads UMI present	19517	30295	33340	46105	34496	63413	43923	54681	105522	142956	158130	242267	
Reads Constant region	19417	29788	33118	45893	34261	62926	43584	54394	5017	5574	9032	25718	
Reads mapped to V at least 3x	10474	14116	16635	9991	19308	22554	20351	13356	2400	2326	3871	6272	
cDNA molecules	115	69	132	217	245	112	278	374	54	33	84	112	
<i>tcrδ</i>													
Total reads	48492	37062	50553	39849	44837	56826	41566	60425	132877	181153	196474	306217	
Reads UMI present	39801	28924	40691	31451	36527	43727	33048	48409	105522	142956	158130	242267	
Reads Constant region	37492	28348	39465	30604	35337	42371	32012	46671	4890	9380	22226	23863	
Reads mapped to V at least 3x	20347	12966	16729	13439	18214	17765	13125	19681	2400	2326	3871	6272	
cDNA molecules	125	107	188	236	259	180	331	397	51	24	98	77	
<i>igh</i>													
Total reads	48228	120613	62359	84393	89816	277839	129357	213585	132877	181153	196474	306217	
Reads UMI present	39665	96820	52005	68480	72613	225784	106255	174604	105522	142956	158130	242267	
Reads Constant region	39076	95159	51141	67375	71496	221507	103800	171856	36630	43199	9707	56149	
Reads mapped to V at least 3x	33329	77359	36982	55936	54711	153124	76476	108873	29057	31984	7437	36130	
cDNA molecules	1660	1283	126	3564	3469	2642	264	6638	761	561	466	1785	

Supplementary Table 3. RNA_seq read statistics. Total reads, total reads generated in the indicated experiment; Reads UMI present, number of reads containing UMI sequence; Reads Constant region, reads containing constant region sequence; Reads mapped to V at least 3x, only reads mapping to a V element at least three times were kept for further analysis; cDNA, number of distinct cDNA molecules.

Publicity	<i>tcra</i>	<i>trb</i>	<i>trg</i>	<i>trd</i>	<i>igm</i>
1	3,418	4,457	390	1,036	8,343
2	766	335	94	45	815
3	257	43	36	8	216
4	75	5	5	3	49

Supplementary Table 4. Publicity of *tcr* and *igm* assemblies. Unique clonotypes for each of the indicated gene assemblies are tabulated as follows: unique to any of the four individuals (publicity 1), or, shared by at least two (publicity 2), at least three (publicity 3), or all four individuals (publicity 4). Note that the total numbers of *tcra* and *trb* (17), and *igm* (18,19), clonotypes are much higher than previously reported for zebrafish, illustrating the sampling problem associated with the repertoire analysis of fish larger than minifish.

Network characteristics

gene	fish	edges	nodes	mean degree	max degree	max csize	mean csize	fraction largest comp	max diam	mean betweenness	mean authority	mean coreness
<i>tcra</i>	Fish 2	1139	736	3.095108696	12	34	5.18309859	0.0462	6	11.55434783	0.01443506	2.34375
<i>tcrb</i>	Fish 2	146	489	0.597137014	7	19	1.32162162	0.0389	8	1.222903885	0.01574039	0.47239264
<i>tcrg</i>	Fish 2	82	70	2.342857143	8	15	3.68421053	0.2143	4	3.171428571	0.11487548	1.85714286
<i>tcrd</i>	Fish 2	10	77	0.25974026	4	8	1.13235294	0.1039	5	0.493506494	0.05397825	0.18181818
<i>igh</i>	Fish 2	4696	2167	4.334102446	36	1516	4.11195446	0.7	18	2632.077065	0.02993988	2.63867097
<i>tcra</i>	Fish 3	894	595	3.005042017	10	27	5.72115385	0.0454	8	7.675630252	0.0204043	2.25042017
<i>tcrb</i>	Fish 3	304	660	0.921212121	9	55	1.49659864	0.0833	10	9.203030303	0.01467566	0.71515152
<i>tcrg</i>	Fish 3	113	70	3.228571429	12	28	4.375	0.4	6	11.11428571	0.13144945	2.37142857
<i>tcrd</i>	Fish 3	12	121	0.198347107	4	6	1.1	0.0496	3	0.107438017	0.03140171	0.15702479
<i>igh</i>	Fish 3	3116	1496	4.165775401	34	1040	4.15555556	0.6952	19	1723.101604	0.03578665	2.57620321
<i>tcra</i>	Fish 4	1174	742	3.164420485	10	31	5.37681159	0.0418	5	9.284366577	0.01749035	2.39892183
<i>tcrb</i>	Fish 4	490	775	1.264516129	11	84	1.77752294	0.1084	17	27.24645161	0.01934753	0.93677419
<i>tcrg</i>	Fish 4	344	160	4.3	19	53	9.41176471	0.3313	8	24.93125	0.08506423	3.18125
<i>tcrd</i>	Fish 4	22	276	0.15942029	4	5	1.078125	0.0181	2	0.065217391	0.01290418	0.13043478
<i>igh</i>	Fish 4	483	510	1.894117647	18	221	2.125	0.4333	16	221.1647059	0.03853532	1.2254902
<i>tcra</i>	Fish 5	2810	1312	4.283536585	12	92	10.9333333	0.0701	13	34.51676829	0.01127461	3.12271341
<i>tcrb</i>	Fish 5	2145	1846	2.323943662	23	367	2.6	0.1988	27	506.9539545	0.01467063	1.57096425
<i>tcrg</i>	Fish 5	471	233	4.042918455	20	62	10.5909091	0.2661	8	26.90128755	0.06053116	3
<i>tcrd</i>	Fish 5	47	326	0.288343558	6	10	1.12802768	0.0307	4	0.288343558	0.01473556	0.23312883
<i>igh</i>	Fish 5	7948	3440	4.620930233	44	2450	4.28393524	0.7122	17	4522.739244	0.02341413	2.78662791

Supplementary Table S5. Network characteristics. mean degree, mean degree of connectivity of a node; max degree, maximum degree of connectivity of a node; max csize; maximum size of cluster; mean csize; mean size of cluster; fraction of largest component; fraction of nodes in largest cluster; max diam; maximum diameter of any cluster. The network parameters were determined as described in (53).

REFERENCES AND NOTES

1. T. Boehm, M. Hirano, S. J. Holland, S. Das, M. Schorpp, M. D. Cooper, Evolution of alternative adaptive immune systems in vertebrates. *Annu. Rev. Immunol.* **36**, 19–42 (2018).
2. G. W. Litman, M. K. Anderson, J. P. Rast, Evolution of antigen binding receptors. *Annu. Rev. Immunol.* **17**, 109–147 (1999).
3. M. D. Cooper, M. N. Alder, The evolution of adaptive immune systems. *Cell* **124**, 815–822 (2006).
4. J. L. Xu, M. M. Davis, Diversity in the CDR3 region of V(H) is sufficient for most antibody specificities. *Immunity* **13**, 37–45 (2000).
5. I. Engel, S. M. Hedrick, Site-directed mutations in the VDJ junctional region of a T cell receptor β chain cause changes in antigenic peptide recognition. *Cell* **54**, 473–484 (1988).
6. P. Bradley, P. G. Thomas, Using T cell receptor repertoires to understand the principles of adaptive immune recognition. *Annu. Rev. Immunol.* **37**, 547–570 (2019).
7. E. Miho, R. Roškar, V. Greiff, S. T. Reddy, Large-scale network analysis reveals the sequence space architecture of antibody repertoires. *Nat. Commun.* **10**, 1321 (2019).
8. A. Fischer, A. Rausell, Primary immunodeficiencies suggest redundancy within the human immune system. *Sci. Immunol.* **1**, eaah5861 (2016).
9. W. S. DeWitt III, A. Smith, G. Schoch, J. A. Hansen, F. A. T. Matsen, P. Bradley, Human T cell receptor occurrence patterns encode immune history, genetic background, and receptor specificity. *Elife* **7**, e38358 (2018).
10. G. Chen, X. Yang, A. Ko, X. Sun, M. Gao, Y. Zhang, A. Shi, R. A. Mariuzza, N.-P. Weng, Sequence and structural analyses reveal distinct and highly diverse human CD8⁺ TCR repertoires to immunodominant viral antigens. *Cell Rep.* **19**, 569–583 (2017).
11. W. S. DeWitt, P. Lindau, T. M. Snyder, A. M. Sherwood, M. Vignali, C. S. Carlson, P. D. Greenberg, N. Duerkopp, R. O. Emerson, H. S. Robins, A public database of memory and naive B-cell receptor sequences. *PLOS ONE* **11**, e0160853 (2016).
12. O. V. Britanova, M. Shugay, E. M. Merzlyak, D. B. Staroverov, E. V. Putintseva, M. A. Turchaninova, I. Z. Mamedov, M. V. Pogorelyy, D. A. Bolotin, M. Izraelson, A. N. Davydov, E. S. Egorov, S. A. Kasatskaya, D. V. Rebrikov, S. Lukyanov, D. M. Chudakov, Dynamics of individual T cell repertoires: From cord blood to centenarians. *J. Immunol.* **196**, 5005–5013 (2016).
13. R. O. Emerson, W. S. DeWitt, M. Vignali, J. Gravley, J. K. Hu, E. J. Osborne, C. Desmarais, M. Klinger, C. S. Carlson, J. A. Hansen, M. Rieder, H. S. Robins, Immunosequencing identifies

signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nat. Genet.* **49**, 659–665 (2017).

14. B. Briney, A. Inderbitzin, C. Joyce, D. R. Burton, Commonality despite exceptional diversity in the baseline human antibody repertoire. *Nature* **566**, 393–397 (2019).
15. C. Soto, R. G. Bombardi, A. Branchizio, N. Kose, P. Matta, A. M. Sevy, R. S. Sinkovits, P. Gilchuk, J. A. Finn, J. E. Crowe Jr., High frequency of shared clonotypes in human B cell receptor repertoires. *Nature* **566**, 398–402 (2019).
16. M. M. Davis, P. J. Bjorkman, T-cell antigen receptor genes and T-cell recognition. *Nature* **334**, 395–402 (1988).
17. R. Covacu, H. Philip, M. Jaronen, J. Almeida, J. E. Kenison, S. Darko, C. C. Chao, G. Yaari, Y. Louzoun, L. Carmel, D. C. Douek, S. Efroni, F. J. Quintana, System-wide analysis of the T cell response. *Cell Rep.* **14**, 2733–2744 (2016).
18. J. A. Weinstein, N. Jiang, R. A. White III, D. S. Fisher, S. R. Quake, High-throughput sequencing of the zebrafish antibody repertoire. *Science* **324**, 807–810 (2009).
19. N. Jiang, J. A. Weinstein, L. Penland, R. A. White III, D. S. Fisher, S. R. Quake, Determinism and stochasticity during maturation of the zebrafish antibody repertoire. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 5348–5353 (2011).
20. F. Rubelt, C. R. Bolen, H. M. McGuire, J. A. Vander Heiden, D. Gadala-Maria, M. Levin, G. M. Euskirchen, M. R. Mamedov, G. E. Swan, C. L. Dekker, L. G. Cowell, S. H. Kleinstein, M. M. Davis, Individual heritable differences result in unique cell lymphocyte receptor repertoires of naïve and antigen-experienced cells. *Nat. Commun.* **7**, 11112 (2016).
21. M. Schroeder, *Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise* (Dover, 2009).
22. M. Kottelat, R. Britz, T. H. Hui, K.-E. Witte, Paedocypris, a new genus of Southeast Asian cyprinid fish with a remarkable sexual dimorphism, comprises the world's smallest vertebrate. *Proc. Biol. Sci.* **273**, 895–899 (2006).
23. S. Liu, T. H. Hui, S. L. Tan, Y. Hong, Chromosome evolution and genome miniaturization in minifish. *PLOS ONE* **7**, e37305 (2012).
24. M. Malstrom, R. Britz, M. Matschiner, O. K. Torresen, R. K. Hadiaty, N. Yaakob, H. H. Tan, K. S. Jakobsen, W. Salzburger, L. Rüber, The most developmentally truncated fishes show extensive hox gene loss and minaturized genomes. *Genome Biol. Evol.* **10**, 1088–1103 (2018).

25. J. S. Swann, A. Weyn, D. Nagakubo, C. C. Bleul, A. Toyoda, C. Happe, N. Netuschil, I. Hess, A. Haas-Assenbaum, Y. Taniguchi, M. Schorpp, T. Boehm, Conversion of the thymus into a bipotent lymphoid organ by replacement of FOXP1 with its paralog, FOXP4. *Cell Rep.* **8**, 1184–1197 (2014).
26. T. Boehm, I. Hess, J. B. Swann, Evolution of lymphoid tissues. *Trends Immunol.* **33**, 315–321 (2012).
27. A. Brendolan, M. M. Rosado, R. Carsetti, L. Salleri, T. N. Dear, Development and function of the mammalian spleen. *Bioessays* **29**, 166–177 (2007).
28. L. Xie, Y. Tao, R. Wu, Q. Ye, H. Xu, Y. Li, Congenital asplenia due to a *tlx1* mutation reduces resistance to *Aeromonas hydrophila* infection in zebrafish. *Fish Shellfish Immunol.* **95**, 538–545 (2019).
29. Y. Chi, Z. Huang, Q. Chen, X. Xiong, K. Chen, J. Xu, Y. Zhang, W. Zhang, Loss of *runx1* function results in B cell immunodeficiency but not T cell in adult zebrafish. *Open Biol.* **8**, 180043 (2018).
30. N. Danilova, J. Bussmann, K. Jekosch, L. A. Steiner, The immunoglobulin heavy-chain locus in zebrafish: Identification and expression of a previously unknown isotype, immunoglobulin Z. *Nat. Immunol.* **6**, 295–302 (2005).
31. S. L. Seelye, P. L. Chen, T. C. Deiss, M. F. Criscitiello, Genomic organization of the zebrafish (*Danio rerio*) T cell receptor alpha/delta locus and analysis of expressed products. *Immunogenetics* **68**, 365–379 (2016).
32. N. D. Meeker, A. C. Smith, J. K. Frazer, D. F. Bradley, L. A. Rudner, C. Love, N. S. Trede, Characterization of the zebrafish T cell receptor β locus. *Immunogenetics* **62**, 23–29 (2010).
33. F. Wan, C.-B. Hu, J. X. Ma, K. Gao, L.-X. Xiang, J.-Z. Shao, Characterization of $\gamma\delta$ T cells from zebrafish provides insights into their important role in adaptive humoral immunity. *Front. Immunol.* **7**, 675 (2016).
34. H. Li, C. Ye, G. Ji, X. Wu, Z. Xiang, Y. Li, Y. Cao, X. Liu, D. C. Douek, D. A. Price, J. Han, Recombinatorial biases and convergent recombination determine interindividual TCR β sharing in murine thymocytes. *J. Immunol.* **189**, 2404–2413 (2012).
35. N. L. La Gruta, S. Gras, S. R. Daley, P. G. Thomas, J. Rossjohn, Understanding the drivers of MHC restriction of T cell receptors. *Nat. Rev. Immunol.* **18**, 467–478 (2018).
36. H. Tanno, T. M. Gould, J. R. McDaniel, W. Cao, Y. Tanno, R. E. Durrett, D. Park, S. J. Cate, W. H. Hildebrand, C. L. Dekker, L. Tian, C. M. Weyand, G. Georgiou, J. J. Goronzy, Determinants

- governing T cell receptor α/β -chain pairing in repertoire formation of identical twins. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 532–540 (2020).
37. P. J. Brennan, M. Brigl, M. B. Brenner, Invariant natural killer T cells: An innate activation scheme linked to diverse effector functions. *Nat. Rev. Immunol.* **13**, 101–117 (2013).
38. D. Dong, L. Zheng, J. Lin, B. Zhang, Y. Zhu, N. Li, S. Xie, Y. Wang, N. Gao, Z. Huang, Structural basis of assembly of the human T cell receptor–CD3 complex. *Nature*, 546–552 (2019).
39. J. A. Yoder, T. M. Orcutt, D. Traver, G. W. Litman, Structural characteristics of zebrafish orthologs of adaptor molecules that associate with transmembrane immune receptors. *Gene* **401**, 154–164 (2007).
40. G. Gaud, R. Lesourne, P. E. Love, Regulatory mechanisms in T cell receptor signalling. *Nat. Rev. Immunol.* **18**, 485–497 (2018).
41. V. I. Levenshtein, Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* **10**, 707–710 (1966)
42. A. Madi, A. Poran, E. Shifrut, S. Reich-Zeliger, E. Greenstein, I. Zaretsky, T. Arnon, F. Van Laethem, A. Singer, J. Lu, P. D. Sun, I. R. Cohen, N. Friedman, T cell receptor repertoires of mice and humans are clustered in similarity networks around conserved public CDR3 sequences. *eLife* **6**, e22057 (2017).
43. J. Glanville, H. Huang, A. Nau, O. Hatton, L. E. Wagar, F. Rubelt, X. Ji, A. Han, S. M. Krams, C. Pettus, N. Haas, C. S. Lindestam Arlehamn, A. Sette, S. D. Boyd, T. J. Scriba, O. M. Martinez, M. M. Davis, Identifying specificity groups in the T cell repertoire. *Nature* **547**, 94–98 (2017).
44. H. Huang, C. Wang, F. Rubelt, T. J. Scriba, M. M. Davis, Analyzing the Mycobacterium tuberculosis immune response by T-cell receptor clustering with GLIPH2 and genome-wide screening. *Nat. Biotechnol.* **38**, 1194–1202 (2020).
45. P. Dash, A. J. Fiore-Gartland, T. Hertz, G. C. Wang, S. Sharma, A. Souquette, J. C. Crawford, E. B. Clemens, T. H. O. Nguyen, K. Kedzierska, N. L. La Gruta, P. Bradley, P. G. Thomas, Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature* **647**, 89–93 (2017).
46. J. Robert, E.-S. Edholm, A prominent role for invariant T cells in the amphibian *Xenopus laevis* tadpoles. *Immunogenetics* **66**, 513–523 (2014).
47. E. N. Rittmeyer, A. Allison, M. C. Gründler, D. K. Thompson, C. C. Austin, Ecological guild evolution and the discovery of the world’s smallest vertebrate. *PLOS ONE* **7**, e29797 (2012).

48. H. S. Robins, S. K. Srivastava, P. V. Campregher, C. J. Turtle, J. Andriesen, S. R. Riddell, C. S. Carlson, E. H. Warren, Overlap and effective size of the human CD8⁺ T cell receptor repertoire. *Sci. Transl. Med.* **2**, 47ra64 (2010).
49. N. M. Provine, P. Klenerman, MAIT cells in health and disease. *Annu. Rev. Immunol.* **38**, 203–228 (2020).
50. C. Song, S. Havlin, H. A. Makse, Self-similarity of complex networks. *Nature* **433**, 392–395 (2005).
51. R. Albert, H. Jeong, A. L. Barabasi, Error and attack tolerance of complex networks. *Nature* **406**, 378–382 (2000).
52. M. A. Turchaninova, O. V. Britanova, D. A. Bolotin, M. Shugay, E. V. Putintseva, D. B. Staroverov, G. Sharonov, D. Shcherbo, I. V. Zvyagin, I. Z. Mamedov, C. Linnemann, T. N. Schumacher, D. M. Chudakov, Pairing of T-cell receptor chains via emulsion PCR. *Eur. J. Immunol.* **43**, 2507–2515 (2013).
53. G. Csardi, T. Nepusz, The igraph software package for complex network research. *InterJournal Complex Systems*, 1695 (2006).
54. D. M. Langenau, A. A. Ferrando, D. Traver, J. L. Kutok, J.-P. Hezel, J. P. Kanki, L. I. Zon, A. T. Look, N. S. Trede, In vivo tracking of T cell development, ablation, and engraftment in transgenic zebrafish. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 7369–7374 (2004).
55. I. Hess, T. Boehm, Intravital imaging of thymopoiesis reveals dynamic lympho-epithelial interactions. *Immunity* **36**, 298–309 (2012).
56. G. Marçais, C. Kingsford, A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. *Bioinformatics* **27**, 764–770 (2011).
57. E. S. Lander, M. S. Waterman, Genomic mapping by fingerprinting random clones: A mathematical analysis. *Genomics* **2**, 231–239 (1988).
58. G. W. Vulture, F. J. Sedlazeck, M. Nattestad, C. J. Underwood, H. Fang, J. Gurtowski, M. C. Schatz, GenomeScope: Fast reference-free genome profiling from short reads. *Bioinformatics* **33**, 2202–2204 (2017).
59. N. I. Weisenfeld, S. Yin, T. Sharpe, B. Lau, R. Hegarty, L. Holmes, B. Sogoloff, D. Tabbaa, L. Williams, C. Russ, C. Nusbaum, E. S. Lander, I. MacCallum, D. B. Jaffe, Comprehensive variation discovery in single human genomes. *Nat. Genet.* **46**, 1350–1355 (2014).
60. Z. Zhang, S. Schwartz, L. Wagner, W. Miller, A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.* **7**, 203–214 (2000).

61. X. Huang, A. Madan, CAP3: A DNA sequence assembly program. *Genome Res.* **9**, 868–877 (1999).
62. N. H. Putnam, B. L. O'Connell, J. C. Stites, B. J. Rice, M. Blanchette, R. Calef, C. J. Troll, A. Fields, P. D. Hartley, C. W. Sugnet, D. Haussler, D. S. Rokhsar, R. E. Green, Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res.* **26**, 342–350 (2016).
63. A. V. Zimin, G. Marcais, D. Puiu, M. Roberts, S. L. Salzberg, J. A. Yorke, The MaSuRCA genome assembler. *Bioinformatics* **29**, 2669–2677 (2013).
64. A. C. English, S. Richards, Y. Han, M. Wang, V. Vee, J. Qu, X. Qin, D. M. Muzny, J. G. Reid, K. C. Worley, R. A. Gibbs, Mind the gap: Upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLOS ONE* **7**, e47768 (2012).
65. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
66. M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, A. Regev, Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
67. W. Li, A. Godzik, Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
68. A. Yates, W. Akanni, M. R. Amode, D. Barrell, K. Billis, D. Carvalho-Silva, C. Cummins, P. Clapham, S. Fitzgerald, L. Gil, C. G. Giron, L. Gordon, T. Hourlier, S. E. Hunt, S. H. Janacek, N. Johnson, T. Juettemann, S. Keenan, I. Lavidas, F. J. Martin, T. Maurel, W. McLaren, D. N. Murphy, R. Nag, M. Nuhn, A. Parker, M. Patricio, M. Pignatelli, M. Rahtz, H. S. Riat, D. Sheppard, K. Taylor, A. Thormann, A. Vullo, S. P. Wilder, A. Zadissa, E. Birney, J. Harrow, M. Muffato, E. Perry, M. Ruffier, G. Spudich, S. J. Trevanion, F. Cunningham, B. L. Aken, D. R. Zerbino, P. Flicek, Ensembl 2016. *Nucleic Acids Res.* **44**, D710–D716 (2016).
69. F. A. Simão, R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, E. M. Zdobnov, BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
70. A. Smit, R. Hubley, *RepeatModeler Open-1.0* (2008–2015); www.repeatmasker.org.
71. G. Abrusán, N. Grundmann, L. DeMester, W. Makalowski, TEclass—A tool for automated classification of unknown eukaryotic transposable elements. *Bioinformatics* **25**, 1329–1330 (2009).

72. W. Bao, K. K. Kojima, O. Kohany, Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* **6**, 11 (2015).
73. A. Smit, R. Hubley, *RepeatMasker Open-4.0* (2013–2015); www.repeatmasker.org.
74. B. L. Cantarel, I. Korf, S. M. Robb, G. Parra, E. Ross, B. Moore, C. Holt, A. Sanchez Alvarado, M. Yandell, MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* **18**, 188–196 (2008)..
75. M. Stanke, S. Waack, Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19 Suppl 2**, ii215–ii225 (2003)
76. U. Grimholt, K. Tsukamoto, T. Azuma, J. Leong, B. F. Koop, J. M. Dijkstra, A comprehensive analysis of teleost MHC class I sequences. *BMC Evol. Biol.* **15**, 32 (2015).