doi:10.1038/nature19840

Supplementary Materials

Contents

Supplementary Note 1. Polyploidy and the allotetraploidy hypothesis 4
1.1 Allo vs. autopolyploidy 4
1.2. Cytogenetics and inheritance in allopolyploids 5
1.3 Mechanisms of allotetraploid formation 5
1.4 Experimental formation of higher polyploids in *Xenopus* 6
1.5 Impact of neoallotetraploidization 7
1.6 Formation and features of neoautopolyploids 7
1.7 Long-term response to tetraploidization after disomy 8
1.8 Definitive identification of an allopolyploid 9

Supplementary Note 2. *X. laevis* shotgun sequencing and assembly 10
2.1 History of the J strain 10
2.2 Preparation of genomic DNA 10
2.3 Plasmid library preparation and shotgun sequencing 10
2.4 BAC and fosmid library preparation and sequencing 10
2.5 Fosmid pool sequencing 11
2.6 Shotgun assembly 11
2.7 Assembly summary 13
2.8. Validation of assembled shotgun genome 14
2.9 Self-consistency of shotgun genome 14
2.10 Comparison with finished BACs 14

Supplementary Note 3. Chromosome-scale assembly 15
3.1 Chromosome assignment with BAC-FISH 15
3.2 *In vivo* long-range linkage with tethered conformation capture (HiC) 15
3.3 Chicago long-range linking library from *in vitro* chromatin 17
3.4 Scaffolding the draft genome with HiRISE 17
3.5 Construction of chromosome-scale pseudomolecules 18

Supplementary Note 4. RNA-seq transcriptome resources 19
4.1 Collection of *X. laevis* transcriptome public resources 19
4.2 Collection of large scale J-strain *X. laevis* transcriptome resources 20
4.3 *De novo* assembly of transcriptome data 20

Supplementary Note 5. Annotation of protein-coding genes and miRNAs 21
5.1 *De novo* repeat identification and masking 21
5.2. Protein-coding gene annotation: overview 22
5.3. Initial annotation 22

5.4 Extension of gene models by pita 23
5.5 Final annotation of chromosome scale assembly 23
5.6. Manual validation of gene models 24
5.7 Annotation of microRNAs 24

Supplementary Note 6. Chromosome evolution 24
6.1 Large-scale genomic rearrangements 24
6.2 The fusion of homologs of XTR 9 and 10 25
6.3 Analysis of the *X. laevis* sex locus 26

Supplementary Note 7. Subgenome-specific repeats 26
7.1 Initial identification of subgenome-specific transposable elements 26
7.2 Distribution of subgenome specific elements 27
7.3 Validation of cross mapped subgenome-specific repeat subfamilies 28
7.4 Transposable element chromosome FISH 29
7.5 Timing of L- and S-specific and enriched transposable element activity. 29
7.6 Global analysis of *Xenopus* repetitive element ages 30

Supplementary Note 8. Phylogeny, divergence times, and evolutionary rates 31
8.1 Identification of orthologous and homoeologous protein-coding genes 31
8.2 Comparison to previous estimates of gene retention 31
8.3 Annotated sequence alignments 32
8.4 Phylogeny 33
8.5 Estimate of substitution rate 34
8.6 microRNAs 34
8.7 pan-vertebrate conserved elements 35
8.8 Estimate of nucleotide diversity (polymorphism within *X. laevis*) 35
8.9 Whole-genome alignment 35

Supplementary Note 9: Unitary pseudogenes 36

Supplementary Note 10. Patterns of retention and deletion. 37
10.1 Retention and gene function 37
10.2 Modeling gene loss and neo/sub-functionalization 38
10.3 Retention of duplicated genes in protein complexes 41
10.4 Retention of ohnologs retained from ancient vertebrate duplications 42
10.5 Retention relationship with gene length 42

Supplementary Note 11: Local duplications 43
11.1 Analysis of locally duplicated genes 43
11.2 Nomenclature of duplicated genes 44
11.3 Nomenclature of pseudogenes 45

Supplementary Note 12: Gene expression 45
12.1 Quantification of gene expression levels with RNA-seq 45
12.2 Gene expression vs. developmental time and tissues 45
12.3 Co-expression network inference and analysis of modules 46

12.4 Evolution of homoeologous expression following allopolyploidy....47
12.5 Thanagenes....48
Supplementary Note 13: Analysis of specific gene families and pathways....48
13.1. Six6....48
13.2. Cell cycle....49
13.3. TGF-beta....49
13.4 Immune Genes....50
13.5 Hippo signaling....52
13.6 Hedgehog (Hh) signaling....53
13.7 Hox clusters....53
13.8 *mix*, *mixer*, *bix*....54
13.9 *nodal5*, *nodal3* and *vg1*....54
13.10 Wnt signaling....55
13.11 Germ plasm....55
13.12 Mitochondria....55
Supplementary Note 14: Epigenetics....55
14.1 ChIP-seq experimental protocol....56
14.2 ChIP-seq data analysis....56
14.3 MethylC-seq for whole-genome bisulfite sequencing....56
14.4 Epigenetic differences explain differential expression between the L and S subgenome....57
Supplementary Note 15: Funding and data availability....59
15.1 Funding....59
15.2 Data availability....59
References....61

Supplementary Note 1. Polyploidy and the allotetraploidy hypothesis

Here we summarize basic concepts related to polyploidy. Excellent reviews can be found in other places[1–5].

1.1 Allo vs. autopolyploidy

A polyploid organism possesses three or more "sets" of chromosomes in its somatic cells, where sets are typically recognized cytogenetically or by gene content. Two distinct classes of polyploids are recognized based on the manner in which they are formed:

- **Allopolyploids** arise through interspecific hybridization, which brings together diverged but recognizably related chromosome sets from two distinct diploid progenitor species.
- In contrast, **autopolyploids** arise when individuals are formed with more than two chromosome sets from a single species.

Related chromosomes within a polyploid are referred to as **homoeologues** (or homoeologs, or homoeologues). The base chromosome number, *x*, is the number of chromosomes contained in a complete set of chromosomes.

While "polyploidy" broadly refers to multiple sets of chromosomes, the timing of the polyploidization event is also relevant.

- Recently formed polyploids are referred to as **neopolyploids**, and their diploid progenitor(s) are often identifiable. In some cases neopolyploids have been synthetically produced in the laboratory by hybridization and/or suppression of cytokinesis after replication[6]. Neopolyploids can be isolated individuals or populations and may exhibit reduced fertility or even sterility.
- In contrast, **paleopolyploids** formed in the (possibly distant) past and have undergone subsequent evolutionary changes, notably including gene loss and/or divergence of homoeologous sequences. Depending on the timing of the ancient polyploidy event, the original diploid progenitor(s) may be extinct or not recognizable. Paleopolyploidy is a feature of species, and is therefore generally propagated by sexual reproduction except in rare cases (*e.g.*, bdelloid rotifers).

Whole genome duplication is a generic term that encompasses various kinds of polyploidy, and emphasizes (1) the duplication of a complete set of ancestral chromosomes (resulting in an initial doubling of gene number, subsequently relaxed in paleopolyploids by gene loss) and (2) the unlinked nature of the gene duplicates created by polyploidy. In the case of allopolyploids, the "ancestral" chromosomes refer to those of the common ancestor of the diploid progenitors.

1.2. Cytogenetics and inheritance in allopolyploids

In a diploid, chromosomes that pair in meiosis I are referred to as **homologs** (or homologues). The term "homolog" is also commonly used to refer to genes or chromosomes in different species that share a common ancestor. When we need to distinguish the two meanings of homolog we will say "meiotic" or "evolutionary" homolog, respectively. As noted above, in an allopolyploid, recognizably similar chromosomes from distinct progenitors are referred to as homoeologues (or homoeologs, homoeologues) to recognize their relatedness.

For simplicity let us refer to the somatic chromosome complement of one diploid progenitor as AA and the other as BB. Then an allotetraploid formed from these progenitors has composition (AA,BB). The A and B chromosomes are homoeologous. If A and B are diverged enough that homoeologous pairing does not occur in meiosis I, then only A-A and B-B bivalents will form. A and B alleles will then assort independently, and inheritance is **disomic**.

In the older literature, especially in discussion of allopolyploid animals, homoeologous loci are sometimes referred to as "allo-alleles." We do not use this terminology since it suggests, incorrectly, that alleles at these loci segregate. In fact, the unlinked A and B loci assort independently, and inheritance is conventionally Mendelian (*i.e.*, disomic at each locus).

Neoautopolyploids may show multivalent pairings[1,2], since more than two chromosomes are meiotically homologous. This situation generally compromises fertility, as aberrant resolution of multivalents typically generates unbalanced gametes (*i.e.*, gametes without an integer number of complete chromosome sets). This block can be overcome by parthenogenesis or vegetative reproduction (as in many autopolyploid plants), by reduction in the number of pairing centers to allow only bivalents to form, or by other poorly understood mechanisms that allow resolution of initial multivalent pairing into bivalents that can execute proper meiotic segregation. In wheat, disomic meiotic segregation can be disrupted by mutation of the *ph1* (pairing homoeologous 1) locus[7]. Note that if homoeologues do not have consistent pairing partners, then inheritance is **polysomic**, that is, multiple variant alleles or haplotypes are segregating at each locus.

1.3 Mechanisms of allotetraploid formation

Various mechanisms can lead to the formation of allotetraploids (Extended Data Fig 1; for more discussion see the reference[1]).

- Interspecific hybridization of AA and BB diploids, to form an AB hybrid, followed by genome replication without cell division to form an AABB individual. This can be induced experimentally by treatment with colchicine or similar agents that block cell division.
- Interspecific hybridization of unreduced gametes from both parental species. In some species, unreduced gametes are common, especially under stress.

- Processes that involve a triploid intermediate. For example, interspecific hybridization of unreduced gametes from one diploid species (AA) with normal reduced gametes from another diploid (BB), can produce a triploid individual (AAB). In a more complicated scenario with some support in *Xenopus* species[8,9], interspecific hybridization of an AA and BB diploid yields AB hybrids that produce unreduced AB gametes. (in *Xenopus*, this only occurs in interspecific hybrid females, which can produce AB eggs; males are infertile. When these are fertilized by 1N sperm from one of the progenitors (*e.g.*, species A) the result is a triploid AAB zygote. Regardless of how a triploid is produced, if an AAB female produces unreduced (AAB) eggs that fuse with a normal B sperm, the result is an AABB individual.

AABB individuals are expected to be fully fertile. Ongoing questions involve the reproductive isolation of the incipient allotetraploid population relative to its (initially more common) diploid progenitors, and dosage compensation specifically as it relates to sex determination, which is presumed to account for the absence of polyploids in mammals; the existence of polyploids in fish and amphibians whose sex determination mechanisms may be strongly affected by non-genetic factors; and the prevalence of polyploidy in plants without chromosomal sex determination.

1.4 Experimental formation of higher polyploids in *Xenopus*

The genus *Xenopus* includes species with somatic chromosome number 2N=36 (*X. laevis, X. mulleri, X. borealis*), 54, 72, and 108 (*X. ruwenzoriensis*), and species with 2N=20 (*Xenopus* (formerly *Silurana*) *tropicalis*) and 40 (*X. epitropicalis*)[8–11]. Cytogenetic analysis and comparison of DNA content with other frogs demonstrates that these species belong to polyploid series with a base chromosome number of N=9 and 10, respectively. Yet no species with N=9 has been described, so that diploid relatives of the *X. laevis* series are apparently no longer extant.

An elegant series of experimental manipulations of tetraploid *Xenopus* suggests scenarios for higher order polyploid formation from an *X. laevis*-like species with 2N=36 as the basic diploid (adapted from other references[8,9]). For example, hybridization of pairs of species *Xa* and *Xb*, both with 2N=36, can produce hybrid zygotes that develop normally. While male progeny are generally sterile, females produce large unreduced oocytes with 2N=36 containing one set of *Xa* and one set of *Xb* chromosomes. These eggs can be fertilized by 1N=18 sperm from species *Xa*, producing triploid zygotes (relative to *Xa* and *Xb*) that develop normally with 3N=54 chromosomes. These males are also sterile, but as with interspecific hybrids, such females produce large unreduced oocytes carrying the somatic chromosome complement, in this case 2 sets of *Xa* chromosomes and one set of *Xb* chromosomes. These can in turn be fertilized by normal sperm from *Xb* males, restoring a chromosome complement of paired chromosomes (now two sets from *Xa* and two sets from *Xb*), but with a doubled number relative to the progenitor species. Since each chromosome has a meiotic homolog, the resulting polyploid has 2N=72 and should be fully fertile. It is an allotetraploid relative to the progenitors *Xa* and *Xb*.

Although there are no extant diploid *Xenopus* species with 2N=18, it is plausible that a similar scenario, with two distinct 2N=18 progenitors, could have led to the formation of a 2N=36 species that diversified further to form the extant *X. laevis* group. This is shown schematically in Extended Data Fig. 1d.

1.5 Impact of neoallotetraploidization

The initial formation of an allotetraploid brings two distinct sets of chromosomes into the same nucleus, so that each progenitor genome (now subgenome) confronts a novel *trans*-regulatory environment. Experimental production of interspecific hybrids and neoallotetraploids in the laboratory has demonstrated early responses to this condition within a few generations. Many of these same early responses are found in diploid hybrids (*e.g.*, with AB genotype, sometimes called "homoploid hybrids"[1]), where they include hybrid dysgenesis.

1. **Activation of transposable elements.** This response, called "genomic shock" by McClintock[12], has been discussed in other references[3,13]. Mechanistically, transposons that were epigenetically silenced in the two diploid progenitors may lose their silencing (due, for example, to altered dosage of suppressing factors or in response to changes in DNA methylation and chromatin). This may share some features with hybrid dysgenesis.
2. **Genome-wide changes in gene expression.** Epigenetic changes, including methylation and alteration of chromatin, can lead to genome-wide changes in gene expression. Remarkably, this can in some cases lead to reciprocal silencing of homoeologous genes in different tissues. Since such responses have been documented in the first or second generation after polyploidization, they presumably are due to epigenetic changes (or response to a new *trans*-environment in the neoallopolyploid) rather than rapid mutation, although the mechanisms of this response are still under active investigation (See others[14,15] for review).
3. **Rapid genomic rearrangement and loss.** Aberrant meiotic pairing and/or transposon activation can lead to chromosomal translocations and/or deletions. For example, synthetic wheat allotetraploids (intended to recapitulate the formation of natural *Triticum*/*Aegilops* allotetraploids) show rapid loss of certain sequences in an apparently reproducible manner[16]. Similarly, *Brassica* neoallopolyploids show rapid chromosomal rearrangement, perhaps due to the effects of mispairing, multivalent resolution, and/or reciprocal translocation[14,17].

1.6 Formation and features of neoautopolyploids

Autopolyploids can arise through endoreduplication or by the fusion of unreduced gametes. Meiosis in a neoautopolyploid may encounter challenges not faced by allotetraploids. Specifically, since all chromosome copies in an autopolyploid are (meiotically) homologous, pairing in meiosis I can lead to multivalent structures. If these multivalent structures are not resolved (later in meiosis) in a manner that allows proper segregation, unbalanced gametes will be produced. Thus triploid autopolyploids are typically infertile except for rare unreduced triploid gametes. Mechanisms that avoid the formation of multivalents (*e.g.*, reduction in the number of pairing sites per chromosome) may be favored.

In an autotetraploid, if pairing and segregation occurs randomly between two of the four homoeologous chromosomes, then each locus segregates up to four alleles. This is known as **tetrasomic** inheritance, in

contrast to the conventional Mendelian disomic pattern. Among the four homoeologous chromosomes, there may be "preferential" pairing if not all chromosomes are equally likely, based on differences or similarities between specific homoeologues[18].

"**Diploidization**" (or, sometimes, "rediploidization") in an autopolyploid refers to the process by which chromosomes change to enforce consistent bivalent pairing between a specific pair of homologous chromosomes, resulting in the restoration disomic inheritance. Once disomic inheritance is re-established, recombination can no longer occur between homoeologous chromosomes, and they will diverge from one another, leading to an AAA'A' karyotype that now evolves like a diploid. Chromosomal rearrangements during diploidization can obscure the original subgenome relationships.

1.7 Long-term response to tetraploidization after disomy

As described above, either immediately (for neoallopolyploids) or after some "diploidization" period (for autopolyploids), disomy is reestablished. Once this state is reached tetraploid genomes are shaped by the longer-term evolutionary forces of mutation, drift, selection, and recombination, leading to further differentiation of the subgenomes. This subsequent evolution is also sometimes referred to as "diploidization"[4], but we stress that, as noted above, allotetraploids are expected to show disomic inheritance (genetic diploidy) as soon as they are formed.

The lifetime of a duplicate gene can be estimated by a simple calculation, as discussed by Lynch and Force[19]. They show that if a pair of duplicate loci are completely redundant with each other, then one of the two loci will be lost (*i.e.*, non-functional alleles will become fixed, either through deletion or disrupting mutations) on a time scale of $1/\mu_c$ generations for small populations, and ~4 Ne generations for large populations. Here μ_c is the mutation rate to a non-functional allele, which is roughly 10^{-5} to 10^{-6} per generation in mammals[20,21] and is expected to be comparable in *Xenopus*. Ne is the effective population size, which under a neutral model can be estimated from the nucleotide diversity according to Gillespie[22]. $\pi = 4\ Ne\ \mu_N$ where μ_N is the nucleotide substitution rate per generation. We measure $\pi \sim 0.5\%$, and below estimate μ_N to be ~3×10^{-9}, implying that Ne ~ 500,000. It follows that if a pair of duplicate loci are completely redundant with each other, and in the absence of selection or other classes of mutation, one of the two loci will be "lost" (*i.e.*, non-functional alleles will become fixed, either through deletion or disrupting mutations) on a time scale of 500,000 ~ 1,000,000 generations, or several million years in *Xenopus*. This is consistent with a delayed onset of pseudogene creation of a few million years of the polyploidy event, as reported in Supplementary Note 9.

Several mechanisms allow deviation from this simple scenario.

- Ohno[23] suggested the importance of **neofunctionalization** in which mutations at one of two duplicate loci produce novel functions such that both loci become subject to purifying selection and loss is prevented.
- The importance of **dosage** has been emphasized by Birchler and colleagues[24,25]. Cell size scales with DNA content, so that polyploid cells are larger than diploid cells, placing different metabolic,

transport, and signaling constraints on the genome. This may lead to selection for higher dosage and the enforced retention of two copies for genes whose products are required in higher doses, and selection for lower dosage for other genes. (Lower dosage can be achieved by deleterious mutation in one or both homoeologues, or by complete loss of one homoeologous locus.)

- The importance of **sub-functionalization** was pointed out by Force, Lynch and colleagues[19,26]. They note that genes often have multiple independently mutable functions (*e.g.*, regulatory elements that drive expression in different tissues). Duplicate loci can acquire complementary mutations in different functions, at which point both loci become indispensable if all ancestral functions are to be retained. Under some population-genetic conditions (see ref[19]), sub-functionalization can drive the fixation of duplicate genes with diverged functions. This model has the conceptual advantage that is relies only on the occurrence of disabling (loss-of-sub-function) mutations rather than presumably rarer beneficial neofunctionalizing changes. Of course, once the two loci become immune to loss, additional mutations can accumulate, leading to further divergence in function.

These mechanisms leading to gene retention are not exclusive, and the mechanism for retention or loss of a specific homoeologous gene pair can depend on stochastic factors (drift) and gene structure (via mutability) or function (dosage sensitivity, availability of subfunctionalizing or neofunctionalizing mutation) which can differ between genes.

1.8 Definitive identification of an allopolyploid

The defining feature of an allopolyploid is that its subgenomes once existed as the genomes of two distinct diploid progenitor species. These diploid progenitors in turn descend from some more ancient ancestor, diverging from one another until reunited by hybridization (Fig. 2). During this interval, the two diploid progenitors evolve independently. In particular, they can acquire distinct transposable element complements that we can use to differentiate subgenomes.

1. As the two progenitor species diverge from their common ancestor, they accumulate species-specific transposable elements that mark their chromosome sets. Relicts of these elements will consistently differentiate the two subgenomes of an allotetraploid. In contrast, since the subgenomes of a genetically diploidized autotetraploid have always shared the same nucleus both pre- and post-tetraploidization, they cannot acquire distinguishing transposable elements.
2. Conversely, transposable elements that are shared between subgenomes must be either (a) older than the divergence of the progenitor species, or (b) younger than the allotetraploidization event. While the two progenitors exist as distinct species, their transposable element activity occurs independently. Thus pan-genome elements (active in both subgenomes) cannot originate during this period. Again in contrast, in a diploidized autotetraploid the transposon activity of both subgenomes should occur in parallel.

The timing of transposable element activity is considered in Supplementary Note 7 below.

Supplementary Note 2. *X. laevis* shotgun sequencing and assembly

2.1 History of the J strain

The history of the J strain is illustrated in Extended Data Fig. 1b. The J strain originates from *Xenopus laevis* individuals introduced from Switzerland to the USA, and eventually brought to Japan and bred. In the course of breeding in Japan, frogs exhibited no "short-term skin rejection"[27], indicating that the MHC locus was almost homozygous, and after repeated single-pair mating for a further 11 generations, the 21st generation population was named the J strain exhibiting no "long-term skin rejection," indicating that most genes are homozygous[28]. Frogs of the 30th generation were sent from Japan to the USA, and one female from the descendant frogs was used for shotgun sequencing. Animals of the 32nd ~ 34th generations kept in Japan served to provide materials for construction of BAC and fosmid libraries, FISH analyses and RNA-seq.

2.2 Preparation of genomic DNA

One female of the J strain was used for shotgun sequencing. *Xenopus laevis* genomic DNA was extracted from erythrocytes as described previously[29]. Briefly, erythrocytes were isolated, lysed in hypotonic buffer, and nuclei isolated by centrifugation. The pellet was resuspended and genomic DNA released with detergent and overnight proteinase K treatment. The DNA was spooled after precipitation with NH4OAc/isopropanol, washed several times in 70% EtOH, and resuspended in TE buffer.

2.3 Plasmid library preparation and shotgun sequencing

The libraries we used for genome assembly are summarized in Supplemental Table 1. Illumina prepared the mate-pair libraries for sequencing.

2.4 BAC and fosmid library preparation and sequencing

A single J-strain female (32nd generation) was used for the XLB1-BAC library, and another female (33rd generation) was used for the XLB2-BAC and XLFIC fosmid libraries, according to the procedures previously described[30]. In brief, blood cells were collected from a frog under the anesthetized condition with MS222 (Sigma), and were embedded in 1% agarose gel plug and subjected to pulsed-field gel electrophoresis to obtain DNA fragments ranging from 125 to 225 kb after partial digestion either with *SacI* for XLB1 BAC, or *HindIII* for XLB2 BAC libraries. Cloning vectors, pKS145 and pKS200 were used to construct XLB1 and XLB2, respectively. Each BAC clone grown in *E. coli* DH10B was picked up and

arrayed into standard 384-well titer-plates. The total number of the isolated clones for the XLB1 library was 141,312 (5.6 X coverage of the *X. laevis* genome), and that for the XLB2 library was 19,200 (0.67 X coverage).

The fosmid library, XLFIC, was constructed from sheared genomic DNA and the pKS300-IC cloning vector. After *in vitro* packaging using Gigapack III Gold Packaging Extract (Agilent Technologies, #200203), *E. coli* XL1-BLUE was infected with the phage particles. The total number of the clones in the XLFIC library was 59,904 clones (0.67 X coverage of the genome).

End-sequencing of 153,600 BAC (134,400 and 19,200 clones from the XLB1 and XLB2 libraries, respectively) and 59,904 fosmid (XLFIC) clones was carried out using the BigDye terminator kit version 3 (Applied Biosystems) and the ABI 3730xl capillary sequencers (Applied Biosystems). Out of the end-sequenced BAC/fosmid clones, fifty-two clones that were localized to regions of biological interest were further sequenced to completion by shotgun sequencing and assembly using the KB basecaller/Phrap/Consed systems as previously described[31]. Gaps and low-quality regions in the initial assembly were both closed and re-sequenced by primer walking and direct sequencing of the PCR products produced from the DNA of original clones. Sequence data from BAC and fosmid clones have been deposited to DDBJ/GenBank/EMBL under the accession numbers: (i) GA131508-GA227532, GA228275-GA244139, GA244852-GA274229, GA274976-GA275712, GA277157-GA344957, GA345673-GA350926, and GA351685-GA393223 for the XLB1 end-sequences, (ii) GA720358-GA756840 for the XLB2 end-sequences, (iii) GA756841-GA867435 for the XLFIC end-sequences, and (iv) AP012997-AP013026, AP014660-AP014679, AP017316 and AP017317 for the finished BAC/fosmid sequences.

2.5 Fosmid pool sequencing

Additional large-insert fosmid clone libraries were prepared as previously described[32]. Briefly, high molecular weight genomic DNA was sheared to 20 ~ 50 kbp in a Hydroshear instrument for 20 cycles at speed code 16. Sheared DNA was size-separated by pulsed-field gel electrophoresis, and the gel was stained with SYBR Gold (Invitrogen) for visualization. Two fractions were excised, a “HI” fraction at 38–40 kbp, and a “LO” fraction at 30–38 kbp. DNA was purified from gel slices by beta-agarase treatment, end-polished, and ligated to the fosmid vector backbone pCC1FOS (Epicentre), packaged into phage and infected into the *E. coli* cloning host. Plate titer counts indicated library titers of $3x10^5$ clones (HI library) and $7x10^5$ clones (LO library). The HI library was divided into three fractions of approximately equal size (plate1, plate2, and bulk-HI), and the LO library was kept as a single pool. Each of the resulting clone pools was expanded by outgrowth, and cloned DNAs were isolated by standard alkaline lysis miniprep. Mate-paired clone-end libraries were prepared as previously described[33,34], and sequenced on an Illumina Hiseq 2000 instrument with paired-end 100 bp reads.

2.6 Shotgun assembly

We assembled the *X. laevis* genome using Illumina shotgun sequence from a single individual of the well-documented highly inbred J-strain (see Extended Data Fig. 1b). The use of an inbred line minimizes the impact of allelic heterozygosity on assembly. We expected that *Xenopus* tetraploidy would not confound shotgun assembly based on previous studies of cDNAs[35] and expressed sequence tags (ESTs)[36] which demonstrated that paralogous coding sequences had ~94% nucleotide identity, with an estimated synonymous substitution rate of Ks ~ 0.26. We confirmed these results with homoeologous protein-coding genes from our *X. laevis* genome assembly (Supplemental Table 3). In general, we expect that coding sequences represent the least diverged homoeologous sequences, and that intronic and non-coding sequences will typically show more divergence.

Contigs and scaffolds were constructed with an improved version of Meraculous[37], which performed well in the Assemblathon 2 comparisons[38]. Contigs were formed from 2x151 bp paired-end Illumina sequences from fragment libraries with insert size ~225 bp, ~425 bp, and ~750 bp, totaling 298.2 Gbp, or ~99X raw coverage of an estimated ~3 Gbp genome[39]. Scaffolds were formed using these paired-end sequences; mate-pairs (~1.5 kb and ~4.5 kb); 10 kb jumping library (Illumina); fosmid-ends (~35 kb; Lucigen); and BAC-ends (~120 kb). Datasets are summarized in Supplemental Table 1.

Briefly, meraculous proceeds as follows. First, reads are decomposed into their overlapping k-mers, where empirically a value of k=51 was used. The histogram of 51-mer counts is shown in Extended Data Fig. 1. The peak at d ~ 30x represent 51-mers that are unique (single copy) in the *Xenopus laevis* J-strain genome. Note that this "k-mer depth" is less than the raw sequence depth of 99x because (1) a read of length R contains only R-k+1 distinct k-mers, and (2) each isolated error damages up to k k-mers. The peak near zero counts represents 51-mers that span sequencing errors in the reads, where each error generates up to 51 distinct 51-mers that are typically not found in the genome, and occur at low frequency (predominantly single occurrences) in the shotgun dataset (see ref[37], for example). The absence of a significant peak at double depth (2d~100x) (Extended Data Fig. 1) is consistent with the observation above that the two homoeologous subgenomes of *X. laevis* are diverged at the ~6% level, so there are few 51-mers that occur exactly twice in the genome.

The cumulative count-weighted distribution of 51-mers vs. genomic copy number in the shotgun dataset is shown in Extended Data Fig. 1, where the horizontal x axis has been scaled to genomic copy number by dividing the 51-mer count by peak depth d, and is shown with a logarithmic scale. The vertical axis estimates the fraction of the genome spanned by 51-mers with copy number less than x. While ~75% of the *X. laevis* genome is single copy (based on the knee at genomic copy number ~1), the remaining repetitive sequence is predominantly found in 51-mers with more than 10 copies in the genome (15%). Thus most 51-mers are either single copy, or very high copy number.

Meraculous "UU contigs" correspond to uncontested assembled k-mer walks, and are the starting point for assembly. To build "UU contigs," we identified all "UU" 51-mers that (1) occur three or more times in the shotgun dataset (dmin = 3) and (2) have unique "high quality extensions" in the shotgun reads[37]. By definition, a k-mer has unique high quality extensions when every occurrence in the shotgun reads is flanked at both ends by a unique nucleotide, considering only flanking positions with quality score greater than Q=20. Both orientations are considered for each k-mer. Such "UU" k-mers occur in a unique k+2-mer context in the genome, and their single-base extension in either direction is

uncontested. Starting from any such UU 51-mer, we traverse the linear de Bruijn subgraph formed by these uncontested k-mer extensions. These paths correspond the set of initial "UU contigs."

By construction, each k-mer that occurs in the UU contigs occurs exactly once in the UU contig set. Using these k-mers as seeds, reads can be efficiently mapped to the UU contigs by requiring one or more exact k-mer matches (Chapman, Ho, *et al.,* unpublished). From these mappings, scaffolds are formed by a conservative greedy algorithm that requires multiple consistent paired-end links between UU contigs and fewer than pmin = 2 or 3 inconsistent links, taking into account insert sizes bootstrapped by paired ends mapping to the same UU contig. After scaffolding with fragment libraries (insert size ~225 bp, ~425 bp, and ~750 bp), mate pairs (1.5 kb and 4.5 kb), jumping libraries (10 kb), fosmid ends (~35 kb), and BAC ends were used sequentially to form progressively longer scaffolds. Links formed by 1 or 2 fosmid and/or BAC ends that were consistent with *X. tropicalis* synteny were also accepted.

To complete the shotgun assembly, intra-scaffold gaps were closed by using reads that either (1) extend from flanking UU-contigs into the gap (including reads that align to UU contigs on both sides of the gap), or (2) are inferred to lie in the gap based on the placement of their paired-end sequence. Note that the orientations of reads placed in a gap are known. After collecting these two classes of reads, k-mer paths were sought that traverse the gap between flanking contigs. An adaptive choice of k depending on the complexity of reads placed in a gap was used to find unique traversals.

Gap closure allows regions with two-copy 51-mers to be assembled if they are within ~500 bp of single-copy sequence. The resulting gap-closed assembly thus can also capture exon-sized sequences that are highly similar between subgenomes. The resulting gap-closed contigs have N50 length ~20 kb. The sizes of remaining gaps were estimated based on spanning paired ends or mate pairs. A gaps size of ten N's is used the gap cannot be closed or the implied gaps size is <= 10.

In the early stages of the project, we also assembled shotgun data with SOAPdenovo[40]. These assemblies were more demanding computationally than our custom meraculous and produced comparable contiguity. Most gaps in meraculous were also gaps in SOAPdenovo assembly, and contigs often ended at or near the same position as the meraculous assembly, indicating limitations fo the dataset rather than of the algorithms used, but with an increased rate of misjoins (gross assembly errors) in SOAPdenovo. Mapping genome shotgun reads to individually sequenced BACs shows that intra-scaffold gaps typically correspond to regions with genomic copy count >100, accounting for the high copy repetitive sequences.

2.7 Assembly summary

The *X. laevis* assembly 7.1 (Xenla7.1) is summarized in Supplemental Table 1. Assembly 7.1 was the basis for developing the chromosome-scale assembly incorporating BAC-FISH, "Hi-C" chromatin conformation capture from *X. laevis* embryos, and an in vitro analog of HiC[41], along with supporting information based on conserved synteny between *X. laevis* and *X. tropicalis*, as described in Supplementary Note 3.

2.8. Validation of assembled shotgun genome

To assess the completeness of the Xenla7.1 assembly, we aligned 11,515 full length insert *X. laevis* cDNAs obtained from NCBI[35] using BLASTN[42]. 11,472 (99.6%) aligned with better than 98% identity over more than 80% of their length, including gaps in the assembly. The 43 cDNAs that did not align to the assembly were aligned to NCBI NR database using BLAST and found to be contaminants (other vertebrates, or known parasites of frogs) (Supplemental Table 4). Thus the draft assembly captures the annotated expressed genome. Of the 11,472 *bona fide Xenopus laevis* cDNAs, 11,194/11,472 (97.5%) of cDNAs have their entire sequence on a single scaffold. The remaining 278 include genes split across scaffolds in the 7.1 assembly or genes that map to scaffolds with an exon-containing gap where the missing exon is on a short scaffold. These data are being considered to improve future assemblies and annotations.

2.9 Self-consistency of shotgun genome

We mapped reads from the Hi-C experiment (Supplementary Note 3) and other long-range paired-end reads (mate-pair libraries, fosmid-end sequencing data) on the genome using bwa mem (version 0.7.10)[43] with the paired-end option, after filtering reads using a procedure similar to that used in RNA-seq quantification. For each 1,000 bp, we counted the number of paired reads crossing that position (called 'xover-score'), to figure out whether it is misassembled or not (the misassembled region will have few or no crossed paired reads).

We evaluated this method with misjoined scaffolds we identified by manual curation of the version 7.1 assembly, and optimized the parameter to determine the break point, based on 15 misjoins we identified in the v7.1 assembly by comparison with BAC-FISH. We also reduced a long gap (represented by N's) on a scaffold to maximum length of 20 kbp, to prevent the failure of paired-end sequence mapping across the gap. As a result, 1,382 mis-join candidates among 1,285 scaffolds were broken prior to forming the chromosome-scale assembly.

This analysis was also applied to the chromosome-scale assembly to prevent the possibility of reintroducing misjoins. All scripts used in this analysis are available at https://github.com/taejoonlab/HTseq-toolbox/.

2.10 Comparison with finished BACs

A set of 36 BAC clones were sequenced in order to assess the accuracy of the v7.1 assembly. Minor variations were detected in the comparison of the BAC clones and the assembly, but overall the BACs are colinear with the genome assembly. A total of 6 of the 36 clones were found to bridge scaffolds in the assembly and were not included in this analysis. In 17 of the remaining 30 contiguous BAC clones, the alignments were of high quality (< 0.3% bp error) with an overall error rate of less than 1 in 100,000

bp. These clone are presumed to be derived from the J-strain. 13 additional clones have a total discrepancy with the assembly of 7,918 aligned based (including marked gap bases) across 2.03 Mb, or a total discrepancy rate of 0.4%. This is comparable to the polymorphism rate estimated from resequencing. Dotplots (generated using Gepard[44] show that BAC sequences are nearly perfectly colinear with the assembly for the first 17, and highly colinear for the remaining 13 clones, with minor deletions and insertions as expected for an alternate haplotype. See Supplemental Table 1 for further information of sequenced BAC clones.

Supplementary Note 3. Chromosome-scale assembly

3.1 Chromosome assignment with BAC-FISH

FISH and chromosome preparation were performed as described previously[45,46]. Heart, lung, and kidney tissues were taken from adult J strain females (mainly 33rd generation) and used for fibroblast cell culture. Cultured fibroblast cells were harvested after 6 h of treatment with BrdU (25 µg/ml) including 1 h of treatment with colcemid (0.17 µg/ml), suspended in 0.075 M KCl, fixed in methanol/acetic acid (3:1), and spread onto clean glass slides using a standard air-drying method. Replication-banded chromosomes were obtained by exposing chromosome slides to UV light after staining with Hoechst 33258 (1 µg/ml) for 5 min.

We mapped 798 BAC clones containing the 198 homoeologous gene pairs using dual-color BAC FISH (Supplemental Table 1). Two BAC clones were individually labeled with biotin-16-dUTP (Roche Diagnostics) and digoxigenin-11-dUTP (Roche Diagnostics) by nick translation. The labeled FISH probes were ethanol precipitated with 100 times the amount of sonicated female genomic DNA for suppression of cross-hybridization to interspersed repetitive sequences. After hybridization, the biotin- and digoxigenin-labeled probes were stained with FITC-avidin (Vector Laboratories) and rhodamine-conjugated anti-digoxigenin Fab fragments (Roche Diagnostics), respectively. The digital FISH images were captured with a cooled CCD camera (Leica DFC360 FX, Leica Microsystems) mounted on a Leica DMRA microscope and processed using the 550CW-QFISH application program of Leica Microsystems Imaging Solutions Ltd. (Cambridge, UK). Chromosomal locations of FISH signals were assigned according to the Hoechst 33258-stained banding patterns and ideogram of *X. laevis* chromosomes as reported[47].

The BAC-FISH alignment revealed a set of 15 scaffold misjoins in the v7.1 assembly (consistent with paired end analysis described above) that were broken before proceeding to the chromosome-scaled assemblies (v7.2 mentioned below).

3.2 *In vivo* long-range linkage with tethered conformation capture (HiC)

Pair linkages produced by chromatin conformation capture have been shown to be useful in long-range scaffolding, as the vast majority of chromatin contacts are between sequences on the same chromosome arm[48,49]. Tethered chromatin conformation capture was performed as previously described[50], with minor modifications. Briefly, for each experiment, 100 *X. laevis* embryos (stage 10.5, about 10,000 cells) were fixed for 30 minutes with 1% formaldehyde/PBS, washed twice with 0.125 M glycine in PBS, washed twice in PBS, and frozen at –80°C.

Embryos were thawed on ice and disrupted with by vigorous pipetting in 550 µl lysis buffer (10 mM HEPES pH=8.0, 10 mM NaCl, 0.2% IGEPAL CA-630), and 1x protease inhibitors solution (Roche) and pelleted at 11000 g, 4°C for 5 min. Cells were washed twice with 1.5 ml ice-cold wash buffer (50 mM Tris.HCl pH=8.0, 50 mM NaCl, 1 mM EDTA) and resuspended in 250 µl wash buffer and 15 µl 10% SDS. 105 µl 25 mM EZlink Iodoacetyl-PEG2-Biotin (Pierce) was then added and samples rocked at RT for 75 min. Samples were diluted with 650 µl 1x NEB buffer 2, incubated on ice for 5 min, then 150 µl 10% Triton X-100 were added, followed by incubation on ice for 5 min, then at 37°C for an additional 10 min. Chromatin was digested overnight at 37°C after adding 85 µl 10x NEB2, 30 µl 1 M DTT, 200 µl water and 35 µl 25 U/µl MboI (NEB).

The sample was dialyzed against 0.5 l TE pH 8.0/sample for 3 h at room temperature, then another 1 h with fresh 0.5 l TE, in a G2 Slide-A-Lyzer cassette with a 20 kDa size cutoff (Pierce). In the meantime, 400 µl T1 Streptavidin Dynabeads (Invitrogen) were washed 3x with 2 ml 0.01% Tween 20/PBS (0.01TPBS) each and resuspended in 1.5 ml 0.01% Tween 20/PBS. The sample was divided into 5 aliquots in 1.5 ml tubes, 300 µl Dynabeads each were added and the protein-DNA complex collected by rocking at room temperature for 60 mins. Beads were blocked by adding 5 µl of 20 mM biotin solution (15x molar excess over streptavidin) and rotating for 15 mins at room temperature. Beads were washed once with 600 µl each of 0.01% Tween 20/PBS and STRP wash buffer (10 mM Tris pH 8.0, 50 mM NaCl, 0.4% Triton X-100). Beads were resuspended in 100 µl of STRP wash buffer, and overhangs were filled in by adding 63 µl water, 1 µl 1 M MgCl2, 10 µl of 10X NEBuffer2, 0.7 µl of 10 mM dATP, 0.7 µl of 10 mM dTTP, 0.7 µl of 10 mM dGTPαS (AXXORA), 15 µl 0.4 mM Biotin-14-dCTP (Invitrogen), 4 µl 10% Triton X-100, and 5 µl of 5 U/µl Klenow enzyme (Enzymatics) and rotating for 40 min at room temperature. The reaction was stopped with 5 µl 0.5 M EDTA, beads were washed twice with 600 µl each Klenow wash buffer (50 mM Tris pH 7.5, 0.4% Triton X-100, 0.1 mM EDTA).

Beads were suspended in 500 µl Klenow wash buffer, transferred to a 15 ml conical tube, and DNA was ligated under rotation for 4 hours at 16°C in a total volume of 4 ml containing 250 µl 10x T4 DNA ligase buffer (Enzymatics), 180 µl 10 % Triton X-100, 100 µl 1 M Tris pH 7.5, 50 µl 100x BSA (10 mg/ml), 2 µl (1200 U) T4 DNA ligase (Enzymatics). The reaction was terminated by adding 200 µl 0.5 M EDTA (4x molar excess over 5 mM Mg2+). Beads were resuspended in 400 µl extraction buffer (50 mM Tris pH 8.0, 0.2% SDS, 1 mM EDTA, 100 mM NaCl), 20 µl 20 mg/ml proteinase K (Ambion) were added and crosslinks reversed for 8 h at 65°C, and another 5 µl proteinase K added and further incubated at 65°C overnight.

DNA was extracted once with 400 µl phenol/chloroform/isoamylalcohol (25:24:1) tris-buffered to pH 8.0 and once with 200 µl CHCl3, and precipitated for two hours at –80°C with 20 µl 5 M NaCl, 1.5 µl 15 mg/ml glycoblue, and 1060 µl 100% EtOH. DNA was pelleted for 25 minutes at 20000 g, 4°C, and washed twice with 1 ml 80% EtOH for 5 minutes, 8000 g, 4°C. Pellets were dissolved in 25 µl 10 mM Tris pH 8.0 each, and aliquots were pooled and digested for 30 min at 37°C with 1 µl 2 µg/µl RNase A. Reactions

were cleaned up on a Zymo DNA Clean & Concentrator purification column (Zymo) and eluted with 50 µl elution buffer. Five micrograms DNA were treated with 300 U exonuclease III in 90 µl 1x NEBuffer 1 for 1 hour at 26°C and then 37°C, then the enzyme was inactivated with 2 µl 0.5 M EDTA, 2 µl 5 M NaCl and heating to 70°C for 20 minutes.

Water was added to 100 µl and DNA was sheared to 100–500 bp in a Covaris E220 in a 6x16 AFA fiber microtube at 5% duty cycle, intensity 5 (= 175 W), 200 cycles/burst for 180 sec total time. DNA was cleaned up with Ampure or equivalent. The DNA fragment ends were polished with 29 µl water, 10 µl 10x T4 DNA ligase buffer (Enzymatics), 4 µl 10 mM dNTP, and 5 µl (15 U) T4 DNA polymerase, 1 µl (5 U) Klenow, 5 µl (50 U) T4 PNK for 30' at 20°C. After DNA cleanup with Ampure or equivalent and elution into 50 µl Tris, fragments were A-tailed with 5.9 µl 10x NEBuffer 2, 0.12 µl 100 mM dATP and 3 µl (15 U) exo-Klenow enzyme (Enzymatics) for 40' at 37°C.

The reaction was stopped with 1.5 µl 0.5 M EDTA and DNA was captured with 56 µl 2x Bind & Wash buffer containing 0.2% Tween 20 and 15 µl T1 Dynabeads (Invitrogen, washed twice with 1x B&W buffer (5 mM Tris-HCl pH 7.5, 0.5 mM EDTA, 1 M NaCl, then suspended in 56 µl 2x B&W buffer), rotating for 30 minutes at room temperature. Beads were washed once with 500 µl each of 1x B&W/0.1% Triton-X100, once with TE. Sequencing adapters were ligated to the bead-bound DNA in 100 µl 1x rapid ligation buffer (Enzymatics) containing 0.1% Tween 20, 0.5 µl NextFlex adapters (Bioo, 1:20 diluted), 5 µl (3000 U) T4 DNA ligase (Enzymatics) for 20 minutes at room temperature. The reaction was stopped with 6 µl 0.5 M EDTA, beads washed twice with 1x B&W, twice with 0.1% Tween 20/TE, then resuspended in 40 µl 0.033% Tween20/LoTE (TE diluted 1:4 with water). Libraries were PCR-amplified using the 20 µl of the bead suspension as template for 10 cycles, cleaned with Ampure or equivalent to remove adapter dimers, and paired-end sequenced on an Illumina NextSeq.

3.3 Chicago long-range linking library from *in vitro* chromatin

Proximity-ligation libraries from in vitro reconstituted chromatin has been shown to produce linkages of up to ~150 kb suitable for long-range scaffolding. To this end a "Chicago" library was prepared as described previously (Dovetail Genomics)[41]. Briefly, 5.5 µg of high molecular weight genomic DNA was reconstituted into chromatin *in vitro*, and fixed with formaldehyde. Fixed chromatin was then digested with *MboI*, the 5' overhangs were filled in with biotinylated and thiolated nucleotides, and then free blunt ends were ligated. After ligation, formaldehyde crosslinks were reversed and the DNA was purified to remove biotin not internal to ligated fragments. The DNA was then pulled down with streptavidin beads to enrich for biotin-containing fragments and sequencing libraries were generated using established protocols[51]. Sequencing was performed using an Illumina HiSeq with 2x150 paired ends.

3.4 Scaffolding the draft genome with HiRISE

The *Xenopus laevis* draft genome in FASTA format (XENLA_JGIv72; scaffold N50 of 3.47 Mb), and Chicago library sequencing reads in FASTQ format were used as input data for HiRise, a software pipeline

designed for using Dovetail Chicago library sequence data to assemble genomes[41]. In a subsequent pass, TCC pairs were used. At the core of the HiRISE pipeline is a likelihood model that predicts the unique distribution of proximity ligation read pair separations[41].

Chicago library sequences were aligned to XENLA_JGIv72 using the SNAP read mapper (http://snap.cs.berkeley.edu)[52] and marked as PCR duplicates using Novosort (http://www.novocraft.com/products/novosort/). Aligned reads were not penalized for having unusual orientation (Chicago paired-end reads may be on the same or opposite strand), insert size, or for mapping to different scaffolds. Additionally, when a read contained the restriction site junction "GATCGATC," the sequence after the first "GATC" was ignored for mapping purposes.

An initial HiRISE assembly was done with Chicago mate pairs. Of the 176.7 million Chicago read pairs, 84.2 million were aligned such that both reads had map quality score greater than 20 and were not marked as duplicates. Of these high-confidence read pairs, 27.7 million spanned between 0 and 2Kb, 3.7 million spanned between 2 kb and 10 kb, and 2.5 million spanned between 10 kb and 100 kb. The Chicago reads yield 4.1X, 6.4X, and 25.6X effective physical coverage between 0 - 2 kb, 2 kb - 10 kb, and 10 kb - 100 kb, respectively.

The mapped Chicago read pair separations were then scored with the HiRISE likelihood model and these scores were subsequently used to identify and break misjoins in the input assembly. HiRISE was run and produced scaffolds (XENLA_HiRISE_v1; scaffold N50 of 8.98 Mb).

The final shotgun assembly was made with HiRise using TCC data. TCC library sequences were aligned to the XENLA_HiRISE_v1 assembly using the methods described above. 228.2 million of the 547.1 million TCC read pairs were aligned such that both reads had map quality greater than 20 and were not marked as duplicates. Of these high-confidence read pairs, 80.0 million spanned between 0 and 2 kb, 9.9 million spanned between 2 kb and 10 kb, and 7.9 million spanned between 10 kb and 100 kb. The TCC reads yield 7.9X, 19.0X, and 317.6X effective physical coverage between 0–2 kb, 2 kb–10 kb, and 10 kb–100 kb, respectively.

The mapped TCC read pair separations were then scored with the HiRISE likelihood model and these scores were subsequently used to identify and break misjoins in the input assembly. HiRISE was run and produced scaffolds (XENLA_HiRISE_v2; scaffold N50 of 34 Mb).

3.5 Construction of chromosome-scale pseudomolecules

The HiRise Chicago-plus-"HiC" assembly described in Supplementary Note 3.3 was used as the basis for the chromosome-scale assembly of *X. laevis*. To produce chromosome-scale sequences we assigned HiRise super-scaffolds to chromosomal positions based on BAC-FISH results (Supplementary Note 3.1). 132 HiRise super-scaffolds, accounting for 2.52 gigabases of assembled sequence (90.6% of the total assembled sequence) were assigned in this manner. Superscaffolds with multiple BAC-FISH markers at distinct cytological positions were oriented accordingly. Superscaffolds placed by only a single marker, or markers with a single cytological position, were oriented based on synteny and the large-scale chromosomal organization shown in Fig 1.

To identify syntenic blocks we used MCScanX[53] to find colinear blocks of three or more interrupted genes between the *X. laevis* chrL, *X. laevis* chrS, and *X. tropicalis* genomes in theorthologue list generated in Supplementary Note 2.4. We restricted ourselves to these blocks to be certain that the units of synteny would not be subject to the noise of individual elements transposing in the genome. Synteny maps for each L and S subgenome compared to the full-length *X. tropicalis* chromosome were compared to BAC-FISH maps (Fig. 1a) to recapitulate any breaks in the conserved synteny, specifically on *X. laevis* chromosomes 3S and 8S. Custom scripts were used to stitch HiRise super-scaffolds together (with spacers of 10,000 N's between them) to form the *X. laevis* v9.1 chromosome-scale assembly (summarized in Supplemental Table 2).

We also provisionally assigned 56 HiRISE superscaffolds without BAC-FISH markers to chromosomes based on (1) conserved synteny with an orthologous segment of *X. tropicalis*; (2) conserved synteny with a homoeologous segment on *X. laevis* (allowing assignment to L or S according to whether the homoeologous segment was S or L, respectively); and (3) repeat content diagnostic of L or S identity (see Supplementary Note 7).

Supplementary Note 4. RNA-seq transcriptome resources

4.1 Collection of *X. laevis* transcriptome public resources

For annotation we collected 697,015 *Xenopus laevis* EST sequences from a diverse set of cDNA libraries, summarized in Supplemental Table 4, deposited in Genbank by numerous groups (see other references[35,54]). These sequences represent an estimated 13,141 genes (data from NCBI-UniGene, *Xenopus laevis* build 94; assuming one UniGene cluster equals one gene) although many genes are incomplete. EST data and full-length sequences are also available in the *Xenopus* Gene Collection. http://genecollections.nci.nih.gov/XGC/.

We relied on raw EST data for gene annotation rather than the *X. laevis* UniGene clusters because the UniGene clusters were formed without recourse to genomic positions, and contain misjoins that incorrectly splice together homoeologous sequences. Our analysis avoids this by allowing the ESTs and RNA-seq reads to map to their appropriate loci of our complete genome sequence. The ESTs provide a rich resource for the characterization of *X. laevis* genes, and since many libraries were constructed in expression-ready vectors, they also provide an excellent resource for functional experiments with individual clones, or for screening by expression cloning.

Most transcripts were generated from the J strain (Supplementary Note 2), but some come from outbred populations. The degree of polymorphism between these libraries (0.04%) is much lower than the divergence between homoeologous genes (~6% Extended Data Fig. 1d), allowing us to confidently map ESTs from various populations and outbred individuals to their corresponding locus in the assembly.

4.2 Collection of large scale J-strain *X. laevis* transcriptome resources

We complemented the existing EST collection with more than 1 billion RNA-seq reads that sample a useful range of developmental stages and adult organs and tissues, as summarized in Supplemental Table 6. For RNA-seq, RNA was extracted from a series of developmental stages, or from a collection of adult tissues. Both stages and tissue samples were collected twice independently. Embryos from fourteen different developmental stages (including 3 oocyte stages, unfertilized egg, and st8 to NF stage 40, from J-strain 34th generation, cultured at 20 degrees except as noted). Thirteen adult tissues and oocytes of different stages (stages I & II, III &IV, V & VI) were collected from a single female, and a testis was harvested from a single male (J-strain 33rd generation).

Total RNA was extracted using Isogen (Nippon Gene). Quality of the total RNA was evaluated by a spectrophotometer and Agilent 2100 Bioanalyzer (Agilent Technologies). cDNA libraries were constructed using Illumina Truseq RNA sample prep kit V2 (Illumina), with the standard non-strand specific mRNA library preparation protocol. Independent samplings were performed from embryos of two crossings or from two female and male adults, separately, providing cDNA library sets for Taira201203_stage, Taira201203_tissue, Ueno201210_stage, Ueno201210_tissue series of RNA-seqs (see Supplemental Table 6). Additionally, to add reads to Ueno201210_stage (for stage 35), their siblings were analyzed to produce Ueno201302_stage series.

Paired-end (100 bp × 2 101 bp × 2) sequencing was performed using an Illumina HiSeq 2000 instrument (Illumina). Datasets of the short reads were deposited in NCBI Gene Expression Omnibus (GEO) database (accession number GSE73430 for stages, GSE73419 for tissues). These RNA-seq data were used for the expression analysis in this manuscript.

4.3 *De novo* assembly of transcriptome data

For comprehensive genome annotation, we also collected over 283 billion bases of RNA-seq data from the Xenopus research community, mostly from outbred wild-type samples (see 'RNA-seq' page of Supplemental Table 6 for the list of libraries). Since most of these data are part of other studies, especially focused on differential gene expression under certain conditions, we did not analyze their expression patterns. Instead, we used those libraries to construct transcripts with a genome-free, *de novo* assembly approach.

After filtering reads with low quality (either a read containing a no-call, or a read with low complexity mostly from poly-A tail or sequencing errors), we ran velvet[55] (version 1.2.03) and oases[56] (version 0.2.06) to construct transcripts. After running BLASTN for all-against-all comparison for each library, we removed identical or shorter redundant sequences with more than 99% identity.

We reasoned that incorrectly assembled transcripts would not produce highly orthologous proteins, so we translated assemblies in all six frames and mapped the resulting peptides to the proteomes of human, mouse, chicken, zebrafish, and *X. tropicalis* from EnsEMBL database (mainly version 72). We determined the most likely reading frame by a simple voting scheme, and assigned orthologous

gene/protein names in other species. Transcripts that did not matched to these proteome databases were marked as non-coding transcripts, and collected separately.

The final set of assembled transcripts were confirmed on the genome assembly by mapping with GMAP[57], removing the redundant sequences based on their coverage on the genome. More information about assembled transcripts are available at the supplementary website at http://www.taejoonlab.org/index.php/XenopusGenomes.

Supplementary Note 5. Annotation of protein-coding genes and miRNAs

We annotated the protein-coding genes of the *X. laevis* genome using a modified version of the DOE Joint Genome Institute annotation pipeline, which integrates transcriptome data, homology, and *ab initio* methods and has been previously applied to numerous metazoan genomes (see ref[58], for example). RNA-seq and chromatin modification data were used to improve these annotations, especially at the 5' end, using pita (version v1.72. doi:10.5281/zenodo.34942). Prior to annotating genes, *X. laevis* genome sequences were repeat-masked by RepeatMasker[59] using a custom *X. laevis* transposable element database. Details are provided below.

5.1 *De novo* repeat identification and masking

First, we used RepeatScout[60] to detect all fragments of the frog genome coding for proteins similar to catalytic cores of transposases, reverse transcriptases, and DNA polymerases representing all known classes of transposable elements (TEs) collected in Repbase[61]. The detected DNA sequences were clustered based on their pairwise identities by using the BLASTclust algorithm from the NCBI BLAST package(the pairwise DNA identity threshold was equal to 80%). Each cluster was then treated as a candidate TE family, described by its consensus sequence.

The consensus sequences were built automatically based on multiple alignments of the cluster sequences expanded in both directions and manually modified based on structural characteristics of known TEs. We then produced a TE library by merging these consensus sequences with tetrapod TE sequences reported previously in literature and collected in Repbase. Using RepeatModeler[62], we identified genomic copies of TEs similar to the library sequences. These were clustered based on their pairwise DNA identities using BLASTclust. In each cluster, a consensus sequence was derived based on multiple alignment of the cluster sequences.

After refinements of these consensus sequences, the identified families of TEs were classified based on their structural hallmarks, including target site duplications, terminal repeats, encoded proteins and similarities to TEs classified previously[59]. Identified TEs are deposited in Repbase[61].

This final set of repeats were used by RepeatMasker[59] to mask the assembly. Using the previously annotated RepBase set of transposons masked only ~10% of the *X. laevis* assembly. Our more complete *de novo* repeat set masked ~40% of the assembly. This is comparable to the repetitive content of other tetrapod genomes.

Analysis of specific families of transposable element are described below (Supplementary Note 7).

5.2. Protein-coding gene annotation: overview

The v1.8 protein-coding annotation described here, of the v9.1 chromosome-scale assembly, was performed by (1) initially applying the DOE Joint Genome Institute (JGI) annotation pipeline with transcriptome and homology support to the v7.1 shotgun assembly; (2) applying the "pita" annotation pipeline to incorporate RNA-seq and H3K4me3 data on v7.1; and (3) a final round of applying the JGI pipeline with pita predictions as additional input on the v9.1 chromosome-scale assembly, allowing the selection of gene models from multiple options with different levels of support. The v7.1 and v9.1 have the same underlying sequence, with the v9.1 assembly organized into chromosomes. Selected genes were manually reviewed for quality control and correction.

5.3. Initial annotation

The JGI annotation pipeline utilizes transcriptome support, similarity to genes in related species, and *ab initio* methods to predict protein-coding genes. The span of gene loci was identified as overlapping regions of aligned transcriptome and homology data on the shotgun assembly:

- ESTs and cDNAs: In support of gene annotation we aligned 697,015 *X. laevis* ESTs and cDNAs from NCBI to the chromosome-scale *X. laevis* genome assembly Xenla7.1, requiring a minimum of 98% identity and 50% coverage (*X. laevis* PASA)[63].
- Homology: Peptide sequences from *X. tropicalis*, human, mouse, and chicken (UniProt) were used.

Briefly, gene locus spans were defined by the overlap of BLAT alignments of EST and cDNA data and BLASTX alignments of both homology and RNA-seq transcript assembly peptides, with an added extension of 500 bp at both ends of each locus. At each such locus, *X. tropicalis*, human, mouse, and chicken peptides, and RNA-seq transcript assembly ORFs were used as protein templates for both GenomeScan[64] and Fgenesh+[65] gene predictions. These predictions were then merged with EST and cDNA data using PASA[66], which corrects exon-intron boundaries and adds untranslated regions (UTRs) based on transcriptome alignments. The longest ORF predictions at each locus was retained, along with alternate splice isoforms accepted if supported by PASA. This defined the JGIv1.6 annotation(Note that the 7.1 assembly has the same nucleotide sequence as the 9.1 chromosome scale assembly, and differs only by the organization of the sequence into chromosomes).

5.4 Extension of gene models by pita

We used the "pita" software package (van Heeringen *et al*., in preparation; version v1.72. doi:10.5281/zenodo.34942), to integrate information from RNA-seq and chromatin data to improve predicted gene models based on information about promoter location.

Briefly, gene models were generated by combining transcript and predicted gene information from other pipelines with H3K4me3 ChIP-seq. Predicted transcripts from the JGIv1.6 annotation (Supplementary Note 5.2) and mRNA/EST/cDNA sequences from the raw annotations were mapped to the *X. laevis* genome Xenla7.1 using GMAP[57]. All hits giving ≥90% identity were kept. *X. laevis* and *X. tropicalis* protein sequences were downloaded from Xenbase (ftp://ftp.xenbase.org/pub/Genomics/Sequences/NCBI/) and mapped to the *X. laevis* genome using blat[67]. The blat alignments were processed using scipio[68], which corrects intron-exon borders and splice sites. In addition, transcript models predicted by Cufflinks (egg and stage 10.5) were included[69]. Except for those produced by the JGIv1.6 annotation, all single-exon models were removed.

All transcripts that shared at least one identical exon were combined in a transcript collection. A transcript collection was represented as a directed graph of all exons in the collection. All exons longer than 2 kb that were present in just one annotation source were removed. In addition, all splice junctions present in only one or two annotation source(s) were removed if they were covered by fewer than 10 reads and if the number of reads was less than 10% of the mean of number of reads of neighboring splice sites in the transcript.

We called one optimal transcript per collection by calculating the optimal path through the graph based on the following criteria: number of H3K4me3 ChIP-seq reads at the 5'-end (see Supplementary Note 14 for experimental details), level of RNA-seq expression in exons, RNA-seq reads covering splice junctions, length of the longest predicted open reading frame and number of different annotation sources including an exon. The EST and RNA-seq sources used are listed in Supplementary Tables 5, 6. The pita models were used as full-length transcripts in the final v1.8 annotations.

5.5 Final annotation of chromosome scale assembly

The annotation of the chromosome-scale assembly v9.1 was performed with the JGI annotation pipeline as described in Supplementary Note 5.3 with the addition of pita models from Supplementary Note 5.4 treated as an *in silico* set of full-length transcripts. The resulting annotation is referred to as annotation v1.8 (See Supplemental Table 3 for more detail). This is the annotation discussed here and deposited in Genbank.

5.6. Manual validation of gene models

412 gene models of particular interest to the *Xenopus* community were curated to validate the gene models. Validation procedures included comparison of gene models with published reference cDNA sequences deposited in the NCBI Genbank database, analysis of splice junctions, and verification of whether usages of the initiation methionine and termination codons were correct. These analyses confirmed that 396 out of 412 gene models (96%) were accurate. In addition, six of the inaccurate models contained only a minor error that would not have significant effects on RNA-seq analysis nor gene annotation. We conclude that the vast majority of gene models predicted in this study (~98%) are appropriate for further analyses.

5.7 Annotation of microRNAs

microRNA (miRNA) precursor sequences were identified by aligning experimentally-confirmed *X. tropicalis* miRNA precursor sequences to the *X. laevis* genome via BLASTN with E-value cutoff 1e-10. The highest percent-identity of each unique sequence per subgenome was chosen as the (co)-ortholog. In all intergenic cases, both homoeologous miRNA sequences showed evidence for expression of their flanking primary-miRNA sequence.

Due to the high degree of similarity between homoeologous precursor miRNAs we could not rely on small RNA sequencing to confirm expression of both homoeologues. We therefore confirmed the expression of intergenic miRNAs by querying the RNA-seq alignments to the genome, looking for reads aligning +/- 1kb of the primary-sequences. To confirm that these alignments were not background, they were compared to 10,000 randomly chosen regions from the genome to confirm they were expressed more often and at a higher level (Extended Data Fig. 5e).

Supplementary Note 6. Chromosome evolution

6.1 Large-scale genomic rearrangements

To elucidate large-scale genome rearrangements within *X. laevis* and between *X. laevis* and *X. tropicalis*, we compared BAC-FISH data for *X. laevis* homoeologous chromosomes (both XLA_L and XLA_S) in this study with a previous study of cDNA FISH data for *X. tropicalis* chromosomes (XTR)[70]. As shown in Fig. 1a, overall syntenies are well conserved between chromosomes of XTR, XLA_L and XLA_S, and no gross translocations were detected. Several intra-chromosomal inversions, however, were found as indicated by arrow bars. Notably, most of the inversions probably occurred in XLA_S chromosomes, because gene orders on XTR and XLA_L chromosomes are largely conserved in these regions. Detailed synteny analysis

for some of the inverted regions was then performed manually. We describe here two large inversions found in XLA3S and XLA8S.

In XLA3S, a large inversion covers almost the entire region of its q arm. Comparing with XLA3L, the proximal break point (3q-*prxBP1*) in proto-XLA3S corresponds to the 152 kb region between *sra1* and *slc35a4* genes in XLA3L, while the distal break point (3q-*disBP2*) in proto-XLA3S corresponds to the 23 kb region between *tecr* and *Xelaev18018955m* genes. The 3q-*prxBP1* is probably near a large gene cluster of *pcdhg* with more than 30 copies, whereas the 3q-*disBP2* was near a large cluster of olfactory receptor (OR) genes.

Large-scale rearrangements seen in XLA8S looks complicated, but can be explained by considering that two rounds of inversions occurred in the ancient XLA8S that was initially colinear with XLA8L and XTR8. Based on the gene synteny analysis, we speculate that the first inversion occurred in the p-arm of proto-XLA8S corresponding to the 8.2 kb region between the *Xelaev18038105m* and *Xelaev18038106m* genes in XLA8L and in the q-arm of proto-XLA8S corresponding to the 106 kb region between the *hectd1* and *arhgap5-like* genes. The second inversion possibly occurred in the 16 kb region between the *gata1* and *Xelaev18038129m* genes and in the 15 kb region between the *timm50* and *dlc* genes.

6.2 The fusion of homologs of XTR 9 and 10

Two *X. tropicalis* chromosomes (XTR9 and XTR10) correspond to the single homoeologous chromosome pair, XLA9_10L and S (Fig. 1a)[46]. We investigated in detail the colinearity between XTR9 and 10, XLA9_10L, and XLA9_10S using MCScanX with default values[53]. The positions of centromeres were estimated by BAC-FISH and positions of frog centromeric repeat-1[71] in XLA9_10L and S and cDNA-FISH and p-/q-arm ratio in XTR9 and 10[70]. As a result, the prospective junction by fusion of ancestral chromosomes homologous to XTR9 and 10 in XLA9_10L and S was determined to fall between regions *rpl13a* to *rps11* on one side and *lypd1* to *actr3* on the other, which are syntenic to the terminal regions of XTR9q and XTR10q, respectively (Extended Data Fig. 2b, upper panel). Furthermore, *X. laevis* genes on either side of this junction correspond to discrete syntenic blocks of human or chicken (Extended Data Fig. 2b, lower panel), supporting the idea that the ancestral chromosome of XLA9_10 originated from chromosome fusion before divergence of the L and S progenitors.

The gene order of XTR9 is highly conserved in XLA9_10L and S, whereas the gene order in the pericentromeric region of XTR10 is different from XLA9_10L or S (Fig. 1a; Extended Data Fig. 2b lower panel)[46]. These results suggest that XLA9_10L and S resulted from telomere-to-telomere tandem fusion followed by inactivation of the centromere on the ancestral XTR9 portion and large pericentric inversions on the ancestral XTR10 portion.

Two processes of chromosomal rearrangements (fusion and inversion) that occurred between the hypothetical proto-XTR9 and 10 to produce proto-XLA9_10, and eventually XLA9_10L and S can be hypothesized (Extended Data Fig. 2d). The two models differ in the organization of the proto-XTR10 and the timing and location of the pericentric inversions. In the first model, a tandem fusion occurred between the proto-XTR9 and 10, and the centromere derived from the proto-XTR9 was inactivated. A

large pericentric inversion happened in this fused proto-XTR9_10, leading to the production of a fused proto-XLA9_10 before allotetraploidization (duplication of proto-XLA9_10). In this scheme, the present XTR9 and 10 resembles the proto-XTR9 and 10. The alternative process is that the proto-XLA9_10 was formed from tandem fusion between proto-XTR9 and 10, followed by inactivation of the centromere derived from the proto-XTR9, and then allotetraploidization occurred. Pericentric inversions of the proto-XTR10 occurred during evolution to produce the present XTR10, while XTR9 retains the ancestral structure.

6.3 Analysis of the *X. laevis* sex locus

Sex determination in *X. laevis* follows a female heterogametic ZZ/ZW system[72]. The female-determining gene *dmw*, a truncated paralog of *dmrt1*, is located in the q-subtelomeric region of chromosome 2L[73]. We fully sequenced BAC clones representing both W and Z haplotypes, with or without *dmw*, respectively, and identified both W- and Z-specific sequences (Extended Data Fig. 2a). The existence of the Z-specific sequence was unexpected and therefore verified by PCR analysis using specific primer sets and DNA from gynogenetic frogs having either W or Z locus. The homoeologous XLA2Sq has no such locus. The W-specific region harbored the *dmw* gene. We also found two genes, *scanw* and *ccdc69w*, in the W-specific region and one gene, *capn5z* in the Z-specific region. The synteny analysis of chromosome 2s in *Xenopus* indicated that the W- or the Z-specific region was inserted into the region between OR genes and *cdk4.L* in the proto-chromosome 2L. W and Z chromosomes were defined as chromosomes 2L possessing W- and Z-specific regions, respectively.

Supplementary Note 7. Subgenome-specific repeats

7.1 Initial identification of subgenome-specific transposable elements

RepeatMasker output was used to calculate the total coverage length (bp) of each repeat family on each scaffold (Xenla7.1 assembly). For each repeat family, a scatter plot was drawn to show the correlation between the total length (bp) of scaffold (x-axis) and the coverage length of the repeat family on each scaffold (y-axis) using R. If a repeat family shows a uniform distribution across the genome, the coverage length will, on the whole, positively correlate with the scaffold length. In contrast, if a repeat family is specific to one subgenome, the positive correlation will be seen in about only half of the scaffolds, while the repeat family will be almost absent in the rest. Repeat families showing uneven distribution on the scaffolds could be subgenome specific and were further analyzed as described below.

The v7.1 scaffolds that were assigned to specific chromosomes by BAC-FISH were collected and used to calculate the approximate density of these repeats on each chromosome. The density was compared between homoeologous chromosomes known from BAC-FISH (e.g., 1L vs. 1S) to confirm specificity of the repeats to one of the homoeologous chromosomes. By this approach, each unevenly distributed

repeat family was confirmed to be specific to either L-chromosomes (type L) or S-chromosomes (type S). This result demonstrates that the L and S chromosome sets correspond to two subgenomes originating from two distinct progenitor species.

The repeats, corresponding to partial fragments of subgenome specific transposons, were used to identify the full-length L- or S-specific transposon sequences as follows. Each consensus sequence of type L or type S repeat was used as a query for a BLASTN search. HSP (high-scoring segment pair) sequences were collected with their flanking sequences and they were compared by multiple alignment to identify the range of sequence similarity. The longest such aligned range was assumed to correspond to the full length of a type L or type S transposon. All sub-genome-specific type L or type S transposons were found to be DNA (class II) transposons. They were then classified by their target site and the terminal inverted repeat (TIR) sequences.

The L-specific fragments were found to be partial sequences of a miniature inverted-repeat transposable element (MITE) family whose TIR starts with 'GG' and whose target site duplication (TSD) sequence is TTA (TAA). This family likely belongs to the PIF/Harbinger superfamily and is therefore named Xl-TpL_Harb. One of the S-specific fragments was similarly found to be a partial sequence of a DNA transposon family whose TIR starts with 'GG' and whose target site is TTA (TAA), but whose internal sequence is distinct from Xl-TpL_Harb. This family also probably belongs to the PIF/Harbinger superfamily and is therefore named Xl-TpS_Harb. Other S-specific fragments were identified as partial sequences of an autonomous element belonging to the Tc1/mariner superfamily (tentatively named Xl-TpS_Mar). Consensus sequences for these elements were curated manually.

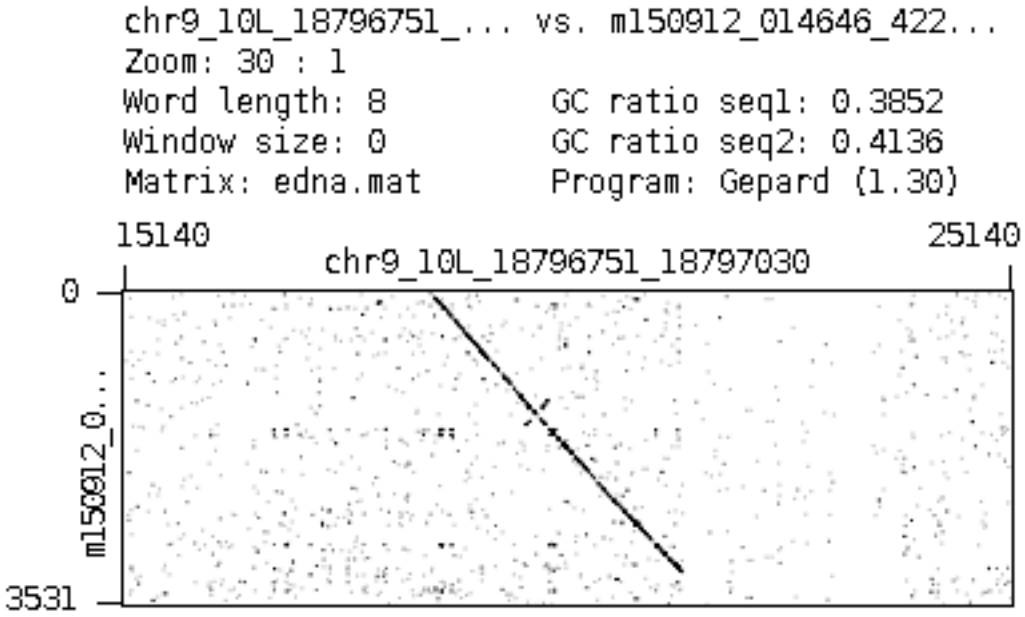
In addition to these three manually curated families of transposable elements, we found additional RepeatModeler-defined TEs that were enriched in L or S. Repetitive genomic loci of at least 400 bp defined by the RepeatModeler library were considered and their total counts were obtained. L has a slightly larger absolute amount of transposons (due to its larger size). We therefore defined repeat families enriched in S by the criterion S/(L+S) >= 0.55, resulting in 33 families. Similarly, L-enriched elements were defined as having L/(S+L) >= 0.65 (resulting in 28 families). The most prominent sub-genome-enriched TE families are Harbinger and Mariner elements. Their distribution per subgenome is detailed in Extended Data Fig. 3 based on subfamily refinement described below. Since RepeatModeler is effective at merging related elements into a single consensus based on a relaxed cutoff, estimates of the timing of repeat activity used sub-family consensus sequences as a proxy for the founding element.

7.2 Distribution of subgenome specific elements

The coverage lengths of the subgenome specific transposons on each chromosome (Xenla9.1) were calculated from the results of blastn search (E value less than 1E-5) using the consensus sequences of Xl-TpL_Harb, Xl-TpS_Harb, and Xl-TpS_Mar as queries (Extended Data Fig. 3a). Because Xl-TpL_Harb and Xl-TpS_Harb share a common 17 bp TIR sequence, this region was removed from the queries.

7.3 Validation of cross mapped subgenome-specific repeat subfamilies

We queried the alignments in Fig. 1b for the occasional subgenome-specific loci which cross-mapped to the opposite subgenome to ask if there was evidence for localized misassembly. There were 24/6,543 occurrences of the L harbinger element on scaffolds assigned to the S subgenome, 35/1,820 occurrences of the S harbinger element on L scaffolds, and 21/5,008 occurrences of the S mariner element on L scaffolds. We analyzed these 80 regions in question using the initial genome surveys with PacBio data (data available on request), and verified that 62 of these regions have supporting PacBio coverage. Our basic approach was to pull 100 mers from either side of the regions and align the 100 mers to the entire PacBio dataset (~10x raw coverage). PacBio reads with a sufficient number of 100mers aligning to them from both sides of the region were selected. We then extracted each region +/- 20 kb on either side and aligned the PacBio reads to these smaller regions. The use of the 40 kb regions facilitates the rapid creation of dot plots, which show the continuity across the region.



The remaining 18/80 cases of "L-specific" elements assigned to S chromosomes, and *vice versa*, are either (1) *bona fide* transpositions between L and S chromosomes after allotetraploidy, (2) localized assembly errors, or (3) misassignments of genomic loci to the L- and S-specific elements hindered by accumulated mutations. We note that these localized regions are dwarfed by the >10,000 cases of L-enriched elements found on L and S-enriched elements found on S, and do not affect our overarching conclusion that L- and S- subgenomes have maintained their integrity since allotetraploidization, and that these subgenomes can be recognized by their transposable element complement.

7.4 Transposable element chromosome FISH

To confirm the chromosomal distribution of the S-specific mariner transposon Xl-TpS_Mar we performed chromosome mapping with mariner-specific probes specific to this element using FISH. FISH probes were prepared by PCR-amplifying three copies of the Xl-TpS_Mar transposon from the genomic DNA of the J-strain and cloning the products in the pBluescript II SK(-) vector. Cell culture, chromosome preparation and FISH were performed as described in Supplementary Note 3.1, with modifications of labeling of the FISH probes and staining with the antibody. The DNA fragments were labeled with digoxigenin-11-dUTP (Roche Diagnostics) by nick translation. After hybridization, the slides were incubated with rhodamine-conjugated anti-digoxigenin Fab fragments (Roche Diagnostics). Presence of the Xl-TpS_Mar was confirmed on S-chromosomes, and no signal was detected on L chromosomes, confirming the rarity of this S-specific element on the L subgenome (Fig. 1c).

7.5 Timing of L- and S-specific and enriched transposable element activity.

As described in Supplementary Note 1.8, we reasoned that in an allotetraploid scenario the activity of subgenome-specific elements should be limited to the interval where the L- and S-progenitors were evolving independently (and therefore unable to exchange transposable elements). To test this hypothesis, we estimated the age of TE relicts on the genome by comparing the extant mutated sequences against their consensus sequence, which approximates the original active element. This analysis focused on the three subgenome specific elements described above: Xl-TpL_Harb, Xl-TpS_Harb, and Xl-TpS_Mar but also included other enriched elements.

Subfamilies were computed manually as follows. We aligned all sequences within each family using MAFFT[74], then constructed phylogenies using FastTree[75]. Such phylogenies (sample for TpS-Mariner shown in Extended Data Fig. 3c) reveal multiple 'star' topologies corresponding to individual subfamilies. Each 'star' represents a single ancestral TE that was active at a defined time. Only 'star' topologies with at least 80% of either S or L copies were considered further. In total, this identified 13 S-specific subfamilies (coming from 6 RepeatModeler consensus sequences) and 37 L-specific subfamilies (coming from 8 RepeatModeler consensus sequences) with at least 10 copies of 400 bp and longer. The repeat families that include these subfamilies are Harbinger and Mariner, and their subgenome-specific enrichment is illustrated in Fig. 1b and Extended Data Fig. 3a.

To infer the timing of repeat activity we computed a sub-family consensus for each star and measured the substitution distance between genomic element and their respective consensus. We aligned the sequences within manually identified 'stars' using MAFFT, then trimmed the alignment to keep positions that show more than 80% sequence coverage in at least 5 nt blocks, and removed CpG positions. Consensus was constructed based on this filtered alignment and the distance to it was computed for each sequence in the alignment by simply counting substitutions in gap-less regions. The distances were then adjusted for multiple substitutions according to the Jukes-Cantor formula[76] and histograms were plotted for the three major S and L specific repeat families (Extended Data Fig. 3b). As expected, the

median values for distance-to-consensus are half the median value of pairwise distances of extant sequences (data not shown). This test suggests that the estimates are not biased by consensus building.

The L and S-specific Harbinger MITEs have a median divergence to consensus of 0.108, and the S-specific Tc1/Mariner element has a median divergence to consensus of 0.054. To convert from substitution distance to absolute time we used the substitution rate of 3~3.2 x 10^{-9} substitutions per year estimated from Note 8, which was derived from synonymous substitutions in protein-coding genes. These divergence values suggest that the Harbinger MITE was active about 34~36 MYA and mariner about 17~18 my ago. The similarity of the Harbinger activity to the L-S progenitor speciation time raises the possibility that this element could have been activated by the speciation process. The Mariner element activity is close to the time of allotetraploidy, but its activity was confined to the S-progenitor. It is tempting to speculate that the active Mariner element in the S-progenitor could have put that species at a disadvantage in the immediate aftermath of tetraploidization. See Extended Data Fig. 3 for the distribution of S and L enriched transposon families on each chromosome, and a distribution of their distances.

7.6 Global analysis of *Xenopus* repetitive element ages

To obtain age estimates for all repetitive element classes, we conducted automated subfamily identification. Starting from the FastTree trees produced as described above, subfamilies were identified by proceeding from the leaves of the tree and merging nodes if the node with the higher count of sequences (tolerance allowance of 2 genes) had a smaller average branch length (measured from all the sub-nodes to the leaves, tolerance of 0.05 substitutions per site). Each identified subfamily was required have at least 10 members for further analysis. This coarse procedure allows automated labeling of stars. The number of identified stars correlates with the total sequence number in the initial alignment, on average with 10 subfamilies (with at least 10 members) per 200 members.

This method allows us to re-compute L and S divergence timings, independent of the protein coding sequences used in the calibration of divergence to geological time (Fig. 2a, Supplementary Note 8). This is important since the transposable elements have different rates of sequence change than protein-coding genes. We set a cutoff for subgenome specific elements when more than 80% (e.g., 9 out of 10 elements) occur on one subgenome. The distances to consensus were computed as above. The median for S was 0.063 (20-21 mya using the substitution rate of Note 8), while for L it was 0.078 (24-26 mya), suggesting that L specific elements are older on average.

This method also allows us to look at the L/S shared elements to check for the signal for L/S divergence. We selected ‘old’ elements present on both L and S in at least 10 copies each and where the oldest elements on both L and S have a larger corrected Jukes-Cantor distance than 0.2. The results identify a valley in the distribution of distances around 0.1, corresponding to the L/S speciation distance.

Supplementary Note 8. Phylogeny, divergence times, and evolutionary rates

8.1 Identification of orthologous and homoeologous protein-coding genes

To identify orthologues of *X. laevis* genes in *X. tropicalis* we used the BLASTP algorithm from the BLAST+ package[77] with a Smith-Waterman refinement and an E-value cutoff of 1E-10. We accepted alignments with matches of at least 80% identity and covering at least 50% of the length of the *X. laevis* query. The highest percent identity alignment within 90% of the maximum BLAST bit score is chosen as the *X. tropicalis*orthologue to a given *X. laevis* protein. We only accepted *X. tropicalis* loci with 1 or 2 *X. laevis* (co)-orthologues (also called homoeologues) by these criteria. Finally, we only accepted *X. laevis* homoeologues whose synteny and subgenome identity agree with the BAC FISH map, resulting in ~15,613/22,718 (69%) *X. tropicalis* protein-coding loci (including those on scaffolds) available for analysis.

Over 1,000 *X. tropicalis* loci with 3 or more loci aligning could be separated into 3 classes. (1) The earlier annotations masked with RepBase contained a number of transposon sequences whose homologous subfamilies were not masked in *X. laevis*. This class is defined by not having a clear syntenic ortholog, and the homologs aligned to many different sequences across their entire length. (2) *X. laevis* loci where one or both genes are fragmented compared to their *X. tropicalis* ortholog. We are working with the *Xenopus* community to properly annotate these loci. (3) *X. laevis* loci that have had a tandem duplication following the speciation from the *X. tropicalis* ancestor. Chordin is an experimentally-validated example of this type (Extended Data Figure 10a). While it would be interesting to study all of the tandem duplications of *X. laevis*, we must first classify the first two groups to be sure that we have a confident list for the third.

24,419 *X. laevis* protein-coding genes can be placed in 2:1 or 1:1 correspondence with 15,613 *X. tropicalis* genes, defining 8,806 homoeologous pairs of *X. laevis* genes with *X. tropicalis* orthologues, and 6,807 single copy orthologues. The remaining genes are members of larger gene families whose *X. tropicalis* orthology is more complex.

8.2 Comparison to previous estimates of gene retention

	Morin	v9.1
Total	8,049	45,099
Singletons	1,548	6,809
Homoeologue pairs	4,535	17,612
No *X. tropicalis* hit	118	6,864

Complex	1,848	13,814

cDNAs from previous study[35] (referred 'Morin cDNAs') were used as full-length transcripts in the v1.8 (final) genome annotation, and all are present in the assembly. Morin et al. found 1,039 homoeologue pairs after sampling 8,049 *X. laevis* genes (1,039 homoeologues out of 7,010 originally duplicated loci = 14.8% retention). This is much lower than our present estimate and other estimates from *X. laevis* ESTs using the *X. tropicalis* gene set[36]. This underestimate of homoeologue retention by Morin *et al.* is likely due to bias towards genes present in the limited number of genes considered, and incomplete capture of homoeologues in their cDNA libraries.

To assess this effect, we calculated homoeologue andorthologue retention rates for the Morin cDNAs using the full genome sequence and annotations of both *Xenopus* species. We find that the genome-wide retention rate of the Morin cDNA set is much higher than their result (75% using homoeologues found in the genome, vs. 14% using those only found in ref[35]). The 75% retention rate for the Morin et al. genes is higher than our genome-wide estimate (56%), likely due to their bias towards more highly expressed genes. NCBI clones were selected by 5' EST sequencing[78], and higher expressed genes will be more likely to appear in the EST set. We have shown that retention rate depends on expression level, as also found in the analysis of the *Paramecium* genome duplications (Extended Data Fig. 8b)[79].

Similarly, Morin *et al*. found limited evidence for GO terms enriched in singletons (which overlap with ours, Supplemental Table 5), but could find no such enrichment in retained homoeologue pairs as we could with the whole genome sequence. This is in part due to limited power to detect these enrichments when using only a relatively small number of genes. These analyses illustrate the importance of a full genome sequence in understanding the evolution of a polyploid organism.

Peshkin *et al*.[80] identified 164 putative pairs of homoeologous genes via proteomic analysis, relying on peptides that differ at a single amino acid position to infer homoeologuey. For this limited set they found correlated expression between these peptides. Since chromosomal position was not assessed, these may or may not be true homoeologues, as tandem duplicates might also differ by a single amino acid substitution.

8.3 Annotated sequence alignments

The longest transcript from each protein-coding locus was chosen for alignments. CDS alignments between *Xenopus* homologs were done using the MACSE package using default parameters on the longest ORF of each sequence[81]. The CDS sequence content and evolutionary rates were calculated using the seqinR package[82]. The calculation of subgenome-specific rates is explained in Extended Data Fig. 1. We used default two-tailed Wilcoxon-Rank sum test in the R package to determine statistically significant differences between mutation rates. *X. tropicalis* chromosomal locations were determined by the placement of the *X. tropicalis*orthologue on the v9 map.

The alignments of pvCNEs were done using MUSCLE[83]. The alignments of all elements were concatenated. Gaps were removed from the alignment using Gblocks[84] and a neighbor-joining tree was generated using MEGA6[85] with 1,000 bootstraps, the Kimura 2-parameter model, and uniform rates among sites. The evolutionary rates of pan-vertebrate aCNEs were compared for every pair of tetrapods using Tajima's relative rate test with elephant shark as outgroup.

8.4 Phylogeny

For species trees, we first concatenated alignments, then used Gblocks[84] using default parameters for a codon alignment to identify large blocks of conserved sequence for comparison. The concatenated alignments were fed to PhyML[86]. We used the GTR model of nucleotide evolution, and bootstrapped 10,000 times (otherwise default parameters). For r8s[87] we used the penalized-likelihood method, with the truncated Newton algorithm to compute the time trees. The collection of other frog transcriptomes, and identification of their best hits in *X. tropicalis* is detailed below.

Rana pipiens transcripts were downloaded from the *Rana* transcriptome database[88]. We compared the proteins to our *X. tropicalis* proteome via BLASTP, and found 14,175 clear orthologues, 7,109 of which were homoeologous in *X. laevis* and used for alignment.

896 *Pipa carvalhoi* ESTs were downloaded from Genbank and aligned to our *X. tropicalis* proteome via BLASTX. Only the longest EST per *X. tropicalis* protein was accepted. 62 *X. tropicalis* proteins with *X. laevis* homoeologues were used for alignment.

Hymenochirus boettgeri transcripts were kindly provided by Rebecca Heald. The raw reads and initial Trinity assembly output are deposited at GEO (the accession number GSE76089). We used BLASTX to align the transcriptome of *Hymenochirus* to the proteome of *X. tropicalis*. 13,844 *X. tropicalis* proteins has a single *Hymenochirus* transcript covering > 80% of their length. 6,664 of these had 2 homoeologues in *X. laevis*, and were used for alignment.

65,522 nucleotides were used in the 6-way alignment between the above species, and *X. tropicalis*, and *X. laevis* L and S. After Gblocks was run using default parameters, 14,142 positions were left in 112 selected blocks to be used to build a tree in PhyML. The bootstrap consensus tree was exported to r8s, where we used a calibration point of 102Mya for the *Pipa- Hymenochirus* divergence[89]. We set a smoothing parameter of 0.1 based on the r8s cross-validation method for this tree.

Xenopus borealis genomic shotgun reads have been submitted to the SRA (accession number SRP070985). Reads were aligned to XENLA9.1 using bwa mem (version 0.7.6a)[90], default parameters. Depth was calculated using samtools depth -q 0 -Q 0 -l 41 (base and mapping quality of at least 0, aligned fragment length at least 41 bp). Variants were called using the GATK (version 3.3-0)[91] HaplotypeCaller walker, requiring a minimum mapping quality score of 25, and limited to sites with 3 or fewer haplotypes in the population. Regions annotated as repetitive were excluded from calling. GATK CallableLoci was run with similar parameters. We identified *X. laevis* genes covered by *X. borealis* reads with a minimum depth of 5 (as determined by callable loci) X. borealis sequences were inferred using GATK's AlternateReferenceMaker, treating *X. borealis* reads as variants of *X. laevis*. *X. borealis* consensus

sequences for 449 2-copy *X. laevis* homoeologue pairs fit these criteria. We aligned these to their (co)-orthologous sequences in *X. laevis* and *X. tropicalis*. 363,903 positions in the original alignment were exported to Gblocks, where 329,682 positions we conserved in 424 blocks and used to build a tree in PhyML. We used the *Xenopus* calibration points from the pipid tree described above to calibrate this tree in r8s. We set a 0.1 smoothing parameter based on r8s cross-validation method.

To determine the epoch-specific rates of subgenomes, the individual 5-way protein alignments were exported to R to use the seqinR pakage to compute the pairwise Ks and Ka. Comparison of the pairwise measurements allows us to partition rates of substitution between subgenomes and time periods.

8.5 Estimate of substitution rate

We used the species divergence times estimated above, and synonymous substitution levels between *Xenopus tropicalis* and *X. laevis* orthologues, and between *X. laevis* L and S homoeologues, to estimate an absolute substitution rate. In these calculations, CpG sites were consistently excluded from consideration. When using this calibration to estimate the timing of transposable element activity, we also excluded CpGs to provide comparable substitution contexts.

The median *X. tropicalis*-*X. laevis* Ks of 0.286 was measured with the Nei-Gojobori method[92] as implemented by yn00 program in PAML[93]. The Nei-Gojobori method was used because its model most closely parallels the Jukes-Cantor model[76] that we used to measure the divergence of transposable elements from each other. Since we estimate the divergence time between *X. tropicalis* and *X. laevis* to be 48 Myr, this Ks estimate implies an absolute substitution rate of Ks(trop-laevis)/2T(trop-laevis) = 3.0 x 10^{-9} per year.

Similarly, we have estimated the divergence of the L and S subgenomes of *X. laevis* to have occurred 34 Mya, and find a median Ks of 0.218 between homoeologues. This implies a substitution rate of Ks(LS)/2T(LS) = 3.2 x 10^{-9} per year.

8.6 microRNAs

When multiple members of a miRNA family aligned to a single *X. laevis* locus as similar percent identities, synteny of flanking protein-coding genes was considered to determine orthology. With the exception of mir-427, a miRNA known to occur in tandem arrays that are difficult to assemble[94], 156/180 (85%) miRNA gene precursor sequences are retained in both homoeologues. The high degree of similarity between their homoeologues makes it difficult to confirm expression of both copies through small-RNA sequencing, which can only isolate the precursor sequence. While the primary sequences between miRNA homoeologues are divergent enough to distinguish reads between the two copies, they have a short half-life, making them difficult to sequence across their length. RNA-seq data may obtain small fragments of the polyadenylated primary sequence present in each stage. We queried our RNA-seq data for alignments +/- 1 kb of the intergenic precursor-miRNA sequence to confirm expression of primary

sequence of both homoeologues. All duplicated intergenic miRNA pairs show reads aligning to the flanking DNA of both copies; this rate is significantly higher than randomly chosen 2.1 kb segments of the unannotated genome (Extended Data Fig. 5e). We cannot confirm expression of homoeologous intronic miRNAs because it is difficult to distinguish their expression from that of their host genes.

8.7 pan-vertebrate conserved elements

To assess the conservation of ancient non-coding elements, we examined 557 previously identified pan-vertebrate conserved non-coding elements (pvCNEs)[95]. We found that 555/557 of the pvCNEs are present in the *X. laevis* assembly. We could not identify the remaining two pvCNEs by our methods in either *X. tropicalis* or human genomes, likely due to our simpler identification method. Of the 555 elements detected, 533 (95.6%) are present in two copies in *X. laevis*. 5 pvCNEs were not present in the latest v9.0 version of *X. tropicalis* but were present in earlier assemblies, and are not considered in the analysis of homoeologous retention analysis of pvCNEs. We aligned the published human sequences to the elephant shark genome by the same megablast parameters in Lee *et al*[95]. The elephant shark sequences were then used to identify the pvCNEs in different tetrapods. Non-*Xenopus* tetrapod genomes are from EnsEMBL build 77.

8.8 Estimate of nucleotide diversity (polymorphism within *X. laevis*)

We used ~15x 2x100 bp shotgun sequence from a clutch of wild-type embryos to estimate the single nucleotide polymorphism rate in *X. laevis*. This dataset was produced as a control for ChIP-seq described in Supplementary Note 14. Briefly, paired-end shotgun sequence was aligned to the chromosome-reference with BWA-MEM[90] with default settings for short reads. Single nucleotide polymorphisms were called by GATK, requiring a minimum of 10x total coverage and 4x coverage of the alternate allele. Sites with low genotype quality or high depth (>40x) were ignored. The genome-wide SNP rate was estimated to be 0.5%, with lower rate in coding regions (0.4%).

8.9 Whole-genome alignment

Whole-genome alignments were performed using CACTUS[96]. The *X. tropicalis*, *X. laevis* ChrL, and *X. laevis* ChrS genomes were analyzed as distinct species, using default parameters. Each set of masked orthologous chromosomes placed by BAC-FISH was fed to CACTUS to reduce the computational load of aligning non-homologous chromosomes. We filtered alignments for those >50 bp in length, present once and only once in *X. tropicalis*, and at most once in either or both subgenomes of *X. laevis*. We merged elements within +/- 25 bp of each other to assess the question of retention of non-coding elements between *Xenopus* genomes.

We concatenated CACTUS alignments and removed gaps using Gblocks. Trees were built using the R *ape* package[97], and significance of branch lengths computed by a Tajima's relative rate test on the final concatenated/ungapped alignments. This analysis reveals that the S subgenome CNEs are mutating faster than L. Conserved non-coding elements (CNEs) within +/- 100kb of a gene are assigned to that gene as its "neighboring landscape." If two protein-coding genes are within 200 kb of one another, the intergenic distance is halved and CNEs are assigned to the nearest gene.

Supplementary Note 9: Unitary pseudogenes

Unitary pseudogenes are genes that have become nonfunctional and decayed in place, in contrast to retroposed or processed pseudogenes that have been transposed to a new location[98]. We reasoned that some missing homoeologues not detected as *bona fide* genes would be present in the *X. laevis* genome as unitary pseudogenes. To search for pseudogenes we looked for triples of consecutive genes in both *X. tropicalis* and one *X. laevis* subgenome for which the central gene of the triple was missing in the other *X. laevis* subgenome. 1,547 genomic loci fit this pattern. At 984/1,547 (64%) of these loci we found a recognizable unitary pseudogene by aligning the *X. tropicalis* peptide to the *X. laevis* genome using Exonerate[99], which performs peptide-to-genome alignment in a manner that allows frameshifts and internal stop codons.

To infer the approximate age of each pseudogene, we used an approach developed previously[98,100–102] that depends on the excess of non-synonymous substitution in a pseudogene relative to anorthologue of known divergence time. This excess non-synonymous substitution is proportional to (1) the time since non-functionalization and (2) the difference between the synonymous and non-synonymous substitution rate for the gene.

For each synteny-confirmed pseudogene we aligned the genomic sequence identified by Exonerate to the corresponding pair of orthologous *X. tropicalis* (T) and *X. laevis* (L) genes. These triplets of sequence were aligned at the peptide (and codon) level using MACSE, which accounts for frameshifts and stop codons[81]. If we assume that the amino acid changing substitutions Ka in the pseudogene lineage evolved according to the same Ka/Ks ratio as that measured between the functional genes (T and L) up until the pseudogenization time T* and at the synonymous rate Ks afterwards (reflecting the relaxation of selection), we can express the pseudogenization time, in units of Ks, for the pseudogene in each triplet[98,101,102].

$$T^* = (KaLP - KaTL/KsTL * KsLP) / (1 - KaTL/KsTL)$$

The values of KsLP and KsTL are here considered constants, equal to the average observed values from the all the aligned triplets, 0.24 and 0.201, respectively. From these alignments we extracted gap free blocks flanked by fully conserved amino acids and allowing at most two consecutive non-conserved amino acids to avoid aligning non-orthologous sequence. Ka and Ks between all such concatenated codon pairs for each gene were estimated using seqinR[82] which also outputs the variance on the estimates.

To estimate the standard deviation on the estimate of T*, for each gene we drew a sample of 10,000 normally distributed KaLT and KaLP values with mean and standard deviation as found by seqinR, evaluated the corresponding T* values and finally computed the standard deviation of the T* sample. For this calculation we restricted ourselves to 430 pseudogenes that had relatively low standard deviations, less than 0.03. Extended Data Fig. 6c shows a histogram of the T* values for this set of 430 genes (boxes), The median age, shown as a blue horizontal line, is 0.0308 and the standard deviation averages 0.019 and is nearly independent of T*.

To further interpret these results, we investigated whether they could be consistent with a single burst of pseudogene formation, at an epoch equal to the median estimate T* = 0.0308. Indeed, the red curved in Extended Data Fig. 6c is the result of drawing an *in silico* sample of 430 genes with noise representing a standard deviation of 0.019. The overall shape of the curve is quite well characterized by a single normal distribution. Since the distribution is broad, our method does not have the precision to measure the detailed distribution of pseudogene ages, and our observed distribution is concordant with the onset of pseudogenization at T*~15 Mya, soon after the allotetraploidization event. A large acceleration of pseudogene creation within a few million years of the allotetraploidization event is expected (Supplementary Note 1). We note that pseudogene formation is ongoing, and that some of our predicted genes represent incipient pseudogenes, including those we refer to as thanagenes (Supplementary Note 12.5).

769 pseudogene CDS sequences built by Exonerate that were >=100bp (the size of our RNA-seq reads) were entered into our expression analysis to compute TPM (769/985=78% of pseudogenes with age estimates). 133 of these (17.2%) show degraded expression patterns. When compared to extant genes, they have less expression, show less variance in expression, and show no correlation of gene expression with their extant homoeologues. (Extended Data Fig. 6 d-f)

Supplementary Note 10. Patterns of retention and deletion.

10.1 Retention and gene function

To assess functional correlates with gene retention we used several computational methods to assign putative functions to genes.

- PfamScan (Pfam v27.0) was used to assign Pfam domains to gene loci[103].
- InterPro2GO was used to map Pfam assignments to GO terms[104].
- *X. tropicalis* KEGG assignments were extracted from the KEGG database via the REST API, and mapped onto *X. laevis* loci via orthology using triples defined in Supplementary Note 8.

Fig. 4a and Extended Data Fig. 10d, e contain scatterplots of the L retention rate vs. S retention rate for each group in the different types of classifications. A sample of the groups with significantly higher/lower retention rates is included. As is found for other whole genome duplications, DNA repair and RNA polymerase pathways are reduced to single copy more often than other loci, while homeobox,

DNA-binding, and major developmental signaling pathways are retained at significantly higher rates. There was no L/S enrichment of any genetic pathway or functional category, suggesting that interspecific incompatibility has not played a measurable role in the gene loss of *X. laevis*.

We tested for significantly high or low retention by comparing the Singleton/Homoeologue count of each group to all others of the same type (GO, Pfam, KEGG, WGCNA stage, WGCNA tissue) in a 2x2 Fisher's exact test. We tested for significance between L/S retention by comparing the L/S count of each group to all others of the same type in a 2x2 Fisher's exact test.

Mouse loci identified to be associated with the mitochondria by GFP localization[105,106] were mapped onto the *X. tropicalis* annotation via BLASTP (E value less than 1E-10, Smith-Waterman refinement) and mapped onto *X. laevis* via orthology. Germ plasm genes were manually annotated by name, and references in the literature. Here we determined significance by comparing the singleton/homoeologue count to the whole genome background (56%) and determined L/S significance by comparing to the whole-genome (75%). We find no evidence that these groups of genes are significantly highly retained, nor do they show a preference for subgenome ($p>0.01$).

10.2 Modeling gene loss and neo/sub-functionalization

Here we present two simple models for gene loss and redundancy. We partition the genes into four categories – retained only on L, retained only on S, retained on L and S but available for loss, and retained on L and S but recalcitrant to loss (perhaps due to neo- or sub-functionalization). These four categories can be applied to the entire gene set, or to only genes in a particular functional grouping (e.g., genes with a given PFAM or GO annotation):

$n(L)$ = number of genes retained in single copy on *L*-subgenome
$n(S)$ = number of genes retained in single copy on *S*-subgenome
$n(LS)$ = number of genes retained in two copies, but still available for loss
$n^*(LS)$ = number of genes permanently retained in two copies due to neo- or sub-functionalization, or because genes are required in two copies for "balance."

Model 1: progressive loss and transition to permanent retention

For a simple model for loss and progressive transition to permanent retention (*e.g.*, due to sub- or neo-functionalization), let λ_L and λ_S be the sub-genome-dependent rates of gene loss on the *L*- and *S*-subgenome, respectively, and λ^* be the rate at which genes become permanently retained in two copies. Then

$$\frac{dn(L)}{dt} = \lambda_s n(LS)$$

$$\frac{dn(S)}{dt} = \lambda_L n(LS)$$

$$\frac{dn^*(LS)}{dt} = \lambda^* n(LS)$$

$$\frac{dn(LS)}{dt} = -(\lambda_S + \lambda_L + \lambda^*)n(LS) = -\lambda n(LS)$$

These equations can be easily solved, since all are driven by the constant loss rate from the category $n(LS)$available for loss:

$$n(LS) = e^{-\lambda t}$$

which then allows the other categories to be integrated

$$n(L) = \frac{\lambda_S}{\lambda}[1 - e^{-\lambda t}]$$

$$n(S) = \frac{\lambda_L}{\lambda}[1 - e^{-\lambda t}]$$

$$n^*(LS) = \frac{\lambda^*}{\lambda}[1 - e^{-\lambda t}]$$

Then the total number of genes retained on the *L*- and *S*-subgenomes are

$$N(L) = n(L) + n(LS) + n^*(LS)$$

$$N(S) = n(S) + n(LS) + n^*(LS)$$

which can be written as

$$N(L) = \left[1 - \frac{\lambda_L}{\lambda}\right] + \frac{\lambda_L}{\lambda} e^{-\lambda t}$$

$$N(S) = \left[1 - \frac{\lambda_S}{\lambda}\right] + \frac{\lambda_S}{\lambda} e^{-\lambda t}$$

(As a sanity check, for t=0 we have $N_L \sim 1 - \lambda_L t$ as $t \to \infty$ have $N_L \sim 1 - \lambda_L/\lambda$,and similarly for N_s).

Since both $n(L)$ and $n(S)$ depend linearly on $e^{-\lambda t}$, we can eliminate time to find that, at any time (i.e., for any degree of loss), in this model the ratio of genes lost on the S- and L-subgenomes is a constant given by the ratio of loss rates:

$$\frac{1-N(S)}{1-N(L)} = \frac{\lambda_S}{\lambda_L}$$

Model 2: progressive loss with a pre-determined recalcitrant fraction

$n(L)$ = number of genes retained in single copy on *L*-subgenome
$n(S)$ = number of genes retained in single copy on *S*-subgenome
$n(LS)$ = number of genes retained in two copies, but still available for loss
$n^*(LS)$ = number of genes permanently retained in two copies due to neo- or sub-functionalization, or because genes are required in two copies for "balance."

In this model we assume that the number of genes permanently retained in two copies is fixed (unavailable for loss from the start) and unchanging.

Then

$$\frac{dn(L)}{dt} = \lambda_s[n(LS) - n^*(LS)]$$

$$\frac{dn(S)}{dt} = \lambda_L[n(LS) - n^*(LS)]$$

$$\frac{dn(LS)}{dt} = -[\lambda_L + \lambda_s][n(LS) - n^*(LS)].$$

The last equation is easily solved for the number of duplicate genes *n(LS)* retained at time *t*:

$$n(LS) = n*(LS) + [1 - n^*(LS)]e^{-\lambda t}$$

where $\lambda = \lambda_s + \lambda_L$ is the total rate of gene loss across both subgenomes.

This allows the first two equations to be integrated to yield

$$n(L) = \frac{\lambda_S}{\lambda}[1 - n^*(LS)]e^{-\lambda t}$$
$$n(S) = \frac{\lambda_L}{\lambda}[1 - n^*(LS)]e^{-\lambda t}$$

In this model the ratio of genes lost on the S- and L-subgenomes is the same as in model 1:

$$\frac{1 - N(S)}{1 - N(L)} = \frac{\lambda_S}{\lambda_L}$$

Thus both models predict that, when considering gene sets from different functional categories, scatterplots of losses on the two subgenomes should lie along a line of slope λ_S/λ_L, consistent with our observations (Fig. 4b; Extended Data Fig. 10c-e).

10.3 Retention of duplicated genes in protein complexes

For protein complex analysis, we used recently published metazoan conserved protein complex information[107]. Because this dataset was compiled with EnsEMBL human gene as a primary key, we applied the following three steps to analyze *X. laevis* genes in metazoan conserved complexes. First, we extracted 'ortholog_one2one' genes between human and *X. tropicalis* from EnsEMBL Mart (version 80). Additionally, to reduce the noise from false orthology, we discarded '*X. laevis* - *X. tropicalis* orthology' other than '1-to-1' or '1-to-2'. Next, we defined X. laevis - *X. tropicalis* orthology based on BLASTP best hits with greater than 40% of alignment ratio. Finally, we inferred human-*X. laevis* orthology by combining these two tables. As a result, we defined 10,580 human genes orthologous to *X. laevis* (3,270 singletons and 7,310 homoeologue pairs). These numbers differ slightly from other numbers in this manuscript, because here we applied additional one-to-one orthology between human and *X. tropicalis*.

We investigated whether homoeologues are preferentially retained in protein complex. Out of 3,793 proteins available in metazoan protein complexes, 1,977 proteins are orthologous to *X. laevis* homoeologues (73.5%), 715 proteins are orthologous to X. laevis singletons (18.9%), and 1,101 proteins do not have *X. laevis* orthology according to the criteria that we used. The rate of homoeologue

retention in protein complex is significantly higher than expectation (Fisher's exact test p-value = $1.3x10^{-8}$).

Out of 8,422 protein-protein interactions defined in metazoan protein complexes, 4,405 interactions were defined among proteins with *X. laevis* orthologues. Among them, 1,912 retained-homoeologue-homoeologue interactions, 1,280 homoeologue-singleton interactions, and 313 singleton-singleton interactions were identified. Similar to the retention rate of homoeologues in protein complex, the protein interaction among homoeologous proteins is significantly higher than expectation (Chi-square test p-value < $1x10^{-5}$).

More information, including the ratio of homoeologue and singleton in each protein complexes, is available at Supplemental Table 9.

10.4 Retention of ohnologs retained from ancient vertebrate duplications

Human homoeologues from the ancient vertebrate duplication were obtained from previously published studies [108]. We mapped these genes to the *X. tropicalis* v9.0 annotation via BLASTP (E value less than 1E-10, Smith-Waterman refinement). We only analyzed protein-coding genes that were in clear 1-to-1 orthology between *X. tropicalis* and human, defined as mutual-best-hits, to be sure that the genes were homoeologous in both species. This is a stringent requirement, and only 1,268/6,918 ancient vertebrate homoeologues passed these criteria. Their retention rates in *X. laevis* were computed from this gene set (Extended Data Fig. 5e). As found for teleost- and salmonid-specific duplications in rainbow trout[109], genes retained after the early vertebrate duplications are more likely to be retained after subsequent polyploidizations.

10.5 Retention relationship with gene length

We investigated retention as a function of gene length and exon number by compiling the CDS length (mRNA start to stop), genomic footprint (gDNA CDS start to stop), exon number, and the number of orthologues in *X. laevis* for each *X. tropicalis* gene in the v9.0 annotation (Extended Data Fig. 5h-j). We binned the distributions by length, and computed the number of homoeologous genes per bin (h) and the number of singletons (s). Using these variables we computed the retention rate (R), and standard deviation ($\mathcal{S}$) for each bin as follows:

$$n = h + s$$

$$R = h/n$$

$$\sigma = \sigma(\Sigma R)/\sqrt{n})$$

Retention is correlated most strongly with genome span, with longer genes (by CDS, genomic span, and exon number) more likely to be retained. This is consistent with previous reports[26] if we assume that longer genes have a larger number of independently mutable sub-functions.

Supplementary Note 11: Local duplications

11.1 Analysis of locally duplicated genes

Clusters of locally duplicated genes were identified as follows. First, the peptide sequences of all annotated genes (longest variant at each locus) were compared to each other using BLASTP, and gene pairs with sequence similarity with e-value <= 10^{-20} were defined as "substantially similar." Next, pairs of substantially similar genes located in the same vicinity on a chromosome, with at most five intervening annotated genes, were identified as "locally duplicated pairs" (Note that gene orientation was not considered, so locally duplicated pairs includes tandem duplicates as well as nearby duplicated genes coded on opposite strands). Finally, such pairs were linked together into larger clusters using a single-linkage approach in which it is required that at least one of the members of a pair is located in the same vicinity as at least one member of a different pair, with a maximum of five intervening genes.

It is evident that being part of a locally duplicated cluster is quite common. For example, 27.9% of *X. tropicalis* genes are part of such a cluster. 63% of locally duplicated clusters contain exactly two members. The largest cluster found has 123 members (immunoglobulin heavy chain on *X. laevis* chr1L:139,106,886-140,071,209).

To determine the number of tandem duplicates that originated during particular epochs in the past and have survived until the present, we used as a pairwise distance metric the fraction f4DTV of 3rd codon positions at aligned four-fold degenerate (4D) codons (coding for P,T,V,A or G) in which a transversion (purine to pyrimidine and vice versa) is observed. Corrected for multiple substitutions, the expected number of transversions at a 4D site is then S4DTV = -1/2 ln(1-2 f4DTV). Using a short custom PERL script, we built UPGMA (Unweighted Pair Group Method with Arithmetic Mean) trees of all genes in each locally duplicated cluster, starting with the most closely related gene pairs (by S4DTV), merging them into nodes represented by the sequences of both genes in the pair and re-calculating S4DTV of the node to all other nodes and genes as a weighted average of each node member's S4DTV until the tree was traversed up to a specified maximum value of S4DTV. Ages of the nodes were then binned into the nearest multiple of S4DTV = 0.01 and the number of nodes in each bin were evaluated across all clusters. Extended Data Fig. 7d-e shows the normalized numbers of nodes per bin as a function of age, as measured by the epoch bin.

Under the simplest scenario of a constant probability P0 of duplication per gene per time (generation) as well as a constant probability λ of loss per gene per generation, and a constant (approximately) total number of genes, we expect an exponentially declining function P(t) = P0 exp(-λt), or a linearly declining log(P). Indeed, the data appears to be consistent with such a model (with the exception of the first bin, which for both laevis and tropicalis is a factor of ~3 higher than expected, and possibly the 2nd bin in

laevis which appear lower than expected). The least-squares fits to the data shown corresponds to duplication rates of 1.94 and 0.548 per gene per 4DTV for *X. tropicalis* and *X. laevis*, respectively. That is, the insertion rate is more than 3.5 times higher within *X. tropicalis* than *X. laevis*. From calibration of the 4DTV measure based on more than 300,000 4D sites within multiple sequence alignments of 5,862 clusters of *X. tropicalis-X. laevis-Hymenochirus* orthologues, S4DTV = 0.11 between *X. tropicalis* and *X. laevis* orthologues. If the two species diverged 37 Mya as suggested in this work, the rates are 0.0058 and 0.0016 gene-1 Myr-1, respectively.

The mean time-to-loss 1/λ are 0.048 and 0.121 S4DTV units for *X. tropicalis* and *X. laevis*, respectively, that is, the loss rate is 2.5 times higher within *X. tropicalis*. Using the time calibration above, this translates into mean life expectancies for newly (locally) duplicated genes of 16 Myr and 40 Myr, respectively.

New genes therefore appear to be both created and lost faster within *X. tropicalis* compared to *X. laevis*. This is likely related to the faster generation time of *X. tropicalis* (up to a factor of 5 in today's lab environment), since tandem duplications occur by uneven crossovers at meiosis and should correlate inversely with generation time, as should loss of genes, provided most losses are caused by population-genetic selection processes rather than an immediate pseudogenization of one of the copies by the first random disabling mutation. The one exception to this would be the excessive number of observed tandem duplications in the first bin, most of which have S4DTV = 0. We believe this excess represents tandemly created pseudogenes that were disabled at birth, such as partially duplicated genes, or by a truly deleterious point mutation, but whose sequences have not had time to degrade substantially, and hence are still annotated as functional genes. If this view is correct, as many as 1,200 annotated *X. tropicalis* and 700 *X. laevis* genes are, in fact, newly minted pseudogenes.

Genome	# Genes	In clusters	# clusters	% in clusters
X. tropicalis	25,668	7,162	1,763	27.9
X. laevis	44,164	8,195	2,148	18.6
X. laevis chrL	23,667	4,908	1,191	20.7
X. laevis chrS	16,939	2,211	726	13.1

11.2 Nomenclature of duplicated genes

When genes are duplicated independently as a single cluster in each species or subgenome, orthologous genes between such species or homoeologous genes between subgenomes are in one-to-multi or multi-to-multi, not one-to-one, relationships. According to a new version of *Xenopus* nomenclature system (http://www.xenbase.org/common/), expanded genes in each single cluster are named by adding .1, .2, etc. as suffixes to root gene names, in which one-to-one orthologous relationship does not hold. To manifest multi-to-multi orthologous relationships, we considered a slightly modified nomenclature system applicable only for genome/subgenome-specific gene expansion by inserting a special character "e" (indicating gene expansion) before species/subgenome-specific paralog number. In this paper,

following root gene names, we added. e1, .e2, etc. as suffixes for *X. tropicalis*, and .e1.L, .e2.L etc. and .e1.S, .e2.S etc. for *X. laevis*. For example, the *bix* cluster consists of *bix.e1*, *bix.e2*, and *bix.e3*.

11.3 Nomenclature of pseudogenes

According to the gene nomenclature guideline by the *Xenopus* Gene Nomenclature Committee (XGNC; see Xenbase), "p" plus the serial number is added to a gene name as a suffix. When either homoeologues (L or S) of certain genes was pseudogenized, the suffix p was added to gene names, like *hoxb2p.L*. In the case of species/subgenome-specifically expanded genes, two ways were adopted. One way is the original one: for example, pseudogenes of the *nodal3* cluster are named *nodal3p1.L* and *nodal3p2.L*. The other way is that p was added after .e plus a paralog number: for example, pseudogenes of the *bix* clusters in *X. tropicalis* were named *bix.e2p* and *bix.e5p*.

Supplementary Note 12: Gene expression

12.1 Quantification of gene expression levels with RNA-seq

We analyzed gene expression of the RNA-seq data described above in Supplementary Note 4 for a developmental time series and selected adult tissues. After filtering (1) reads with no call ('N') and (2) reads with low complexity (not having all of 'A', 'C', 'G, and 'T') from raw J-strain RNA-seq reads, we mapped them to primary transcript sequences using bwa mem (version 0.7.10) with paired-end option {Li 2014}. We quantified the expression of each transcript using Transcripts Per Million" (TPM) values estimated by RSEM (version 1.2.19) [110].

To prevent the noise derived from reads of homoeologous transcripts, we removed hits either (1) with additional targets including homoeologues, or (2) with partial alignment with insertions/deletions (indels), before running RSEM. As a result, we used highly specific reads mapped only on one copy of homoeologue transcripts in this analysis. This approach may underestimate the expression of homoeologous transcripts by ignoring reads from identical regions, although the expression of most homoeologues can be measured by taking advantage of paired-end reads. All redundant sequences were removed in the database before mapping, to measure at least group-level expression of redundant sequences.

All scripts used in this analysis are available at https://github.com/taejoonlab/HTseq-toolbox/.

12.2 Gene expression vs. developmental time and tissues

Prior to analysis, all TPM values <0.5 were reduced to 0. Any gene whose expression value is <0.5 for all samples was removed from analysis. For each homoeologue pair at each tissue and time point, the L/S expression ratio was calculated and log transformed according to ($\log_{10}$(TPM.L+0.1/TPM.S+0.1)). The boxplot of expression ratios between sub-genomes is included in Fig. 4b. On average the L sub-genome is expressed higher than the S in all tissues and time points, however the magnitude of that expression difference varies. Prior to the maternal-zygotic transition (MZT), and in the adult ovary, genes of the L sub-genome are expressed 11% higher than genes of S, on average. Post-MZT, and in somatic adult tissues, L is expressed 23% higher than S on average. These results imply that maternal and zygotic expression can be differentially regulated.

Maternal and zygotic gene expression can be controlled by different promoters and/or enhancers providing a path to subfunctionalization by complementary loss or degradation of maternal and zygotic regulatory elements such that one homoeologue becomes specific to maternal expression, and the other to zygotic expression. This path to subfunctionalization follows the general scheme outlined by Force *et al*[26]. We scanned homoeologue pairs for the pattern of one gene being on prior to MZT (TPM > 0.5), and the other completely shut off (TPM < 0.5, while both are on after MZT (example in Fig. 4c, d). There are 200 homoeologues where L is expressed early, and S is not, and 191 where S is expressed early, and L is not. Conversely, there are only 19 homoeologue pairs which partition their expression between the embryo and somatic adult tissues (*i.e.*, they have no sub-genome bias). We are currently investigating whether the increased plasticity of maternal expression is due to more rapid turnover of maternal promoters.

12.3 Co-expression network inference and analysis of modules

We analyzed expression variation among homoeologous genes by developing co-expression networks with Weighted gene co-expression network analysis (WCGNA)[111]. Prior to analysis, all TPM values below 0.5 were reduced to 0. Any gene whose expression level was less than 0.5 in all stages or tissues was removed from analysis. For developmental expression data we restricted ourselves to 3,797/8,806 homoeologue pairs that showed differential expression in at least one experiment (minimum 10-fold expression difference). For adult data, all 8,806 homoeologue pairs were used to extract module eigengenes. The observed expression values ($\log_{10}$(TPM+0.1)) for each gene in a homoeologue pair were summed in a homoeologue expression matrix. We then inferred a weighted undirected co-expression network using the WGCNA method[111] with a soft thresholding power of 12. Next, groups of closely connected genes, or modules, were identified by clustering genes based on the topological overlap matrix and cutting the resulting dendrogram with the cutreeDynamic method in R (parameters: deepSplit=2, pamRespectsDendro=FALSE, minModuleSize=30). Non-module genes were grouped into an artificial “grey” module. Initial modules whose expression profiles were very similar (eigengene correlation >= 0.85) were merged. For the heatmap visualization in Extended Data Fig. 10e the genes were organized by group, and expression patterns were visualized by the heatmap function in R.

Single copy genes, and homoeologue pairs that were originally not used in the WGCNA analysis, were assigned to WGCNA modules by computing a correlation matrix between each gene and the eigengene

expression patterns. We then utilized the corPvalueStudent function (with a p-value cutoff of 0.01) of the WGCNA package to test for significant correlations between genes and eigengenes. If the smallest p-value was greater than 0.01, the gene was assigned to the artificial "grey" module.

For each co-expression module, we determined whether the homoeologue retention rate was higher than expected by a Fisher's exact test ($p < 0.01$). Significant differences between L/S retention rates were also determined by a Fisher's exact test. Significant differences in evolutionary rates of different modules were computed using the Wilcoxon-Mann-Whitney Test in R.

12.4 Evolution of homoeologous expression following allopolyploidy

To further explore the divergence of expression between homoeologues, we repeated the analysis of expression divergence performed in the trout genome analysis[109]. The expression data for adult tissues and developmental stages was merged, and TPM <=0.5 was reduced to 0. 7,616/8,806 (86.4%) of homoeologue pairs showed expression for both genes in at least one data set (*i.e.*, each homoeologue was expressed in at least 1 tissue or time point). The TPM values were log-transformed ($\log_{10}$(TPM+0.1)), and the correlation coefficient and p-value were computed by the cor.test function in R, while the mean expression difference p-value were computed by the t.test function in R. For both correlation and t-tests, a p-value <= 1E-5 was chosen as significant (stringent to account for multiple sampling error). The 'HC' group is for homoeologue pairs whose correlation was significant, and 'NC' is for homoeologue pairs whose correlation was not significant. All significant correlations were positive. 'DE' is for homoeologue pairs where the t-test supports a significant difference in the mean expression between two homoeologues, and 'SE' is for homoeologue pairs that do not show a significant difference in their mean expression. Combining these classifications, we determine that there are 3,966 (52.1%) homoeologue pairs in the HCSE group, 169 (2.2%) homoeologue pairs in the HCDE group, 2,745 (36%) pairs in the NCSE group, and 736 (9.7%)pairs in the NCDE group.

To investigate differences in mutation rate and length between homoeologue categories, we computed the 4DTv between homoeologues for each pair, as well as the absolute value of the CDS length difference and K_a/K_s (Extended Data Fig. 8d-f). We assessed significance between the categories for these variables by using a Wilcoxon-ranked sum test of each classification against all others. We find that the HCSE group has a lower 4DTv, length difference, and K_a/K_s than the others, while the NCDE group has a higher 4DTv, length difference, and K_a/K_s than other categories. The NCSE group shows a high K_a/K_s as well, suggesting that uncorrelated expression may be linked to a higher rate of protein sequence change between homoeologues.

We also asked if any of the homoeologue categories were enriched in function, by looking for enrichment of GO, PFAM, and KEGG categories in each group (Supplemental Table 10). We test for enrichment by using a Fisher's exact test ($p<=0.01$). Similar to previous study with rainbow trout[109], we find that DNA-binding and regulation of transcription are strongly associated with the HCSE group. Additionally, we find ribosomal and mitochondrial functions are enriched in the HCDE group, and glutathione metabolism is enriched in the NCDE group. We also find significant enrichment in the larger groups, with Homeobox domains being enriched in the high correlation (HC) group, C2 domains

enriched in the no correlation (NC) group, leucine rich repeats enriched in the same expression (SE) group, and short chain dehydrogenase genes in the differential expression (DE) group (Supplemental Table 10). These results reveal a strong link between the evolution of gene expression following duplication, and gene function. Together with the result that gene expression is strongly correlated with retention (Extended Data Fig. 8b), these data support the hypothesis that gene dosage is an important factor in determining both gene retention, and the potential for differential expression of a locus following polyploidy.

12.5 Thanagenes

Thanagenes in *X. laevis* are defined as well-formed genes predicted by our gene annotation pipeline whose expression never rises above a TPM of 0.5 in all 28 experimental data sets, but whose *X. tropicalis*orthologue and *X. laevis* homoeologue are expressed at a TPM above 1 in at least 1 tissue or timepoint. Sequence alignments of triplets containing a thanagene were concatenated based on which subgenome the thanagene was present. The remaining three-way alignments were also concatenated for comparison. Gaps were removed by Gblocks using default parameters. Maximum-likelihood trees were built by PhyML, using GTR model of evolution and 1,000 bootstrapped trees. Error estimates in Extended Data Fig. 5f are the standard deviation of the 1,000 bootstrapped trees computed by PhyML. We find that thanagenes have accumulated additional non-synonymous substitutions, implying that they have relaxed constraint. For more detailed analysis of a decaying locus, see the Six6 analysis in Supplementary Note 13.1.

Supplementary Note 13: Analysis of specific gene families and pathways

13.1. Six6

Phylogenetic trees were constructed using the neighbor-joining method with human SIX6 protein sequence (GenBank accession number, NP031400), *X. tropicalis* Six6 (deduced from the *six6* gene model of JGI, v9.0), and *X. laevis* Six6.L and Six6.S (deduced from *six6.L* and *six6.S* gene models of JGI, v.1.8). *six6* CNEs were identified using the MultiPipMaker alignment tool[113]. Their locations on genome assemblies are as follows: Homo sapiens (hg38), chr14: 60,507,610–60,507,775; *X. tropicalis* (9.0), chr8: 82,271,421–82,271,586; *X. laevis* (9.1), chr8L: 79,230,888–79,231,053, chr8S: 7,733,127–7,733,291. *X. laevis* CNEs with their short flanking sequences (chr8L: 79,230,877–79,231,172 and chr8S: 7,733,011–7,733,302) were amplified from the J-strain genomic DNA, and linked to a β-actin basal promoter-GFP cassette[114]. The resulting constructs were introduced into *X. laevis* embryos using the nuclear transplantation method as described[29]. Semi-quantitative analysis of the GFP expression was performed by in situ hybridization followed by densitometric measurement of staining signals in the eye region. Statistical analysis of the data was performed using Student's t-test.

13.2. Cell cycle

A comparative analysis of genes encoding cell cycle regulators in *X. laevis* indicates that homoeologueues of *cdk7* and *ccnh* were deleted from XLA1S, whereas homoeologous gene pairs of other *cdks* and *ccns* are mostly retained (Extended Data Fig. 9b). Cdk7 and Cyclin H constitute Cdk-activating kinase (CAK), the key factor that is essential for the progression of cell cycle via activating phosphorylation within the activation-loop of Cdk1, Cdk2, Cdk4 and Cdk6[115].

Homoeologous gene pairs of cdks except cdk7 are retained (Extended Data Fig. 9b), while *cdk* genes including *cdk7* in plants are fundamentally conserved with low copy number after repeated whole genome duplication[116]. Analysis of cdk7 genes in zebrafish and medaka, which experienced a 3rd whole genome duplication during evolution, supports that cdk7 should be low in genome. Also, the duplicate of ccnh gene is not retained and it became a singleton even though other *ccns* are amplified[117]. These examples strongly suggest that the copy number of *cdk7* and *ccnh* genes must be low during evolution. Having become singleton genes may make their gene regulation simplified compared to maintaining homoeologous gene pairs. Supposing that the potential maximum expression level of a singleton is the half of that of paired genes put together, becoming singletons may prevent harmful high expression of key regulators caused by stochastic and accidental perturbations.

Though *cdk* genes are conserved to be low, homoeologous gene pairs of cdks are retained except cdk7 in *X. laevis*. This property is supposed to cause its feature. One of the features of organisms with tetraploid genomes is larger body size than those with diploid genomes (e.g., *Xenopus*, trout, Arabidopsis). This is true for *X. laevis*: its body length is 2.55 times longer than that of *X. tropicalis*[118]. The larger body size is caused in part by larger cell sizes in *X. laevis* than in *X. tropicalis*[119]. It was previously shown that slower progression of the cell cycle induced by treatment of mouse fibroblasts or avian erythroblasts with low concentrations of inhibitors for DNA synthesis results in abnormally larger cell sizes[35]. Since CAK (Cdk7+CcnH) is required for progression of cell cycle[115] and their expression is lower in *X. laevis* than that in *X. tropicalis*[120]. It is possible that their lower expression links to slower progression of cell cycle and larger cell sizes in *X. laevis* than those in *X. tropicalis*.

13.3. TGF-beta

The TGF-beta family of growth factors plays important roles in diverse biological processes. Developmental signaling pathways such as Wnt and Hh tend to have high rates of homoeologue retention, but the TGF-beta pathway is the champion (Fig. 4a). Exceptions include the *nodal3* and *nodal 5* loci which contain at least four or five reiterated copies in *X. laevis*, but in both these case the entire locus is deleted from S (Extended Data Fig. 7b-d). The *vg1* gene is tandemly expanded on L, but is lost on S (Extended Data Fig. 7e)[121]. Moreover, the BMP antagonist chordin gene is specifically duplicated on the L chromosome. A number of extracellular regulators (e.g. antagonists) are differentially regulated during embryogenesis and some of them are lost from the S chromosome (*grem1* and *ltbp4*). Genes

modulating the magnitude of signaling in the cell such as co-receptors, (*hfe2, eng* and *tgfbr3*), the TGF-beta type I receptor (*tgfbr1*), the inhibitory Smad signal transducer (*smad6.2*) and a negative docking protein for activin type I receptor (*dok1*) are retained in both homoeologues, but differentially transcribed during early embryogenesis (Extended Data Fig. 10a).

13.4 Immune Genes

Based on previous studies and current knowledge of basic immune mechanisms, we predicted that immune genes directly contributing to antigen receptor generation would be reduced to single copy to maintain an 'optimal' precursor frequency of B and T cells, providing effective immunity while preserving self-tolerance. Such immune genes include the antigen receptors (AgR) (i.e., T cell receptors (TCR) and immunoglobulins (Ig)) and antigen presentation molecules, including major histocompatibility complex (MHC) class I and II and their processing genes. Other immune genes not related to repertoire generation and maintenance were predicted not to be reduced to single copy. Supplemental Table 8 shows that the results met the expectations. However, in some cases when both gene copies were present in the genome, gene silencing is controlled at the expression level with one gene expressed at higher levels (expression in spleen and liver tissues shown in Supplemental Table 8). In the descriptions below, we categorized immune genes based on their functions.

Antigen Receptors. It was previously shown that the Ig heavy chain locus was reduced to single copy in *X. laevis* as all Vh genes were found only on one chromosome[122]. Thus, we predicted that all antigen receptors would be reduced to single copy in tetraploid *X. laevis*. All chains encoding V, D, and J gene segments (e.g. Ig-heavy, tcr-beta, and tcr-delta) are indeed reduced to single copy as well as some Ig-light chains and tcr-gamma. Yet we found igl-sigma and tcr-alpha genes encoded by both homoeologues. The presence of homoeologous genes for these VJ chains suggests that the number of VDJ chains (e.g. Vh) is under stronger selection. However, expression of one set of chromosomes is much higher than that of the other chromosome, suggesting that there is some level of functional silencing. Note that the second set of tcr-alpha genes were not assigned to a chromosomal location, but the two tcr-alpha genes were in similar genomic linkage groups (e.g. both linked to the homoeologous set of genes for *dad1* and *abhd4*), suggesting that these two tcr-alpha genes were homoeologues. Homoeologous tcr-alpha may be beneficial during the receptor editing process in the thymus; as shown in mice and humans, tcr-alpha is not allelically excluded, consistent with the lack of gene loss in *X. laevis*. In the *X. tropicalis* genome, there are two TCR-alpha genes in *cis*. The *X. laevis* scaffolds containing tcr-alpha genes are not large enough to obtain the entire genomic regions and thus it is not known whether there are two tcr-alpha genes in *cis* in *X. laevis*.

Genes involved in AgR generation and transcription factors. Some genes important for somatic gene rearrangement of Ig/TCR (i.e. *rag2* and *tdt*) are reduced to single copy, suggesting that the regulation of the gene rearrangement process might be tightly controlled in a dose-dependent manner. One exception is *rag1*: as previously reported[123], both homoeologueues are retained in the genome with no recombination between them. The *rag1* and *rag2* genes map next to each other and the proteins form a complex to initiate gene rearrangement. Since expression levels of the two *rag1* homoeologueues are

similar (2-fold; although TPM is below the cut-off threshold), both *rag1* products may form complexes with rag2. However, some *rag1* gene copies were silenced in higher ploidy *Xenopus* species[124], suggesting that *rag2*-non-linked *rag1* gene may be "on the way" to becoming silenced. All genes playing roles in AgR signaling are tetraploid, except *cd3d/g* (CD3d and g are not differentiated into separate genes in frogs, with an intermediate d/g ancestral chain), suggesting that dosage of co-receptors and signaling molecules does not affect immune function. Reduction to single copy of *aire* suggests that AIRE-mediated negative selection during T cell development in thymus must be tightly regulated. AIRE interacts with a large number of proteins involved in gene transcription, consistent with tight regulation of its expression. Other transcription factors for plasma (B) cell differentiation or T helper (Th) cell subset differentiation seem to be lost or differentially silenced. However, expression level needs to be experimentally evaluated.

MHC. As was previously published[122,125,126], MHC genes in the polyploid *Xenopus* species are reduced in number. In fact for some genes such as class II beta, the number of genes seems to matter more for the selection during evolution of *Xenopus*[125]. Specifically, MHC genes directly involved in antigen presentation (i.e., MHC class I, II, *dm*) and class I antigen processing pathway (*lmp2, lmp7, tap1, tap2*) are all reduced to single copy, while genes that do not map to mammalian MHC (i.e., the third immunoproteasome *psmb10* and *B2m*, the light chain necessarily for MHC class I assembly which associates with MHC class I[127] are not lost but are differentially expressed. Thus, close linkage to the MHC seems to play important roles for antigen presentation, perhaps for regulation at a multi-locus level rather than the individual gene level. Interestingly, differential gene loss on the chromosome pairs may also be the result of gene regulation. Two immunoproteasome subunits are differentially lost on one chromosome or the other; *lmp7* is functional on the L chromosome along with the class I, *tap*, and class II, while *lmp2* is functional on the S chromosome (Fig. 3d). Like other transcription factors involved in AgR or Th cell differentiation, transcription factors for MHC class I (*nlrc5*) and Class II (*cIIta*) expressions are not diplodized. Many other genes that map to the MHC, but not involved in antigen presentation, are not reduced to single copy or show differential silencing (e.g., *lta, ltb, tnfa* in the Cytokine category). In addition, there is a cluster of non-classical class I (*XNC*) genes found in the telomere region of the MHC chromosome 8 that previously was reported to be lost[122]. In our more detailed analysis, we found that the *XNC* region is not completely reduced to singel copy, with at least 16 genes on L chromosome and 4 genes on S chromosome so far detected (data not shown). The discrepancy is due to the sensitivity in original Southern blotting[128] and FISH (personal communication with Du Pasquier) analyses by using evolutionarily conserved alpha-3 domain as probe. Note that we tried to compare the MHC architecture of *X. laevis* to that of *X. tropicalis*, but two newer versions of *X. tropicalis* assemblies (v.7.1 and v.8) were both incomplete and matched neither to *X. laevis* nor to each other. Thus, we did not include them in our analysis.

Complement. The Complement system plays important roles in pathogen clearance as a soluble mediator and there are three pathways: classical, alternative, and lectin. Genes involved in the classical pathway (e.g. *c1q, r, s*) are all reduced to single copy except *c3*, while genes in other innate pathways remain tetraploid. *C3* is a major substrate binding molecule, the focal point of all three pathways, and found at the highest levels in serum. The *c3* gene was expanded both in the diploid *X. tropicalis* and *X. laevis*, thus a general feature in *Xenopus*. Interestingly, the various *c3* genes are differentially expressed in various tissues with only one *c3* gene being dominantly expressed in liver, a major complement

component-producing organ; this suggests that some *c3* genes have been subfunctionalized after *cis*-duplication. Such subfunctionalization can be also seen between *cfb* homoeologues: in addition to expression in the liver, *cfb* in L chromosome is also expressed in the lung (~120 tpm), whereas almost no expression was detected for *cfb* in S chromosome in the lung. All members of the membrane attack complex (MAC) (*c5*-*c9*) are lost, perhaps to discourage a 'dominant negative effect' (i.e. so that gene products from different alleles might not inhibit formation of the multi-chain complex). Complement regulators are neither reduced to single copy nor differentially expressed, suggesting that the *Xenopus* complement system, with the increase in c3 isotypes, is more complex and must be tightly regulated with a diverse array of regulatory proteins.

Innate Pattern Recognition Receptors. In general, many genes are lost, but some have expanded in a species-specific manner[129]. *tlr1, 3, 6, 7, 8*, and *9* are lost, while *tlr2* and *4* are not. In some cases (e.g. *tlr12, 13*), multiple genes were found in unassigned scaffolds, thus it is difficult to determine whether they are homoeologous genes or tandem duplicates. There are differences in gene numbers between *X. tropicalis* and *X. laevis,* suggesting that the loci evolved over a short evolutionarily time. Indeed, expansion of pattern recognition receptors is seen in many different species, some with over 1,500 genes in particular families[130]. Thus, similar to *Xenopus c3*, expanded *tlr* genes may have different functions.

Cytokines and Receptors. We identified major cytokines and their receptors and most of them, including innate and adaptive cytokines, are not reduced to single copy. The *il2* and *il2rg* are among the few singleton cytokine genes. *IL2* is a critical lymphocyte growth factor and its receptor *IL2Rg* is shared among at least 5 other cytokine receptors. Expression of multiple *il2*, and especially *il2rg*, might lead to non-specific lymphocyte activation, and contribute to susceptibility to certain infections or in controlling autoimmunity. We could not detect some cytokines, especially the short ones. However, this might be due to the limitation of the gene prediction algorithm.

Costimulation and other genes. Costimulation is important for lymphocyte regulation, depending on a balance between positive and negative regulatory molecules. Therefore, like in the case of *il2*, we expected that most costimulatory genes would be lost to avoid non-specific activation. We specifically looked for the B7 family members and found that, in contrast to expectations, none of these genes was reduced to single copy. Although the expression level from each chromosome is not equal, the number of positive and negative costimulatory molecules (e.g., CD28 and CTLA4) may need to be balanced in order to provide an optimal immune response. A cluster of *perforin* genes was found on *X. laevis* scaffold-22, and without a homoeologous cluster, however one gene tends to be dominantly expressed. Perforin is also expanded in the *X. tropicalis* genome, thus the expansion of this gene is a common feature, for unknown reasons.

13.5 Hippo signaling

Many components of the Hippo pathway are highly conserved from *Drosophila* to mammals. In the whole genome of *Xenopus laevis*, almost all homoeologous genes for the Hippo pathway were identified on chromosomes L and S. RNA sequencing data showed that both of the homoeologous genes are

expressed with the similar pattern during early development and in the adult organs. However, in case of *taz*, a key gene of the Hippo pathway, one of the homoeologous gene pair has been completely lost from chromosome XLA8S (Extended Data Fig. 9c). In contrast to singleton of *taz*, both of duplicated *yap* genes are conserved and expressed. Yorkie is a key regulator in Drosophila Hippo pathway, while the two paralog *yap* and *taz* exist in vertebrates. Although the two paralogs have similar functions, an essential role is played by taz instead of yap in mesenchymal stem cell differentiation[131] and in Wnt signaling[132]. Some evolutional selection may have put pressure on key gene of Hippo pathway, *taz*, to become a singleton. Since scalloped (*tead* ortholog) is a default repressor, its co-factor yorkie is required for normal cell growth[133]. In Drosophila, the Hippo pathway negatively regulates expression of cyclin E to restrict cell proliferation[134]. Interestingly, gene loss in the *X. laevis* genome was also recognized in one of homoeologous pair of cyclin H genes (*ccnh*) and its partner gene cdk7 (Extended Data Fig. 9b). CcnH/Cdk7 plays an initiating role in the cell cycle by activating cyclin E/cdk2, which promotes cell proliferation. Simultaneous gene loss of *taz* and *ccnh/cdk7* might imply that dosage-sensitive regulation occurs in cell cycle regulation after allotetraploidization.

13.6 Hedgehog (Hh) signaling

Like other signaling pathways, for example, mediated by Wnt or TGF-beta families, the Hedgehog (Hh) signaling pathway is also used repeatedly during embryonic development. During vertebrate evolution, for example in mammalian, three (Indian, Sonic and Desert) Hh genes (*ihh*, *shh*, *dhh*) are evolved presumably due to genome duplications, whereas in zebrafish there are five Hh genes[135]. In *Xenopus laevis* we have found both homoeologues of *ihh*, *shh*, and *dhh*. Likewise, most homoeologues for major components of the Hh pathway in *X. laevis* reside in the same syntenic region on both chromosomes L and S with some exceptions. For example, *hhat* is only found in the L chromosome. zebrafish also has only one *hhat*, suggesting that this gene is one of low copy number family genes. RNA-seq data showed that homoeologues are expressed at the comparable levels or that L genes tend to be expressed at higher levels. Some S genes (e.g., *gli1*), however, are expressed at higher levels than L genes.

13.7 Hox clusters

A total of eight Hox clusters, consisting of homoeologous pairs of HoxA, B, C and D clusters were found in the genome, and were named HoxA.L, HoxA.S, HoxB.L, HoxB.S, HoxC.L, HoxC.S, HoxD.L, and HoxD.S. Whereas *X. tropicalis* has four clusters and 38 Hox genes (*X. tropicalis* genome assembly v9), *X. laevis* has twice the number of genes, which is 76 in total, including one pseudogene. This pseudogene is *hoxb2p.L*, and compared to its homoeologueue, *hoxb2.S*, is predicted to have frameshifts which would lead to the production of a truncated protein, if any. The location on the chromosomes are in correspondence between these two species. Since *X. tropicalis* HoxB and HoxD clusters are positioned on XTR9 and XTR10, respectively, two Hox clusters in *X. laevis*, namely HoxB.L and HoxD.L (or HoxB.S and HoxD.S), are located on the same chromosome, XLA9_10L (or XLA9_10S), on the p and q arms, respectively.

13.8 *mix*, *mixer*, *bix*

Gene models of *bix.e1*, *.e2p*, *.e4*, and *.e5p* of *X. tropicalis* were revised manually to adjust exon boundaries. The TGG to TGA point mutation at the 390th nt of the *bix.e5p* CDS resulted in a premature stop codon. Exon 1 of *bix.e4* is not found due to a long N gap in the upstream of exon 2, assuming that *bix.e4* is an active gene. A five nt deletion in exon 1 of *bix.e2p* caused a frameshift mutation. Exon 1 of *bix.e1* is not present in the *X. tropicalis* genome ver. 9, but was found in a BAC clone sequence ISB1-324H4, which was used for reconstructing the full CDS. For phylogenetic analysis using MEGA6[85], full or partial CDSs of 21 sequences from the mix family were aligned using MUSCLE and their phylogenetic relations were inferred using the Neighbor-Joining method with the bootstrap test (1,000 replicates).

Human, chicken, and zebrafish have a single *mix-like 1* (*mixl1*) gene, whereas *X. tropicalis* and *X. laevis* have multiple *mix*-related-genes in a single cluster. The phylogenetic tree suggests that a single *mix* ancestral gene was triplicated in the *Xenopus* ancestor to generate *mix1*, *mixer*, and *bix*, which are orthologous to *mixl1* (Extended Data Fig. 7a). Curiously, *bix* gene clusters on *X. tropicalis* and *X. laevis* subgenomes each fall into the same clade, suggesting that the *bix* gene clusters were generated by species/subgenome-specific gene expansion.

13.9 *nodal5*, *nodal3* and *vg1*

Amplification of *nodal5*, a mesodermal inducer of *X. laevis*, has been reported previously[136]. FISH analysis showed S chromosome specific deletion of *nodal5* gene cluster. Hybridization signals of *nodal5* (arrows) were detected in the long arm of XLA3L, but the entire locus is deleted from XLA3S. The cluster consisted of four active *nodal5* genes and one *nodal6* gene in *X. tropicalis* genome, at least five active *nodal5* genes (*nodal5.e1.L* ~ *nodal5.e5.L*), pseudogenes (*nodal5p1.L* ~ *nodal5p4.L*) and one *nodal6.L* gene in *X. laevis* genome (XLA3L), and are deleted from the corresponding region of XLA3S, except one pseudogene (*nodal5p1.S*) and *nodal6.S*.

X. laevis nodal3 and *vg1* genes are located on chromosomes 3L and 1L, respectively (Extended Data Fig. 7d, e). Among four complete copies of *nodal3*, *nodal3.e1.L* and *.e2.L* correspond to the Genbank sequences *nodal3.1* (NM_001085790.1) and *nodal3.2* (NM_001085596.1), respectively. *nodal3p1.L* is a pseudogene with truncations at both 5'- and 3'- ends of the coding sequence. The truncations of *nodal3p1.L* are confirmed by fosmid full-sequencing. There are two other highly degenerate *nodal3* pseudogenes on the L chromosome. *vg1* and *derrière* are orthologous to mammalian *gdf1*. Three copies of *vg1* genes (*vg1.e1.L*, *.e2.L* and *.e3.L*) were identified by BAC full-sequencing of three independent clones. Based on a previous report[121], these copies can be grouped into two types of *vg1* genes encoding serine (*vg1.e1.L* and *.e3.L*) or proline (*vg1.e2.L*) residues at the position 20 of their protein products. On the S chromosomes, both *nodal3* and *vg1* gene clusters are deleted from the corresponding regions of 3S and 1S, respectively, although there is a pseudogene for *vg1* gene (*vg1p.S*). The gene cluster deletions from the S chromosome were confirmed by cDNA FISH analysis.

13.10 Wnt signaling

Wnt signaling plays important roles in *Xenopus* early embryogenesis including D-V axis formation, A-P neural patterning, and cell proliferation (Extended Data Fig. 9b). In the canonical Wnt pathway, almost all components we analyzed possessed both L and S genes, but *wnt2b*, *wnt11b*, and *lrp5* lost their S genes, and *tcf7* lost its L gene. At the *wnt2b* locus, a long deletion occurred in the S chromosome, resulting in the deletion of *wnt2b.S* and at least two genes.

13.11 Germ plasm

In many animals, the germline is specified by the maternal inheritance of germ plasm[137,138]. Germ plasm contains ribonucleoprotein granules and mitochondria, and since germ plasm proteins evolve relatively quickly, we might expect some incompatibility of L and S alloalleles in directing germ cell formation. Indeed, we note that genes encoding critical maternal germ plasm-localized RNAs appear as single homoeologues on L subgenome chromosomes. These genes include *dazl*, *ddx4* and *ddx25* (*vasa* homologs), and *dnd1*, which have highly conserved roles in multiple aspects of germ cell development in *Xenopus*[139–143], as well as other novel germ plasm-localizing RNAs with unknown functions, *ddx59* and germes (LOC779566)[144,145]. Germ plasm also contains components of the Piwi/piRNA pathway, required for PGC specification and transposon silencing in the germline[146,147]. Many Piwi pathway-related genes are also exclusively L homoeologues in *X. laevis*, including *piwil1*, *piwil2*, *asz1* and *trdr6*. Exceptions to the trend include germ plasm-localized RNAs with functions outside the germline (e.g., *sybu*, *grip2*) which are encoded by both L and S homoeologues, as are the non-germ plasm vegetally-localized RNAs, such as *vegt*, *gdf1* (*vg1*) and *bicc1*[148] suggesting that preferential loss of S homoeologues is limited to core germ plasm components and not to localized RNAs in general.

13.12 Mitochondria

During allopolyploidy, the nuclear genome duplicates while the mitochondrial genome does not. While the two progenitor species were separated they may have gained cytonuclear incompatibilities that would cause the nuclear-encoded mitochondrial components from the same subgenome as the maternal contributer to the polyploidy be preferred. To assess these we aligned proteins annotated by the database as localizing to the mitochondria in mouse and human to *X. tropicalis* to identify their frog orthologues. We found both homoeologue and L retention rates were similar to background (Table 2; Supplemental Table 3).

Supplementary Note 14: Epigenetics

14.1 ChIP-seq experimental protocol

Embryos (n= 35-90) were fixed in 1% formaldehyde for 30 mins at Nieuwkoop-Faber stage 10.5. Embryos were washed once in 125 mM glycine / 25% Marc's modified Ringer's medium (MMR) and twice in 25% MMR, homogenized on ice in sonication buffer (20 mM Tris-HCl, pH 8/10 mM KCl/1mM EDTA/10% glycerol/5 mM DTT/0.125% Nonidet P-40, and protease inhibitor cocktail (Roche)). Homogenized embryos were sonicated for 20 minutes using a Bioruptor sonicator (Diagenode). Sonicated extract was centrifuged at top speed in a cold table-top centrifuge and supernatants (ChIP extracts) were snap frozen in liquid nitrogen and stored at –20°C until used. Prior to assembling the ChIP reaction, the ChIP extract was diluted with IP buffer (50 mM Tris•HCl, pH 8/100 mM NaCl/2mM EDTA/1 mM DTT/1% Nonidet P-40, and protease inhibitor cocktail) and then incubated with 1–5 µg of antibody and 12.5 µl Prot A/G beads (Santa Cruz) for an overnight binding reaction on the rotating wheel in the cold room. The following antibodies were used: H3K4me3 (Abcam ab8580), H3K4me1 (Abcam ab8895), p300 (C-20, Santa Cruz sc-585), H3K36me3 (Abcam ab9050) and RNA polymerase II (Diagenode C15200004).

The beads were sequentially washed, first with ChIP1 buffer (IP buffer plus 0.1% sodium deoxycholate), then ChIP2 buffer (ChIP1 buffer with 500 mM NaCl final concentration), then ChIP3 buffer (ChIP1 buffer with 250 mM LiCl), then again with ChIP1 buffer, and lastly with TE buffer (10 mM Tris, pH 8/1 mM EDTA). The material was eluted in 1% SDS in 0.1 M sodium bicarbonate. Cross-linking was reversed by adding 16 µl of 5 M NaCl and incubating at 65°C for 4–5 hours. DNA was extracted using the Qiagen QIAquick PCR purification kit. About 10 ng input DNA was used for sample preparation for high-throughput sequencing on an Illumina HiSeq 2000 or NextSeq (according to manufacturer's protocol).

14.2 ChIP-seq data analysis

Reads were mapped to the *Xenopus laevis* genome (Xenla9.1) using bwa mem (version 0.7.10-r789) with default settings[90]. Duplicate reads were marked using bamUtil v1.0.2. Where applicable (H3K4me3, p300) peaks were called using MACS (version 2.1.0.20140616)[149] relative to the Input track using the options --broad -g 2.3e9 -q 0.001. --buffer-size 1000. Peaks were combined for replicates using bedtools intersect (version v.2.20.1)[150]. Figures of genomic profiles were generated using fluff v1.62 (Zenodo. 10.5281/zenodo.34209).

14.3 MethylC-seq for whole-genome bisulfite sequencing

Embryos (n=24; NK stage 10.5) were homogenized in 3 volumes STOP-buffer (15 mM EDTA, 10 mM Tris-HCl pH7.5, 1% SDS, 0.5 mg/mL proteinase K). The homogenate was incubated for 4 hours at 37 degrees. Two phenol:chloroform:isoamylalcohol (PCI, 25:24:1) extractions were performed by adding 1 volume of PCI, rotating for 30 minutes at RT and spinning for 5 minutes at 13 krpm. DNA was precipitated in 1/5

volume NH4AC 4M + 3 volumes EtOH during an overnight incubation at 4 degrees. After the DNA spun down for 20 minutes 13 krpm at 4 degrees the pellet was washed with 70% EtOH. The DNA pellet was dissolved in 100 uL for a 2 hours RNase A (0.01 volume of 10 mg/mL) treatment at 37 degrees. Two Mg/SDS precipitations were performed on the RNA depleted DNA: Impurities spun down at 4 degrees for 5 minutes 13 krpm after adding 0.05 volumes of 10% SDS + 0.042 volumes of MgCl2 2M and incubation for 15 minutes on ice. A third PCI extraction was performed, which was followed by a chloroform:isoamylalcohol (CI, 24:1) extraction. DNA was precipitated in 2.5 volumes EtOH + 1/10 volume NaOAc 3M pH 5.2 during an overnight incubation at -20 degrees. After the DNA spun down for 30 minutes 13 krpm at 4 degrees the pellet was washed with 70% EtOH. The purified DNA was dissolved in 50 uL H2O.

MethylC-seq library generation was performed as described previously[151,152]. The genomic DNA was sonicated to an average size of 200 bp, purified and end-repaired followed by the ligation of methylated Illumina TruSeq sequencing adapters. Library amplification was performed with KAPA HiFi HotStart Uracil+ DNA polymerase (Kapa Biosystems, Woburn, MA), using 6 cycles of amplification. MethylC-seq libraries were sequenced in single-end mode on the Illumina HiSeq 1500 platform. The sequenced reads in FASTQ format were mapped to the in silico bisulfite-converted *Xenopus laevis* reference genome (Xenla9.1) using the Bowtie alignment algorithm with the following parameters: -e 120 -l 20 -n 0 as previously reported[153,154]. To estimate the bisulfite non-conversion frequency, the frequency of all cytosine base-calls at reference cytosine positions in the lambda genome (unmethylated spike in control) was normalized by the total number of base-calls at reference cytosine positions in the lambda genome. See below for sequencing and conversion statistics.

- Total sequence: 54,488,091,011
- Genome length: 2,805,684,924
- Total coverage: 19.4 X
- Covered Cs: 894,652,551
- All Cs: 956,308,623
- % Cs covered: 93.553%
- Conversion %: 99.6

14.4 Epigenetic differences explain differential expression between the L and S subgenome

To explore the gene-regulatory changes associated with gene expression differences between the L and S subgenomes, we characterized the chromatin landscape in gastrula-stage (NF stage 10.5) embryos. We determined nucleotide resolution DNA methylation levels (DNAme) by genome-wide bisulfite sequencing, and we generated profiles of the promoter mark histone H3 lysine 4 trimethylation (H3K4me3), the transcription elongation mark H3K36me3, as well as RNA polymerase II (RNAPII) and the enhancer-associated coactivator p300 using ChIP-seq.

As in most vertebrates[155,156], *X. laevis* CpG dinucleotides are globally methylated, with the notable exception of unmethylated islands which are predominantly found at the promoter regions of genes. On average the methylation level is 92% and L and S chromosomes are indistinguishable in this respect. In contrast, the global statistics of the promoter modification H3K4me3 do show a clear difference between the L and S subgenomes: The L subgenome is decorated with 28% more H3K4me3 peaks compared to the S subgenome (respectively 12,509 versus 9,682 peaks). From these peaks, 6,790 are functionally conserved between L and S (54% and 70% of the peaks on L and S, respectively), having conserved sequence as well as the H3K4me3 modification on both homoeologous sites. Compared to *Xenopus tropicalis*, S has fewer conserved H3K4me3 peaks than L: Out of 14,672 peaks in *X. tropicalis* 6,927 are functionally conserved on L, while 5,833 remain functional on S. The median conservation at the sequence level is slightly higher for L (59% and 55% for L and S respectively). Together, these results reflect a higher rate of gene promoter loss on S.

Similar to H3K4me3, the enhancer-associated co-activator p300 is also more widely associated with the L subgenome, featuring 35% more binding sites than the S genome (respectively 13,268 and 9,749 peaks). In line with high evolutionary dynamics of transcription factor binding[157], p300-bound regulatory regions are less conserved than promoter sequences in *Xenopus*; just ~2,540 peaks are functionally conserved between the L and S subgenomes (19% and 26% of the peaks on L and S, respectively). Out of 24,388 stage 10.5 enhancers in *X. tropicalis* 3,457 and 2,702 can be identified as p300 peaks on the L and S subgenomes (14% and11% for L and S respectively).

Many of the genes that are still present as homoeologous duplicates, are expressed at different levels in L and S. We wondered which regulatory features would contribute most to the L versus S expression differences. On the basis of chromatin state properties, a Random Forest machine learning algorithm can accurately predict L versus S expression bias in a set of 1,129 genes with greater than 3-fold expression difference at NF stage 10.5 (Extended Data Fig. 8f, g; mean of ROC area under the curve 0.778 with 10-fold cross-validation). Motif occurrence in gene promoters was determined using 'gimme scan' from the GimmeMotifs package[158] using motifs based on clustering a database of vertebrate motifs[159] (see also Vertebrate motif clusters v3.0. doi:10.6084/m9.figshare.1555851). Features were selected using Linear Support Vector Classification (C=0.25, l1 penalty), and classification was performed using a Random Forest Classifier (2000 trees). The model was implemented in Python using scikit[160]. The relative feature importance (Extended Data Fig. 8g) is based on the Gini importance, which is defined as the total decrease in node impurity, weighted by the probability of reaching that node, averaged over all trees of the ensemble[161]. Among various variables, the ratios of H3K4me3 and DNA methylation at the promoter contributed most to the decision tree model. A difference in p300 binding in the genomic region surrounding the gene also contributed to the Random Forest classification, as did the presence or absence of a number of specific transcription factor motifs in the promoter. These results indicate that changes at promoters and enhancers have contributed to the expression differences observed between the L and S subgenome.

Supplementary Note 15: Funding and data availability

15.1 Funding

This work was supported by the U.S. National Institute of Child Health and Human Development through grants HD065705, HD080708, P41 HD064556, GM086321 DSR, RMH AMS, JBL.TM, SM and P41 HD064556 to Xenbase; Japan Society for the Promotion of Science KAKENHI Grant Numbers, 221S0002 (A.T., Y.K., A.F.), 24590232 (S.T.), 25440180 (A.H.), 15K14521 (M.K.), 15K07082 (H.Ogino), 22570137 (T.T), 25460245 (A.S.), 23370059 (Y.I.), 23113004 (Y.M.), 25251026 (M.T.), and 22127007 (N.U.). Additional support was provided by the UNIST Research Fund Grant Number 1.150094.01, 1.150043.01 and 1.160060.01 (T.K), U.S. National Human Genome Research Institute (J.S.), the Hiroshima University Phoenix Leader Education Program (S.T.), National BioResource Project of MEXT Japan (A.S.), the US National Institute of Health Office of the Director and US National Institute of Allergy and Infectious Diseases (M.F.F., Y.O.), the US National Institute of General Medical Science (I.Q., S.H., E.M.M, J.B.W.), the US National Science Foundation (NSF), Cancer Prevention Research Institute of Texas(CPRIT) and Welch(F1515) (E.M.M.), the US NIHGM and NHLBI (J.B.W.), and the Okinawa Institute for Science and Technology Graduate University (O.S). Work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 DSR AMS, JAC, UH, SS, JC, JJ, JG, JS; J.K. was supported by an NSF graduate research fellowship, and A.M.S. was supported by a NHGRI Training Grant. Grant R01HD069344 to GJCV. Part of this work was carried out on the Dutch national e-infrastructure with the support of SURF Foundation. SJvH is supported by the Netherlands Organization for Scientific research (NWO-ALW, grant 863.12.002). O.B. is supported by an Australian Research Council Discovery Early Career Researcher Award - DECRA (DE140101962). Some sequences were generated using instruments at the UC Berkeley sequencing core purchased by NIH shared instrumentation grant 1S10OD018174.

15.2 Data availability

The XENLAv9.1 genome assembly and annotation are deposited at NCBI (accession LYTH00000000). The DNA read libraries of *X. laevis* and *X. borealis* were deposited at the Short Read Archive under accessions SRP071264 and SRP070985 respectively. Datasets of the *X. laevis* RNA-seq short reads were deposited in NCBI Gene Expression Omnibus (accession number GSE73430 for stages, GSE73419 for tissues). Datasets of the *Hymenochirus* RNA-seq short reads were deposited in NCBI GEO (accession number GSE76089). The epigenetic data have been deposited in NCBI's Gene Expression Omnibus and are accessible through GEO Series accession numbers GSE76059 for ChIP-seq. MethylC-seq data are accessible through GEO Series accession number GSE76247. Those sequence data from BAC and fosmid clones have been deposited to DDBJ/GenBank/EMBL under the accession numbers: (i) GA131508-GA227532, GA228275-GA244139, GA244852-GA274229, GA274976-GA275712, GA277157-GA344957, GA345673-GA350926, and GA351685-GA393223 for the XLB1 end-sequences, (ii) GA720358-GA756840 for the XLB2 end-sequences, (iii) GA756841-GA867435 for the XLFIC end-sequences, and (iv) AP012997-AP013026,AP014660-AP014679, AP017316, and AP017317 for the finished BAC/fosmid sequences.

References

1. Hegarty, M. J. & Hiscock, S. J. Genomic Clues to the Evolutionary Success of Polyploid Plants. *Curr. Biol.* **18,** R435–R444 (2008).
2. Otto, S. P. The evolutionary consequences of polyploidy. *Cell* **131,** 452–62 (2007).
3. Comai, L. The advantages and disadvantages of being polyploid. *Nat. Rev. Genet.* **6,** 836–46 (2005).
4. Wolfe, K. H. Yesterday's polyploids and the mystery of diploidization. *Nat. Rev. Genet.* **2,** 333–341 (2001).
5. Renny-Byfield, S. & Wendel, J. F. Doubling down on genomes: polyploidy and crop plants. *Am. J. Bot.* **101,** 1711–25 (2014).
6. Madlung, A. *et al.* Genomic changes in synthetic Arabidopsis polyploids. *Plant J.* **41,** 221–30 (2005).
7. Griffiths, S. *et al.* Molecular characterization of Ph1 as a major chromosome pairing locus in polyploid wheat. *Nature* **439,** 749–52 (2006).
8. Kobel, H. R. & Du Pasquier, L. Genetics of polyploid Xenopus. *Trends Genet.* **2,** 310–315 (1986).
9. Kobel, H. R. in *The Biology of Xenopus* (eds. Tinsley, R. C. & Kobel, H. R.) 391–401 (Oxford University Press, 1996).
10. Graf, J. D. & Kobel, H. R. Genetics of Xenopus laevis. *Methods Cell Biol.* **36,** 19–34 (1991).
11. Schmid, M., Evans, B. J. & Bogart, J. P. Polyploidy in Amphibia. *Cytogenet. Genome Res.* **145,** 315–30 (2015).
12. McClintock, B. The significance of responses of the genome to challenge. *Science* **226,** 792–801 (1984).
13. Comai, L., Madlung, A., Josefsson, C. & Tyagi, A. Do the different parental 'heteromes' cause genomic shock in newly formed allopolyploids? *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **358,** 1149–55 (2003).
14. Chen, Z. J. & Ni, Z. Mechanisms of genomic rearrangements and gene expression changes in plant polyploids. *Bioessays* **28,** 240–52 (2006).
15. Song, Q. & Chen, Z. J. Epigenetic and developmental regulation in plant polyploids. *Curr. Opin. Plant Biol.* **24,** 101–9 (2015).
16. Feldman, M. *et al.* Rapid Elimination of Low-Copy DNA Sequences in Polyploid Wheat: A Possible Mechanism for Differentiation of Homoeologous Chromosomes. *Genetics* **147,** 1381–1387 (1997).
17. Xiong, Z., Gaeta, R. T. & Pires, J. C. Homoeologous shuffling and chromosome compensation maintain genome balance in resynthesized allopolyploid Brassica napus. *Proc. Natl. Acad. Sci.* **108,** 7908–7913 (2011).

18. Cao, D., Osborn, T. C. & Doerge, R. W. Correct estimation of preferential chromosome pairing in autotetraploids. *Genome Res.* **14,** 459–62 (2004).

19. Lynch, M. & Force, A. The probability of duplicate gene preservation by subfunctionalization. *Genetics* **154,** 459–73 (2000).

20. Haldane, J. B. S. The rate of spontaneous mutation of a human gene. *J. Genet.* **31,** 317–326 (1935).

21. Schlager, G. & Dickie, M. M. Spontaneous mutation rates at five coat-color loci in mice. *Science* **151,** 205–6 (1966).

22. Gillespie, J. H. Population Genetics: A Concise Guide. *Johns Hopkins University Press* (2004). at <http://www.amazon.com/Population-Genetics-A-Concise-Guide/dp/0801880092>

23. Ohno, S. *Evolution by Gene Duplication*. (Springer Berlin Heidelberg, 1970). doi:10.1007/978-3-642-86659-3

24. Birchler, J. A. & Veitia, R. A. Gene balance hypothesis: connecting issues of dosage sensitivity across biological disciplines. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 14746–53 (2012).

25. Conant, G. C., Birchler, J. A. & Pires, J. C. Dosage, duplication, and diploidization: clarifying the interplay of multiple models for duplicate gene evolution over time. *Curr. Opin. Plant Biol.* **19,** 91–8 (2014).

26. Force, A. *et al.* Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151,** 1531–45 (1999).

27. Tochinai, S. & Katagiri, C. Complete abrogation of immune response to skin allografts and rabbit erythrocytes in the early thymectomized Xenopus. *Dev. Growth …* **17,** 383–394 (1975).

28. Izutsu, Y. & Yoshizato, K. Metamorphosis-dependent recognition of larval skin as non-self by inbred adult frogs (Xenopus laevis). *J. Exp. Zool.* **266,** 163–7 (1993).

29. Sive, H. L., Grainger, R. M. & Harland, R. M. *Early development of Xenopus laevis: a laboratory manual*. (Cold Spring Harbor Laboratory Press Cold Spring Harbor, NY:, 2000).

30. Fujiyama, A. *et al.* Construction and analysis of a human-chimpanzee comparative clone map. *Science* **295,** 131–4 (2002).

31. Hattori, M. *et al.* The DNA sequence of human chromosome 21. *Nature* **405,** 311–9 (2000).

32. Kitzman, J. O. *et al.* Haplotype-resolved genome sequencing of a Gujarati Indian individual. *Nat. Biotechnol.* **29,** 59–63 (2011).

33. Gnerre, S. *et al.* High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 1513–8 (2011).

34. Adey, A. *et al.* The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line. *Nature* **500,** 207–11 (2013).

35. Morin, R. D. *et al.* Sequencing and analysis of 10,967 full-length cDNA clones from Xenopus laevis and Xenopus tropicalis reveals post-tetraploidization transcriptome remodeling. *Genome Res.* **16,** 796–803 (2006).

36. Hellsten, U. *et al.* Accelerated gene evolution and subfunctionalization in the pseudotetraploid frog Xenopus laevis. *BMC Biol.* **5,** 31 (2007).

37. Chapman, J. A. *et al.* Meraculous: de novo genome assembly with short paired-end reads. *PLoS One* **6,** e23501 (2011).

38. Bradnam, K. R. *et al.* Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *Gigascience* **2,** 10 (2013).

39. Thiébaud, C. H. & Fischberg, M. DNA content in the genus Xenopus. *Chromosoma* **59,** 253–7 (1977).

40. Luo, R. *et al.* SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* **1,** 18 (2012).

41. Putnam, N. H. *et al.* Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res.* **26,** 342–50 (2016).

42. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–10 (1990).

43. Li, H. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* **30,** 2843–51 (2014).

44. Krumsiek, J., Arnold, R. & Rattei, T. Gepard: a rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics* **23,** 1026–8 (2007).

45. Matsuda, Y. & Chapman, V. M. Application of fluorescence in situ hybridization in genome analysis of the mouse. *Electrophoresis* **16,** 261–72 (1995).

46. Uno, Y., Nishida, C., Takagi, C., Ueno, N. & Matsuda, Y. Homoeologous chromosomes of Xenopus laevis are highly conserved after whole-genome duplication. *Heredity (Edinb).* **111,** 430–6 (2013).

47. Matsuda, Y. *et al.* A New Nomenclature of Xenopus laevis Chromosomes Based on the Phylogenetic Relationship to Silurana/Xenopus tropicalis. *Cytogenet. Genome Res.* **145,** 187–91 (2015).

48. Burton, J. N. *et al.* Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat. Biotechnol.* **31,** 1119–25 (2013).

49. Kaplan, N. & Dekker, J. High-throughput genome scaffolding from in vivo DNA interaction frequency. *Nat. Biotechnol.* **31,** 1143–7 (2013).

50. Kalhor, R., Tjong, H., Jayathilaka, N., Alber, F. & Chen, L. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat. Biotechnol.* **30,** 90–8 (2012).

51. Meyer, M. & Kircher, M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb. Protoc.* **2010,** pdb.prot5448 (2010).

52. Zaharia, M. *et al.* Faster and More Accurate Sequence Alignment with SNAP. (2011).

53. Wang, Y. *et al.* MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40,** e49 (2012).

54. Gilchrist, M. J. From expression cloning to gene modeling: the development of Xenopus gene sequence resources. *Genesis* **50,** 143–54 (2012).

55. Zerbino, D. R. Using the Velvet de novo assembler for short-read sequencing technologies. *Curr. Protoc. Bioinformatics* **Chapter 11,** Unit 11.5 (2010).

56. Schulz, M. H., Zerbino, D. R., Vingron, M. & Birney, E. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* **28,** 1086–92 (2012).

57. Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21,** 1859–1875 (2005).

58. Simakov, O. *et al.* Insights into bilaterian evolution from three spiralian genomes. *Nature* **493,** 526–31 (2013).

59. Smit, AFA, Hubley, R & Green, P. RepeatMasker Open-4.0. http://www.repeatmasker.org

60. Price, A. L., Jones, N. C. & Pevzner, P. A. De novo identification of repeat families in large genomes. *Bioinformatics* **21 Suppl 1,** i351-8 (2005).

61. Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110,** 462–467 (2005).

62. Smit, AFA, Hubley, R. RepeatModeler Open-1.0. http://www.repeatmasker.org

63. Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9,** R7 (2008).

64. Yeh, R.-F., Lim, L. P. & Burge, C. B. Computational Inference of Homologous Gene Structures in the Human Genome. *Genome Res.* **11,** 803–816 (2001).

65. Salamov, A. A. & Solovyev, V. V. Ab initio Gene Finding in Drosophila Genomic DNA. *Genome Res.* **10,** 516–522 (2000).

66. Haas, B. J. *et al.* Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31,** 5654–66 (2003).

67. Kent, W. J. BLAT---The BLAST-Like Alignment Tool. *Genome Res.* **12,** 656–664 (2002).

68. Keller, O., Odronitz, F., Stanke, M., Kollmar, M. & Waack, S. Scipio: Using protein sequences to determine the precise exon/intron structures of genes and their orthologues in closely related species. *BMC Bioinformatics* **9,** 278 (2008).

69. Smits, A. H. *et al.* Global absolute quantification reveals tight regulation of protein expression in single Xenopus eggs. *Nucleic Acids Res.* **42,** 9880–9891 (2014).

70. Uno, Y. *et al.* Inference of the Protokaryotypes of Amniotes and Tetrapods and the Evolutionary Processes of Microchromosomes from Comparative Gene Mapping. *PLoS One* **7,** e53027 (2012).

71. Edwards, N. S. & Murray, A. W. Identification of xenopus CENP-A and an associated centromeric DNA repeat. *Mol. Biol. Cell* **16,** 1800–10 (2005).

72. Chang, C. Y. & Witschi, E. Genic control and hormonal reversal of sex differentiation in Xenopus. *Proc. Soc. Exp. Biol. Med.* **93,** 140–4 (1956).

73. Yoshimoto, S. *et al.* A W-linked DM-domain gene, DM-W, participates in primary ovary development in Xenopus laevis. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 2469–74 (2008).

74. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30,** 3059–66 (2002).

75. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **5,** e9490 (2010).

76. Jukes, T. H. & Cantor, C. R. Evolution of protein molecules. *Mamm. protein Metab.* **3,** 21–132 (1969).

77. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10,** 421 (2009).

78. Gerhard, D. S. *et al.* The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC). *Genome Res.* **14,** 2121–7 (2004).

79. Gout, J.-F., Kahn, D., Duret, L. & Paramecium Post-Genomics Consortium. The relationship among gene expression, the evolution of gene dosage, and the rate of protein evolution. *PLoS Genet.* **6,** e1000944 (2010).

80. Peshkin, L. *et al.* On the Relationship of Protein and mRNA Dynamics in Vertebrate Embryonic Development. *Dev. Cell* **35,** 383–94 (2015).

81. Ranwez, V., Harispe, S., Delsuc, F. & Douzery, E. J. P. MACSE: Multiple Alignment of Coding SEquences accounting for frameshifts and stop codons. *PLoS One* **6,** e22594 (2011).

82. Charif, D., Thioulouse, J., Lobry, J. R. & Perrière, G. Online synonymous codon usage analyses with the ade4 and seqinR packages. *Bioinformatics* **21,** 545–7 (2005).

83. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32,** 1792–7 (2004).

84. Talavera, G. & Castresana, J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.* **56,** 564–77 (2007).

85. Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol. Biol. Evol.* **30,** 2725–9 (2013).

86. Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59,** 307–21 (2010).

87. Sanderson, M. J. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* **19,** 301–2 (2003).

88. Christenson, M. K. *et al.* De novo Assembly and Analysis of the Northern Leopard Frog Rana pipiens Transcriptome. *J. genomics* **2,** 141–9 (2014).

89. Cannatella, D. Xenopus in Space and Time: Fossils, Node Calibrations, Tip-Dating, and Paleobiogeography. *Cytogenet. Genome Res.* **145,** 283–301 (2015).

90. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. (2013).

91. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-

generation DNA sequencing data. *Genome Res.* **20,** 1297–1303 (2010).

92. Nei, M. & Gojobori, T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3,** 418–26 (1986).

93. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24,** 1586–91 (2007).

94. Lund, E., Liu, M., Hartley, R. S., Sheets, M. D. & Dahlberg, J. E. Deadenylation of maternal mRNAs mediated by miR-427 in Xenopus laevis embryos. *RNA* **15,** 2351–63 (2009).

95. Lee, A. P., Kerk, S. Y., Tan, Y. Y., Brenner, S. & Venkatesh, B. Ancient vertebrate conserved noncoding elements have been evolving rapidly in teleost fishes. *Mol. Biol. Evol.* **28,** 1205–15 (2011).

96. Paten, B. *et al.* Cactus: Algorithms for genome multiple sequence alignment. *Genome Res.* **21,** 1512–28 (2011).

97. Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20,** 289–90 (2004).

98. Zhang, Z. D., Frankish, A., Hunt, T., Harrow, J. & Gerstein, M. Identification and analysis of unitary pseudogenes: historic and contemporary gene losses in humans and other primates. *Genome Biol.* **11,** R26 (2010).

99. Slater, G. S. C. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6,** 31 (2005).

100. Bustamante, C. D., Nielsen, R. & Hartl, D. L. A maximum likelihood method for analyzing pseudogene evolution: implications for silent site evolution in humans and rodents. *Mol. Biol. Evol.* **19,** 110–7 (2002).

101. Chou, H.-H. *et al.* Inactivation of CMP-N-acetylneuraminic acid hydroxylase occurred prior to brain expansion during human evolution. *Proc. Natl. Acad. Sci. U. S. A.* **99,** 11736–41 (2002).

102. Meredith, R. W., Gatesy, J., Murphy, W. J., Ryder, O. A. & Springer, M. S. Molecular Decay of the Tooth Gene Enamelin (ENAM) Mirrors the Loss of Enamel in the Fossil Record of Placental Mammals. *PLoS Genet.* **5,** e1000634 (2009).

103. Finn, R. D. *et al.* Pfam: the protein families database. *Nucleic Acids Res.* **42,** D222-30 (2014).

104. Mitchell, A. *et al.* The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res.* **43,** D213–D221 (2015).

105. Calvo, S. E., Clauser, K. R. & Mootha, V. K. MitoCarta2.0: an updated inventory of mammalian mitochondrial proteins. *Nucleic Acids Res.* (2015). doi:10.1093/nar/gkv1003

106. Pagliarini, D. J. *et al.* A mitochondrial protein compendium elucidates complex I disease biology. *Cell* **134,** 112–23 (2008).

107. Wan, C. *et al.* Panorama of ancient metazoan macromolecular complexes. *Nature* **525,** 339–44 (2015).

108. Makino, T. & McLysaght, A. Ohnologs in the human genome are dosage balanced and frequently

associated with disease. *Proc. Natl. Acad. Sci.* **107,** 9270–9274 (2010).

109. Berthelot, C. *et al.* The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. *Nat. Commun.* **5,** 3657 (2014).

110. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12,** 323 (2011).

111. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9,** 559 (2008).

112. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4,** 406–25 (1987).

113. Schwartz, S. PipMaker---A Web Server for Aligning Two Genomic DNA Sequences. *Genome Res.* **10,** 577–586 (2000).

114. Ogino, H., Fisher, M. & Grainger, R. M. Convergence of a head-field selector Otx2 and Notch signaling: a mechanism for lens specification. *Development* **135,** 249–58 (2008).

115. Lolli, G. & Johnson, L. N. CAK-Cyclin-dependent Activating Kinase: a key kinase in cell cycle control and a target for drugs? *Cell Cycle* **4,** 572–7 (2005).

116. Lehti-Shiu, M. D. & Shiu, S.-H. S.-H. Diversity, classification and function of the plant protein kinase superfamily. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **367,** 2619–39 (2012).

117. Wang, G. *et al.* Genome-wide analysis of the cyclin family in Arabidopsis and comparative phylogenetic analysis of plant cyclin-like proteins. *Plant Physiol.* **135,** 1084–99 (2004).

118. Kobel, H. R. & Tinsley, R. C. *The Biology of Xenopus*. *The Biology of Xenopus* (Oxford University Press, 1996).

119. Edens, L. J. & Levy, D. L. Size scaling of subcellular organelles and structures in Xenopus laevis and Xenopus tropicalis. *Xenopus Dev.* 325–345 (2014).

120. Yanai, I., Peshkin, L., Jorgensen, P. & Kirschner, M. W. Mapping gene expression in two Xenopus species: evolutionary constraints and developmental flexibility. *Dev. Cell* **20,** 483–96 (2011).

121. Birsoy, B., Kofron, M., Schaible, K., Wylie, C. & Heasman, J. Vg1 is an essential signaling molecule in Xenopus development. *Development* **133,** 15–20 (2006).

122. Courtet, M., Flajnik, M. & Du Pasquier, L. Major histocompatibility complex and immunoglobulin loci visualized by in situ hybridization on Xenopus chromosomes. *Dev. Comp. Immunol.* **25,** 149–57 (2001).

123. Evans, B. J., Kelley, D. B., Melnick, D. J. & Cannatella, D. C. Evolution of RAG-1 in polyploid clawed frogs. *Mol. Biol. Evol.* **22,** 1193–207 (2005).

124. Anderson, D. W. & Evans, B. J. Regulatory evolution of a duplicated heterodimer across species and tissues of allopolyploid clawed frogs (Xenopus). *J. Mol. Evol.* **68,** 236–47 (2009).

125. Du Pasquier, L., Wilson, M. & Sammut, B. The fate of duplicated immunity genes in the dodecaploid Xenopus ruwenzoriensis. *Front. Biosci. (Landmark Ed.* **14,** 177–91 (2009).

126. Nonaka, M. *et al.* Major histocompatibility complex gene mapping in the amphibian Xenopus implies a primordial organization. *Proc. Natl. Acad. Sci. U. S. A.* **94,** 5789–91 (1997).

127. Stewart, R. *et al.* Cloning and characterization of Xenopus beta2-microglobulin. *Dev. Comp. Immunol.* **29,** 723–32 (2005).

128. Flajnik, M. F. *et al.* A novel type of class I gene organization in vertebrates: a large family of non-MHC-linked class I genes is expressed at the RNA level in the amphibian Xenopus. *EMBO J.* **12,** 4385–96 (1993).

129. Ishii, A., Kawasaki, M., Matsumoto, M., Tochinai, S. & Seya, T. Phylogenetic and expression analysis of amphibian Xenopus Toll-like receptors. *Immunogenetics* **59,** 281–93 (2007).

130. Buckley, K. M. & Rast, J. P. Dynamic evolution of toll-like receptor multigene families in echinoderms. *Front. Immunol.* **3,** 136 (2012).

131. Hong, J.-H. *et al.* TAZ, a transcriptional modulator of mesenchymal stem cell differentiation. *Science* **309,** 1074–8 (2005).

132. Azzolin, L. *et al.* Role of TAZ as mediator of Wnt signaling. *Cell* **151,** 1443–56 (2012).

133. Koontz, L. M. *et al.* The Hippo effector Yorkie controls normal tissue growth by antagonizing scalloped-mediated default repression. *Dev. Cell* **25,** 388–401 (2013).

134. Udan, R. S., Kango-Singh, M., Nolo, R., Tao, C. & Halder, G. Hippo promotes proliferation arrest and apoptosis in the Salvador/Warts pathway. *Nat. Cell Biol.* **5,** 914–20 (2003).

135. Ingham, P. W. & Placzek, M. Orchestrating ontogenesis: variations on a theme by sonic hedgehog. *Nat. Rev. Genet.* **7,** 841–50 (2006).

136. Takahashi, S. *et al.* Nodal-related gene Xnr5 is amplified in the Xenopus genome. *Genesis* **44,** 309–21 (2006).

137. Houston, D. W. & King, M. L. A critical role for Xdazl, a germ plasm-localized RNA, in the differentiation of primordial germ cells in Xenopus. *Development* **127,** 447–56 (2000).

138. Seydoux, G. & Braun, R. E. Pathway to Totipotency: Lessons from Germ Cells. *Cell* **127,** 891–904 (2006).

139. Houston, D. W. & King, M. L. Germ plasm and molecular determinants of germ cell fate. *Curr. Top. Dev. Biol.* **50,** 155–81 (2000).

140. Yamaguchi, T., Taguchi, A., Watanabe, K. & Orii, H. DEADSouth protein localizes to germ plasm and is required for the development of primordial germ cells in Xenopus laevis. *Biol. Open* **2,** 191–199 (2013).

141. Horvay, K., Claußen, M., Katzer, M., Landgrebe, J. & Pieler, T. Xenopus Dead end mRNA is a localized maternal determinant that serves a conserved function in germ cell development. *Dev. Biol.* **291,** 1–11 (2006).

142. Weidinger, G. *et al.* dead end, a novel vertebrate germ plasm component, is required for zebrafish primordial germ cell migration and survival. *Curr. Biol.* **13,** 1429–34 (2003).

143. Lai, F., Singh, A. & King, M. Lou. Xenopus Nanos1 is required to prevent endoderm gene

expression and apoptosis in primordial germ cells. *Development* **139,** 1476–86 (2012).

144. Kloc, M. & Chan, A. P. Centroid, a novel putative DEAD-box RNA helicase maternal mRNA, is localized in the mitochondrial cloud in Xenopus laevis oocytes. *Int. J. Dev. Biol.* **51,** 701–6 (2007).

145. Berekelya, L. A. *et al.* The protein encoded by the germ plasm RNA Germes associates with dynein light chains and functions in Xenopus germline development. *Differentiation.* **75,** 546–58 (2007).

146. Lau, N. C., Ohsumi, T., Borowsky, M., Kingston, R. E. & Blower, M. D. Systematic and single cell analysis of Xenopus Piwi-interacting RNAs and Xiwi. *EMBO J.* **28,** 2945–2958 (2009).

147. Houwing, S. *et al.* A Role for Piwi and piRNAs in Germ Cell Maintenance and Transposon Silencing in Zebrafish. *Cell* **129,** 69–82 (2007).

148. Houston, D. W. Regulation of cell polarity and RNA localization in vertebrate oocytes. *Int. Rev. Cell Mol. Biol.* **306,** 127–85 (2013).

149. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.* **9,** R137 (2008).

150. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–2 (2010).

151. Lister, R. *et al.* Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature* **471,** 68–73 (2011).

152. Urich, M. A., Nery, J. R., Lister, R., Schmitz, R. J. & Ecker, J. R. MethylC-seq library preparation for base-resolution whole-genome bisulfite sequencing. *Nat. Protoc.* **10,** 475–483 (2015).

153. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).

154. Lister, R. *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462,** 315–322 (2009).

155. Long, H. K. *et al.* Epigenetic conservation at gene regulatory elements revealed by non-methylated DNA profiling in seven vertebrates. *Elife* **2,** e00348 (2013).

156. Deaton, A. M. & Bird, A. CpG islands and the regulation of transcription. *Genes Dev.* **25,** 1010–22 (2011).

157. Schmidt, D. *et al.* Five-Vertebrate ChIP-seq Reveals the Evolutionary Dynamics of Transcription Factor Binding. *Science* **328,** 1036–1040 (2010).

158. van Heeringen, S. J. & Veenstra, G. J. C. GimmeMotifs: a de novo motif prediction pipeline for ChIP-sequencing experiments. *Bioinformatics* **27,** 270–271 (2011).

159. Weirauch, M. T. *et al.* Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity. *Cell* **158,** 1431–1443 (2014).

160. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12,** 2825–2830 (2011).

161. Breiman, L., Friedman, J., Stone, C. J. & Olshen, R. A. *Classification and regression trees*. (CRC

press, 1984).