Transcriptomic Insights into Genetic Diversity of Protein-Coding Genes in *X. laevis*.

Authors: Virginia Savova[1,3], Esther J Pearl[2], Elvan Boke[1], Anwesha Nag[3], Ivan Adzhubei[4], Marko E. Horb[2], Leonid Peshkin[1#]

Affiliations:

[1] Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA.

[2] National Xenopus Resource and Bell Center for Regenerative Biology and Tissue Engineering, Marine Biological Laboratory, Woods Hole, MA 02543, USA.

[3] Dana-Farber Cancer Institute and Department of Genetics, Harvard Medical School, 450 Brookline Ave., Boston, MA 02215, USA☒

[4] Division of Genetics, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA

[#] Corresponding author

Abstract

We characterize the genetic diversity of *Xenopus laevis* strains using RNA-seq data and allele-specific analysis. This data provides a catalogue of coding variation, which can be used for improving the genomic sequence, as well as for better sequence alignment, probe design, and proteomic analysis. In addition, we paint a broad picture of the genetic landscape of the species by functionally annotating different classes of mutations with a well-established prediction tool (PolyPhen-2). Further, we specifically compare the variation in the progeny of four crosses: inbred genomic (J)-strain, outbred albino (B)-strain, and two hybrid crosses of J and B strains. We identify a subset of mutations specific to the B strain, which allows us to investigate the selection pressures affecting duplicated genes in this allotetraploid. From these crosses we find the ratio of non-synonymous to synonymous mutations is lower in duplicated genes, which suggests that they are under greater purifying selection. Surprisingly, we also find that function-altering ("damaging") mutations constitute a greater fraction of the non-synonymous variants in this group, which suggests a role for subfunctionalization in coding variation affecting duplicated genes.

Introduction

A systematic understanding of the genetic basis of human disease and its underlying cellular and molecular mechanisms is dependent on model organisms that can both capture the pathology under investigation and provide tools for functional studies. From mammalian to invertebrate models, each organism offers tradeoffs between genomic and physiologic similarity to humans as well as the availability, scalability and cost of functional assays. Historically, *Xenopus* has been an excellent model for cell biological and embryological studies, such as for functional analysis of cell cycle control and cloning (Gurdon, 2013; Hunt, 2002), but has been less utilized as a genetic model. The recent arrival of high quality chromosome level genomic sequence and protein coding gene models for *X. laevis* (and *X. tropicalis*) is once again bringing *Xenopus* to the forefront of biological research allowing for more genomic and genetic analyses. Already remarkably useful in enabling many unique biochemical approaches in cell biology, embryology, and cell type differentiation, the availability of high quality genome sequence data will only reinforce its strength as a model organism.

Xenopus is a valuable vertebrate model system because of its large size, availability of several hundred embryos from each mating, rapid and synchronous *ex vivo* development (major organ formation occurs 48 hours post-fertilization), similar organ systems (e.g. lungs, limb) and ease of care. In addition, a large fraction of *Xenopus* genes have identifiable human orthologs for which both transcriptomic and proteomic temporal expression in early development have been well characterized (Peshkin et al., 2015; Yanai et al., 2011). In addition, orthologous human mRNAs (mutant or wild type) are commonly used to rescue the phenotype of *Xenopus* mutants illustrating the benefits of studying human disease causing genes in *Xenopus* (Davis et al., 2014; Pearl et al., 2011). These features make complementation studies (del Viso et al., 2012) in *Xenopus* both possible and attractive.

In this paper, we begin to examine how such genetic diversity might affect the use of different inbred, transgenic and wild type animals. Immunology studies notwithstanding (Gantress et al., 2003; Izutsu and Maéno, 2005), *Xenopus* researchers have historically used animals that are genetically diverse, with little regard to their specific background. To this day, many laboratories maintain colonies of *Xenopus* animals from non-genomic lines, transgenic animals or simply wild type animals. Such genetic diversity in laboratory populations creates both issues and opportunities for research, but a systemic analysis of these issues has not been done. Our study presents a rich resource of data, enabling initial characterization of high-confidence single-nucleotide polymorphisms (SNPs) and other variation (insertions and deletions a.k.a. INDELs). We discuss the general features of the data and review strategies of how researchers can take advantage of these genetic variations.

The *Xenopus laevis* genome assembly is based on sequencing of the inbred J strain (Gantress et al., 2003) and characterizing the diversity of the genomic line itself can contribute significantly to experimental design, since accurate sequence is crucial for design and implementation of RNAi, morpholinos, and more recently for CRISPR and mass spectrometry. To help address this diversity, we investigate the residual heterozygosity of genomic J strain colonies that is due to the fact that the J strain is not a completely inbred population. We also provide a characterization of the level of polymorphisms between the J strain and an outbred line, using transcriptome sequencing of the albino B strain. This characterization is important for two reasons. First, with advances in genomic

analyses in *Xenopus* (e.g. RNA-Seq, proteomics, genome editing) our data underlines the importance of using a defined genetic strain for such experiments when trying to analyze the data. Second, crossing an inbred strain with another line whose genomic variation is reproducible and characterized in depth may provide insights into functional and regulatory variation driven by differences between alleles. This type of experimental design has been utilized extensively in other model organisms (Cui et al., 2006).

Our analysis of several SNPs in protein coding genes led us to re-examine the gene duplication phenomenon. *X. laevis* genome underwent allotetraploidization between 18 Mya (Session et al., 2016) to 40 Mya (Hellsten et al., 2007). As a result, duplicated gene function may either be a) retained in both copies making them functionally redundant; b) lost in one of the copies where it becomes a pseudogene; c) a novel and divergent gene function is acquired by one of the two copies; or d) the original gene function becomes split between the two duplicated copies, either in location or effect. Since mutations in only one of two functionally redundant orthologs might not display a phenotype, one may hypothesize that purifying selection would be relaxed. We asked to what extent potentially deleterious mutations occur on duplicated genes. Our results confirm previous reports that duplicated genes have lower mutation rates, however, further analyses of the nonsynonymous variants reveal unexpected differences potentially attributable to subfunctionalization.

Materials and Methods

Animals**.** We chose two strains of *X. laevis*: "J strain" (RRID:NXR_0.0024) (Gantress et al., 2003) as provided by co-author Marko Horb from the National *Xenopus* Resource Center (NXR, RRID:SCR_013731), and "B strain" as provided by co-author Leonid Peshkin from the Harvard Medical School (**Figure 1**). The J strain was obtained from Jacques Robert (University of Rochester Medical Center) specifically for genome sequencing due to its high level of inbreeding. It is therefore expected to show significantly lower heterozygosity than wildtype. It is estimated to be inbred for 32 generations (Gantress et al., 2003; Tochinai and Katagiri, 1975). The J strain animals used in this study were second generation (F34) offspring of parents from the same clutch as those used for the sequencing characterization (J. Robert's personal communication). The B strain is a line of inbred albino animals with estimated inbreeding of at least 10 generations. According to Olga Hoperskaya (Hoperskaya, 1975), albinos appeared in the *Xenopus* colony at the Institute of Developmental Biology, Moscow, in 1972. These were indirectly imported to the Berkeley colony in the 1980's. At Berkeley the strain had some edema and was occasionally outbred to pigmented frogs in the mid-1980s, and crossed to get back the albino without edema (John Gerhart, personal communications). The frog facility at HMS was founded in 1993-1994 when Marc Kirschner moved there from UCSF and brought a small group of albino frogs with him. To the best of our knowledge, no animals were brought into the facility between 1993 and 2000 and the group was inbred every 2-3 years. Since 2000, no animals were brought into the HMS colony from an outside source, which leads us to estimate that they have been inbred for approximately 10 generations since 1993, with moderate inbreeding before that.

Hybrid crosses and sequencing**.** We successfully performed natural mating of the two *Xenopus* strains: two first-generation hybrid (F1) and two straight self (JxJ, BxB) crosses. We then collected tadpoles at a single developmental time-point (stage NF-42), pooled ten tadpoles per cross, and

isolated RNA from each pool. After RiboZero treatment, we constructed four Illumina libraries, and performed RNAseq on HiSeq 2000 platform, resulting in approximately 30 to 47 million reads per library with paired-end 100 base reads (see **Table S1** for details). The F1 libraries were pooled *in-silico* and analyzed together.

Data analysis. The *X. laevis* genome sequence data was downloaded from Xenbase (RRID:SCR_003280), including the gene models and respective names**.** We used the most up to date genome assembly (Session et al., 2016) from files named "XL9.1_annot_v1.8.1.primary*" released on Dec 9, 2015 on Xenbase (ftp.xenbase.org/pub/Genomics/JGI) . RNA-Seq reads were soft-trimmed by quality, and stripped of remaining Illumina adapter sequences using Trimmomatic (Bolger et al., 2014). We used paired end alignment which means that the mapping information from both the forward and the reverse read was used simultaneously to determine the mapping location and possibly disambiguate mapping in situation when one of the reads in a pair was a match to multiple mapping loci. We aligned the paired reads and aligned to the JGI9.1 assembly using *tophat* ver. 2.1.0 (Trapnell et al., 2009), allowing up to five mismatches, read-gap length of 3, and edit-distance of 5 (command line parameters used: tophat -N 5 --read-gap-length 3 --read-edit-dist 5.) Edit-distance for *tophat* is a parameter that controls simultaneously the permitted mismatches and the permitted INDELS and refers to the minimal number of changes that need to be made to the mapped read in order to obtain a sequence perfectly matched to the genome. The --N parameter controls only the number of mismatches. So for example, with read-edit-dist of 5 we allow 3 mismatches and 2 deletions, but not 5 mismatches and 3 deletions). We used the samtools ver. 1.3 (Li et al., 2009) *mpileup* tool to count the number of bases observed in each covered position. We then processed the output with *bcftools* (Li et al., 2009) to obtain and characterize variant calls by depth (number of reads covering the base) and quality (a phred-scaled score designed to reflect confidence in the identification of the nucleobases by the sequencer).

For a conservative approach to SNP discovery in an allotetraploid genome we discarded ambiguous reads, defined as reads with multiple acceptable alignments under the specified parameters (see above). To do so, we used a mapping quality filter on the alignments used for SNP calling; the mapping quality of ambiguously mapped reads produced by the aligner is indicated as 0, while we required mapping quality of at least 30 (switch –q 30; specifying minimal alignment quality of the read) to discard any ambiguous or low quality alignments (7% of the mapped reads, **Table S1**); otherwise, mappings across duplicated genes would have resulted in many false SNPs resulting from cross-mappings of reads from the two homeologs. The variants were further annotated as synonymous or missense using *SNPeff* (Cingolani et al., 2012) software and further qualified as benign or damaging using PolyPhen-2 web server (Adzhubei et al., 2010). SNPeff uses the gene models and genetic code to label each SNP as non-coding or coding; coding SNPs were then labeled synonymous or non-synonymous. SNPeff annotation thus could be easily reproduced from VCF files, which are made available as part of this publication via NCBI GEO entry GSE74470. PolyPhen is a method and a server which assigns a score between 0 and 1 to a missense SNP, reflecting the likelihood that a given variant will affect the protein's function. The score is further categorized as "benign", "uncertain", "possibly damaging" and "probably damaging". Values of dN/dS were obtained by first calculating the ratio of nonsynonymous to synonymous mutations (D_n/D_s). The number of

nonsynonymous and synonymous sites was calculated using the PAML software package (Yang, 2007) from the primary transcript files.

Results

Remaining genetic diversity in the colonies of the reference *Xenopus laevis* strain (J strain)

To gauge the RNA-encoding sequence diversity remaining within the J strain colony, we performed RNA-sequencing and compared cDNA sequences to the *X. laevis* reference genome, which is based on that strain. We prepared cDNA libraries from pooled stage NF-42 tadpoles (Nieuwkoop and Faber, 1994) of J strain, and aligned the sequence reads to genome assembly version 9.1, available on Xenbase. There were 347,850 high-confidence (QUAL 30, 97% probability of correct call according to *bcftools*) single-nucleotide variants in cDNA of J-strain animals compared to the reference J strain genome sequence.

Of these, transitions outnumbered transversions (215,529 ts, 132,321 tv) and constituted 62% of the total. This rate is lower than would be expected in human (67%, $p < 0.000001$) based on the 1000 Genomes Project data (DePristo et al., 2011). Since the ratio is stable across the quality range and the per-base sequence quality is high, we interpret the difference to be due to the relatively higher GC content of the *Xenopus* genome, which has been shown to influence this measure (Wang et al., 2015a). There were 38,354 INDELs. Many more possible variants were detected at lower confidence thresholds (**Figure 2**), but researchers with specific interest in any such candidate loci should validate the target loci.

Of the SNPs covered by 10 or more reads, which afford a >99% chance of detecting both parental variants, 74% were heterozygous, indicating that the diversity within the J strain colony remains significant, constituting approximately 0.05% of all bases covered at that level (27,688 of 61,667,356 bases). However, at this depth threshold, the fraction of heterozygous SNPs is higher than expected (Wang et al., 2015b) from the Hardy-Weinberg equilibrium (0.67%), suggesting that we are overestimating heterozygous SNPs, possibly due to alignment errors and/or deviations from diploidy. Indeed, higher depth thresholds provide a better estimate of true heterozygosity (**Figure S1**). As for the homozygous variants, these derive from three possible sources: 1) sequencing and assembly errors; 2) wildtype variants that were not expunged by inbreeding and may be homozygous in the particular cross; 3) true homozygous variants (presumably mostly derived from wildtype) that appeared post-genome sequencing specifically in the MBL colony. Since the MBL J strain colony was started from a single pair, there is some possibility that a fraction of the homozygous SNPs may be truly homozygous in that colony. Overall, the fraction of homozygous SNPs was observed to be higher at highly covered positions (32% of all SNPs and 0.02% of the genome, 1,129/3,564/6,114,919). Based on these positions, our estimate of the rate of homozygous SNPs is 0.02%, and the fraction of heterozygous SNPs among all SNPs is 69%, which is comparable to an occasionally outbred human population [cf. Figure 1 of (Wang et al., 2015b)].

Thus, our findings improve our understanding of the genetic landscape of *Xenopus laevis* by adding SNP information that will be useful for planning of future experiments that require precise sequence alignment, such as CRISPR-Cas genome editing applications. The comparable levels of

heterozygosity of J strain and human populations also suggest *Xenopus* can be used to model some aspects of variation in human populations.

A catalogue of SNPs in the B strain

Next, we wanted to characterize SNPs in the albino strain, which we call the "B strain". Albino animals are very useful for imaging studies in *Xenopus*, and can be used to create different albino transgenic and mutant lines by breeding of albino and pigmented animals. For work in such animals, accurate information about variants harbored in the albino colony should assist in primer and morpholino design as well as genome editing. Additionally, known polymorphisms could be used to pursue allele-specific characterization. For example, the temporal and spatial distribution of maternally deposited mRNA can be studied in a F1 hybrid cross. The SNPs we report are a useful resource for genomic research with that strain.

As expected for a divergent line, the amount of overall variation observed in the B strain is significantly higher than the J strain (708,820 SNPs of QUAL 30, ts: 403,886, tv: 304,934, indel: 62,748). In fact, the number of SNPs is more than twice than that observed in J strain, and this remains the case when we control for depth of coverage (**Figure S1**). Furthermore, the vast majority of identified SNPs are heterozygous, which was expected because the strain has not been strictly inbred and has been subject to periodic outbreeding. Interestingly, the ts/tv ratio for the B strain is even lower than the ratio observed for J strain (1.3 versus 1.6, p<0.0001, **Figure 2**). Since the samples were sequenced on the same lane, the lower ratio cannot be due to variation in sequencing quality but may indicate higher mismapping rates. Thus, as expected, the non-genomic inbred strain presents a deeper reservoir of genetic diversity that can be exploited in genetic studies.

Characterization of differences between J strain and occasionally outbred B strain

Using the SNPs identified in the RNA-Seq experiments, we estimated the expected sequence diversity between the J strain and an occasionally outbred line, the B strain. Specifically, within the cDNA of protein-coding genes, we find 626,321 SNPs in the B strain that are not observed in J strain over a total of ~57.8*10^6 nucleotides covered, or 1.1% difference. SNPs common to B and J strains constitute only a modest fraction of B strain SNPs (<29%), but the majority (>70%) of J-line SNPs. A SNP shared between J-line and B-line is very likely a wild-type variant persisting in the J-line colony. A SNP unique to J-line may indicate either a false positive in our data or a reference inaccuracy.

Characterization of diversity in a first generation hybrid

The genomic diversity data from an F1-hybrid cross of the J and B strains is a further resource. The number of SNPs and INDELs observed in the JxB F1 generation is more similar to that observed in J strain than in the B strain, particularly at higher quality thresholds. This is due to a combination of factors, specifically a qualitative and quantitative decrease in the representation of alternative variants in the F1 compared to the pure B strain, and a mapping bias favoring the alignment of reference reads. However, when taking into account only the coverage depth (and not the proportion of the alternative allele observed), the amount of observed F1 variation increased substantially (**Figure 2**). This indicates that F1 data can be used to validate and refine variant calls from the pure strains.

Further, it is possible to use the data to discover expression quantitative trait loci (eQTLs) by analyzing the relative expression of the two alleles at heterozygous variant calls.

Validation and technical noise estimation

Our approach to SNP calling has a number of inherent limitations due to the fact that it uses RNA rather than DNA. To understand the potential for inaccuracies inherent in the method, we used the data from an additional experiment in which one tadpole was cut into two parts, and the resulting samples underwent all steps of the protocol from RNA extraction to sequencing separately. We then assessed the validation rate of all SNPs covered with at least 10 reads in both samples. This minimal coverage restriction was necessary in order for us to reduce the influence of unrelated factors such as differences in gene expression level on the chance of confirmation When a variant was called in both replicates, >99.7% of genotype calls were identical showing that the bases were accurately called; When a variant was called in one sample, it was also called in the other 70% of the time. When we restricted the analysis to SNPS with quality 30 or higher (which was used for cutoff for statistics cited in the paper) in the sample where they were detected, the agreement jumped to 82%. The quality call itself corresponds to 97% chance of correct call, which means the discrepancy between the expected and actual confirmation rate is 15 %.

It is important to note however that the unconfirmed calls are not necessarily equivalent to a false positive rate. A SNP may remain undetected in one of the samples even at a minimal coverage of 10 reads due to a number of factors, including amplification bias, base quality differences, and tissue-specific differences in regulatory regions (eQTLs) which may introduce different allelic bias at heterozygous loci in different parts of the sample.

Missense mutations

Having acquired the SNP data, we asked how many of these SNPs cause a change in the protein sequence they encode, and what is the nature of these changes, i.e. are they neutral, or function-altering (a.k.a. damaging). For this analysis, we used SNPs covered at depth >30. At the protein level, of about 20M amino acid residues covered, we detect 29,008 missense changes, or approximately 0.15% difference. Assuming the average tryptic peptide length of 13 amino acids, we can estimate ~2% probability of missing a peptide in proteomic mass-spec spectra-to-peptide match because of a substitution. The degree of sequence difference between J strain and wild type *Xenopus* is especially relevant in proteomic experiments, since a single amino acid substitution will result in a peptide not being identified in a peptide-spectra match. Provided an estimate that 15-50% of all proteins are identified and quantified based on a single peptide, the SNP distinction will lead to a substantially reduced number of characterized proteins and reduced accuracy for proteins registered with more than a single peptide. An average peptide length for proteins detected in our previous *Xenopus* experiment is 13 amino acids (Peshkin et al., 2015; Wühr et al., 2014). We analyzed nsSNPs using our previously published server for predicting damaging missense mutations (Adzhubei et al., 2010). A total of 29,008 nsSNPs in 11,603 proteins were identified. These proteins correspond to 9,012 uniquely named genes, where we use gene symbols assigned by our published pipeline (Wühr et al., 2014). Homeologous genes located on the longer and shorter paired

chromosomes (Matsuda et al., 2015; Session et al., 2016) are respectively labeled with suffixes ".L" and ".S".

There is no bias between "L" and "S" chromosomes (respectively 10,858 vs 9,059 SNPs in 5,774 vs 5,566 genes), and 3,353 genes have nsSNPs in both "L" and "S" homeologs. The functional effect annotation is mainly based on a positional amino acid conservation score.

PolyPhen-2 category	Number of variants	% *X. laevis*	% *H. sapiens* (WHESS)
Benign	19307	66	40
possibly damaging	3124	11	20
probably damaging	1679	6	40
Unknown	4898	17	0

Table 1: The classification of variants in *X. laevis* coding sequence by PolyPhen-2 category as based on "HumVar" model, and their relative share as compared to *WHESS* (*H. sapiens*) dataset.

We next set out to determine how this compares to what is observed in humans. It is estimated that a typical individual differs from the reference genome sequence at approximately 10,000-12,000 synonymous and 10,000 non-synonymous sites (1000 Genomes Project Consortium et al., 2012), with 250 to 300 loss-of-function variants. All nsSNPs are categorized into "benign", "unknown", "possibly damaging" and "probably damaging" categories based on the PoyPhen-2 model (**Table 1**). The percentage of damaging mutations we observe is naturally much lower than in the whole *human* exome sequence space (Adzhubei et al., 2010) (*WHESS*). This is expected since our data represents only a tiny sample from the population, thus capturing primarily the most frequent putatively neutral or mildly damaging variants. However, we observe that there is a significant number of function-altering mutations in *X. laevis* potentially making it a useful model organism for developmental genetics.

Genome duplication and subfunctionalization

Xenopus laevis underwent speciation (~34 million years ago), followed by allotetraploidization approximately 18 million years ago. Subsequently, an estimated 40% of duplicate gene copies were lost (Session et al., 2016). One common question for mutation analysis is whether the genes with retained duplicates evolve at a different rate than singleton genes. It has been observed across species that genes that retain their duplicated copy after whole genome duplication are functionally more important and therefore subject to strong selection (Jordan et al., 2004). If functional importance did not play a role in duplicate copy loss, one may hypothesize that genes with duplicates are more robust to mutation due to the presence of an extra copy. To assess the strength of purifying selection in the two groups of genes, we used SNPs unique to the B strain which we took to have appeared at a time later than genome duplication. We compared genes with existing duplicates (paired) with apparent single allele genes (singletons) in two measures – the ratio of missense (non-synonymous) variants to synonymous variants (D_n/D_s) and the ratio of damaging variants to missense variants. Consistent with other species, we find the results suggest that paired genes appear to evolve at a slower average rate -- a mean D_n/D_s value of 0.87(paired) vs 1.01(singletons), when computed over genes with at least one synonymous variant (n_{paired} = 4484, $mean_{paired}$ = 0.87, std_{paired} = 0.89, $n_{singleton}$ =

4014, $mean_{singleton}$ = 1.01, $std_{singleton}$ = 1.04, upaired t-test p < 10e-05; χ^2 test p < 10e-05; Kolmogorov-Smirnov test for difference in respective distributions, p = 1.5e-09, see **Fig. 3**).

One concern is that genes without synonymous substitutions may be disproportionately distributed among paired and singleton genes, to control for this possibility, we calculated D_n/D_s for each set of sequences as two virtual supergenes (paired and singleton). The results of this calculation uphold the previously observed difference, although taking into consideration sequences of genes without synonymous substitutions brings the overall ratio down, as expected ($mean_{paired}$ = 0.68, CI_{low} = 0.66, CI_{high} = 0.70; $mean_{singleton}$ = 0.84, CI_{low} = 0.81, CI_{high} = 0.86, see **Table 2**). The dN/dS ratio (Yang and Bielawski, 2000), which is the D_n/D_s normalized by the number of potential synonymous and non-synonymous sites, can be interpreted as an indicator for selection pressure. It is lower than 1, suggesting that both groups of genes remain largely under purifying selection. Thus, we can conclude that duplicated genes are under higher purifying selection than singletons.

Type	Genes	Missense (benign + damaging)	Synon. mutations	D_n/D_s	Synon sites	Nonsyn sites	dN/dS	Damaging/ Missense	Damaging/ Synonymous
singleton	27,779	14,990	17,936	**0.84**	6,586,292	24,025,756	0.229	0.151	0.126
paired	17,320	14,018	20,719	**0.68**	6,501,803	23,642,778	0.186	0.181	0.123

Table 2: Variant types in singletons versus paired genes in *Xenopus laevis*. Only variants unique to the B strain are considered and annotated by PolyPhen-2. For this comparison, "paired" and "singleton" genes are considered as a unified supergene, which allows us to take into consideration genes without synonymous substitutions.

An additional question is whether some paired genes underwent subfunctionalization after the split. The process of subfunctionalization would be characterized by a larger than expected fraction of function-altering mutations. We therefore asked if there is a difference between the two gene groups in the fraction of function-altering (damaging) mutations observed in missense mutations. We found that damaging mutations constitute a higher fraction of the missense mutations in these genes as compared to singletons (0.151 vs 0.181, p < 10e-05, Chi-square test). Remarkably, the ratio of damaging to synonymous variants is close, therefore the elevated fraction of damaging among missense offsets the reduced fraction of missense among all.

One caveat is that this analysis is performed on a draft version of the *X. laevis* genome, it is therefore unlikely that all duplicated genes have been positively identified. Thus, the two juxtaposed categories – paired and singletons, are best thought of as a paired subset and a singleton-enriched subset. Even so, our results are relative and based on the comparison of the two classes, and therefore do not rely on perfect separation between the classes and can absorb categorization errors. We should thus expect the results to get more significant once the "singleton-enriched" category becomes better defined with the future releases of the genome assembly and gene models.

How sequence differences affect protein mass spectrometry resolution

To provide an additional and orthologous estimate of the divergence of the J and B strains, we analyzed the spectra from our previous proteomic mass spectrometry experiments using the most recent *Xenopus laevis* genome sequence as a reference dataset. Using the same search parameters as previously reported (Peshkin et al., 2015) but the gene models from the v9.1 genome, approximately 144K spectra were matched to approximately 45K unique peptides --- an overall improvement of 5% compared to using an older *X. laevis* genome JGI6.1 genome reference in our original publication. After creating an alternative reference corresponding to B strain variants of proteins, we found that 1.4% of peptides are missing or about 1 in 75 peptides. At the protein level this corresponds to detecting a total of 8,456 instead of 8,499 proteins or 0.5% of the proteins (of which 2,062 were single-peptide detections with 11 proteins only lost from these).

We noted that for a heterozygous nsSNP proteomic mass spectrometry measurement would show only fractional level of protein expression. This fraction is in turn unknown since there are genes with genetic allelic bias or possibly epigenetically-driven mono-allelic expression (Nag et al., 2013). However, we estimate that none of the aforementioned loss of resolution seems significant at the level of nsSNP. Finally, our numbers are clearly an underestimate since there are variants that we did not detect within the proteins we measured in that particular experiment; in addition, proteins present in early embryonic stages are likely to be under more stringent purifying selection compared to later stages.

Discussion

In summary, we have presented a resource for coding variation in *Xenopus laevis* based on the newly released genome sequence. One caveat is that we study polymorphism indirectly, using transcripts, which means that both noise and systematic bias could affect our observations. While we have provided an upper bound of the estimate of technical noise with an experiment involving technical replicate, there are biological sources of noise we cannot fully address. For example, deamination could systematically produce A/G substitutions in substantial fraction of the transcripts – a phenomenon known as “RNA editing”. A recent study found that as much as 60% of the mRNA in squid brain is edited (Alon et al., 2015). Furthermore, expression biases of genetic or epigenetic nature can affect our ability to detect coding polymorphisms.

In addition, it is possible that some of the newly described variants may have resulted from mRNAs transcribed from the currently missing sections of the genome sequence, which bears resemblance to the existing sequence. Note that cases like this describe actual variants present in mRNA although not necessarily in DNA.

With these caveats, we have shown that there remains residual coding variation in the inbred *Xenopus laevis* J-strain line, likely reflecting in part the differences between existing J-line colonies and the colony used for the genome project.

Unsurprisingly, the genetic diversity in an occasionally outbred strain is much larger that that of the inbred strain. The greater diversity of occasionally outbred strains, such as the B strain constitutes a technical challenge and may need to be taken into account in bioinformatics analysis, e.g. when mapping parameters are set. Specifically, the number of allowed mismatches may be tuned to reflect the expected number of variants in the read given the read length and should be increased if an

occasionally outbred strain (such as the B strain) is used. However, note that increasing the rate of allowed mismatches may present greater difficulty in distinguishing homeolog expression, and the overall mapping rate may be affected since more ambiguous mappings across homeologs are likely to result. This trade-off should be considered in the context of the specific sequencing application. Further, using a SNP-masked reference should be considered when tadpoles from two different strains (e.g. J and B strains) are used in the same experiment. In addition, homozygous variants in J strain can be utilized to improve the quality of the genome sequence.

Of particular interest in this regard are functionally relevant, and more broadly amino-acid changing mutations. The latter have a detrimental effect on efforts to measure the protein abundance. When the use of an inbred strain cannot mitigate the problem, a correction of the reference sequence should be considered based on the known genetic profile of the sample. Another promising direction is functional characterization of non-synonymous variants using models and computational methods with a purpose of identifying variants which are high-confidence mutagenic according to the model, yet clearly present and harbored in the population. Resolving such discordance would lead to improvement of computational functional prediction algorithms, and potentially discovery of compensatory epistasis. Epistatic interactions could be further studied by gene-editing and mutagenesis approaches starting from two distinct strains with characterized differences.

While the timescales over which natural selection shapes the genome of the species are orders of magnitude longer than the timescale of strain generation, we have also shown that strain variation can be used to make claims about evolutionary processes on longer timescales, because samples from that strain reflect the genetic diversity within the species. Analysis of variants exclusive to the occasionally outbred strain allowed us to address the question of subfunctionalization in gene duplications. The results of the comparison reveal an interesting paradox. While paired genes appear to evolve at a slower rate, function-altering (damaging) mutations constitute a higher fraction of the missense mutations in these genes as compared to singletons. Indeed, the first observation is described as an apparently universal phenomenon affecting gene duplication throughout the animal kingdom (Jordan et al., 2004), and is attributed to the fact that gene duplications generally affect genes with higher functional significance. Paired genes are therefore subjected to higher levels of purifying selection weeding out non-beneficial mutations. On this background, the higher prevalence of function-altering mutations in paired genes among non-synonymous mutations might be cautiously interpreted as evidence for subfunctionalization: variants leading to different function are retained at higher rates in paired genes due to the duplicate alleviating the pressure against acquiring new use. This may be in agreement with a dynamic model of gene duplication, in which the novel copy of a duplicated gene is initially subjected to rapid evolution, followed by a return to purifying selection (Pegueroles et al., 2013). Further analysis on a gene-by-gene basis may shed more light on the viability of this interpretation.

In conclusion, this study shows that genetic analysis of *Xenopus laevis* enabled by the publication of its genome can lead to solutions of current technical challenges and to new insights into the functional significance of variation. The long life span of *Xenopus* (over 10 years) and the ease of large-scale developmental screens in this organism provide a valuable tool for assessing the functional relevance of specific gene variants. The list of variants uncovered in this study is an important step towards building up valuable resources for the community.

Supplementary Materials

The raw sequence reads data and the variant call files were deposited to GEO under accession number **GSE74470**.

Acknowledgements

L.P. was supported by the NIH grant R01HD073104, also L.P., A.N. and V.S. were supported by R21HD81675, M.H. and E.P. by P40 OD010997. We thank MBL/NXR for their support of this project.

Figures

Figure 1: Three groups are considered independently for discovery and validation of genomic variants in *X. laevis:* a straight J, a straight B and an F1 cross.

Figure 2: Assessment of diversity in *Xenopus laevis* by RNA-seq of clutches of inbred (J), and occasionally outbred (B) lines, and F1 crosses. **A-B**. Number of SNPs (A) and INDELs (B) as a function of *bcftools* call quality. Call quality corresponds to $-10\log_{10}$ probability that call is wrong). **C**. Ratio of transitions to transversions as a function of call quality (see A-B). **D**. Number of calls as a function of coverage depth.

Figure 3: A comparison between D_n/D_s distribution for paired and singleton genes, illustrating a shift between two groups of genes. X-axis: log2 D_n/D_s

References

1000 Genomes Project Consortium, Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., McVean, G.A., 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 491, 56–65. doi:10.1038/nature11632

Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., Sunyaev, S.R., 2010. A method and server for predicting damaging missense mutations. Nat. Methods 7, 248–249. doi:10.1038/nmeth0410-248

Alon, S., Garrett, S.C., Levanon, E.Y., Olson, S., Graveley, B.R., Rosenthal, J.J.C., Eisenberg, E., 2015. The majority of transcripts in the squid nervous system are extensively recoded by A-to-I RNA editing. eLife 4. doi:10.7554/eLife.05198

Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., Ruden, D.M., 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 6, 80–92. doi:10.4161/fly.19695

Cui, X., Affourtit, J., Shockley, K.R., Woo, Y., Churchill, G.A., 2006. Inheritance Patterns of Transcript Levels in F1 Hybrid Mice. Genetics 174, 627–637. doi:10.1534/genetics.106.060251

Davis, E.E., Frangakis, S., Katsanis, N., 2014. Interpreting human genetic variation with in vivo zebrafish assays. Biochim. Biophys. Acta 1842, 1960–1970. doi:10.1016/j.bbadis.2014.05.024

del Viso, F., Bhattacharya, D., Kong, Y., Gilchrist, M.J., Khokha, M.K., 2012. Exon capture and bulk segregant analysis: rapid discovery of causative mutations using high-throughput sequencing. BMC Genomics 13, 649. doi:10.1186/1471-2164-13-649

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., McKenna, A., Fennell, T.J., Kernytsky, A.M., Sivachenko, A.Y., Cibulskis, K., Gabriel, S.B., Altshuler, D., Daly, M.J., 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43, 491–498. doi:10.1038/ng.806

Gantress, J., Maniero, G.D., Cohen, N., Robert, J., 2003. Development and characterization of a model system to study amphibian immune responses to iridoviruses. Virology 311, 254–262.

Gurdon, J.B., 2013. The egg and the nucleus: a battle for supremacy. Dev. Camb. Engl. 140, 2449–2456. doi:10.1242/dev.097170

Hellsten, U., Khokha, M.K., Grammer, T.C., Harland, R.M., Richardson, P., Rokhsar, D.S., 2007. Accelerated gene evolution and subfunctionalization in the pseudotetraploid frog Xenopus laevis. BMC Biol. 5, 31. doi:10.1186/1741-7007-5-31

Hoperskaya, O.A., 1975. The development of animals homozygous for a mutation causing periodic albinism (ap) in Xenopus laevis. J. Embryol. Exp. Morphol. 34, 253–264.

Hunt, T., 2002. Nobel Lecture. Protein synthesis, proteolysis, and cell cycle transitions. Biosci. Rep. 22, 465–486.

Izutsu, Y., Maéno, M., 2005. Analyses of immune responses to ontogeny-specific antigens using an inbred strain of Xenopus laevis (J strain). Methods Mol. Med. 105, 149–158.

Jordan, I.K., Wolf, Y.I., Koonin, E.V., 2004. Duplicated genes evolve slower than singletons despite the initial rate increase. BMC Evol. Biol. 4, 22. doi:10.1186/1471-2148-4-22

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 1000 Genome Project Data Processing Subgroup, 2009. The Sequence Alignment/Map format and SAMtools. Bioinforma. Oxf. Engl. 25, 2078–2079. doi:10.1093/bioinformatics/btp352

Matsuda, Y., Uno, Y., Kondo, M., Gilchrist, M.J., Zorn, A.M., Rokhsar, D.S., Schmid, M., Taira, M., 2015. A New Nomenclature of Xenopus laevis Chromosomes Based on the Phylogenetic Relationship to Silurana/Xenopus tropicalis. Cytogenet. Genome Res. 145, 187–191. doi:10.1159/000381292

Nag, A., Savova, V., Fung, H.-L., Miron, A., Yuan, G.-C., Zhang, K., Gimelbrant, A.A., 2013. Chromatin signature of widespread monoallelic expression. eLife 2, e01256. doi:10.7554/eLife.01256

Nieuwkoop, P.D., Faber, J., 1994. Normal table of Xenopus laevis (Daudin): a systematical and chronological survey of the development from the fertilized egg till the end of metamorphosis. Garland Pub, New York.

Pearl, E.J., Jarikji, Z., Horb, M.E., 2011. Functional analysis of Rfx6 and mutant variants associated with neonatal diabetes. Dev. Biol. 351, 135–145. doi:10.1016/j.ydbio.2010.12.043

Pegueroles, C., Laurie, S., Albà, M.M., 2013. Accelerated evolution after gene duplication: a time-dependent process affecting just one copy. Mol. Biol. Evol. 30, 1830–1842. doi:10.1093/molbev/mst083

Peshkin, L., Wühr, M., Pearl, E., Haas, W., Freeman, R.M., Gerhart, J.C., Klein, A.M., Horb, M., Gygi, S.P., Kirschner, M.W., 2015. On the Relationship of Protein and mRNA Dynamics in Vertebrate Embryonic Development. Dev. Cell 35, 383–394. doi:10.1016/j.devcel.2015.10.010

Session, A.M., Uno, Y., Chapman, J.A., Toyoda, A., Takahashi, S., Fukui, A., Hikosaka, A., Suzuki, A., Kondo, M., van Heeringen, S.J., Quigley, I., Heinz, S., Ogino, H., Ochi, H., Hellsten, U., Lyons, J.B., Simakov, O., Putnam, N., Stites, J., Kuroki, Y., Tanaka, T., Michiue, T., Watanabe, M., Bogdanovic, O., Lister, R., Georgiou, G., Paranjpe, S.S., van Kruijsbergen, I., Shu, S., Carlson, J., Kinoshita, T., Ohta, Y., Mawaribuchi, S., Jenkins, J., Grimwood, J., Schmutz, J., Mitros, T., Mozaffari, S.V., Suzuki, Y., Haramoto, Y., Yamamoto, T.S., Takagi, C., Heald, R., Miller, K., Haudenschild, C., Kitzman, J., Nakayama, T., Izutsu, Y., Robert, J., Fortriede, J., Burns, K., Lotay, V., Karimi, K., Yasuoka, Y., Dichmann, D.S., Flajnik, M.F., Houston, D.W., Shendure, J., DuPasquier, L., Vize, P.D., Zorn, A.M., Ito, M., Marcotte, E.M., Wallingford, J.B., Ito, Y., Asashima, M., Ueno, N., Matsuda, Y., Veenstra, G.J.C., Fujiyama, A., Harland, R.M., Taira, M., Rokhsar, D.S., 2016. Genome evolution in the allotetraploid frog Xenopus laevis. Nature 538, 336–343. doi:10.1038/nature19840

Tochinai, S., Katagiri, C., 1975. COMPLETE ABROGATION OF IMMUNE RESPONSE TO SKIN ALLOGRAFTS AND RABBIT ERYTHROCYTES IN THE EARLY THYMECTOMIZED XENOPUS. Dev. Growth Differ. 17, 383–394. doi:10.1111/j.1440-169X.1975.00383.x

Trapnell, C., Pachter, L., Salzberg, S.L., 2009. TopHat: discovering splice junctions with RNA-Seq. Bioinforma. Oxf. Engl. 25, 1105–1111. doi:10.1093/bioinformatics/btp120

Wang, J., Raskin, L., Samuels, D.C., Shyr, Y., Guo, Y., 2015a. Genome measures used for quality control are dependent on gene function and ancestry. Bioinforma. Oxf. Engl. 31, 318–323. doi:10.1093/bioinformatics/btu668

Wang, J., Samuels, D.C., Shyr, Y., Guo, Y., 2015b. Population structure analysis on 2504 individuals across 26 ancestries using bioinformatics approaches. BMC Bioinformatics 16, P19. doi:10.1186/1471-2105-16-S15-P19

Wühr, M., Freeman, R.M., Presler, M., Horb, M.E., Peshkin, L., Gygi, S.P., Kirschner, M.W., 2014. Deep proteomics of the Xenopus laevis egg using an mRNA-derived reference database. Curr. Biol. CB 24, 1467–1475. doi:10.1016/j.cub.2014.05.044

Yanai, I., Peshkin, L., Jorgensen, P., Kirschner, M.W., 2011. Mapping gene expression in two Xenopus species: evolutionary constraints and developmental flexibility. Dev. Cell 20, 483–496. doi:10.1016/j.devcel.2011.03.015

Yang, Bielawski, 2000. Statistical methods for detecting molecular adaptation. Trends Ecol. Evol. 15, 496–503.

Yang, Z., 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol. Biol. Evol. 24, 1586–1591. doi:10.1093/molbev/msm088

Supplementary Figures and Tables

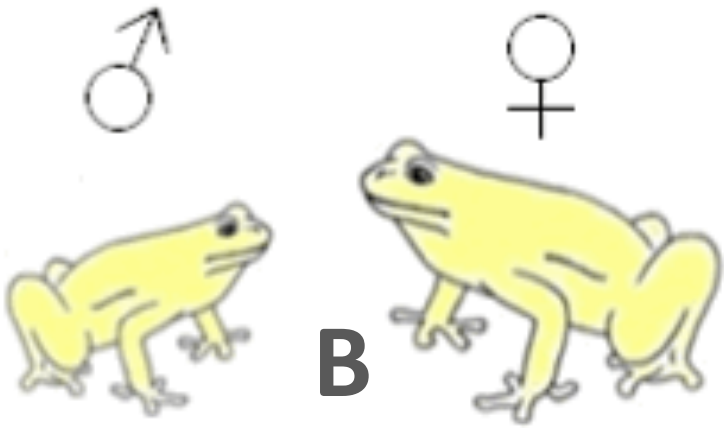
Table S1: Alignment statistics for the RNA-Seq read obtained respectively from J strain, B strain and F1-crosses.

	Paired reads	Aligned Concordantly	Multiple Alignment	Discordant	Concordant alignment rate	Unique concordant alignments
B line	32044487	27468914	1798611	1018479	83%	77%
J line	36403413	32115006	2127964	1224116	85%	79%
F1 crosses	76787153	66608073	4505595	2056413	84%	78%

Figure S1: Number of heterozygous (blue) and homozygous (red) mutations in J strain as a function of call quality (see Figure 2). Only SNPs covered with 10 or more reads were considered.

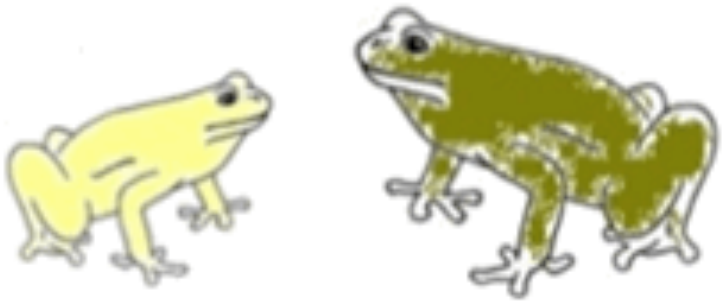
Glossary

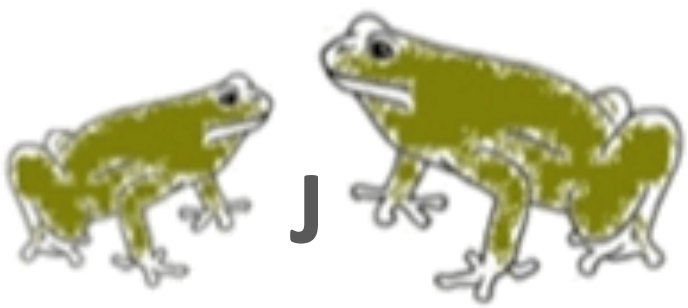
B strain	occasionally outbred albino strain
J strain	genomic stain
F1	first-generation hybrid
Transitions	interchanges of two-ring purines (A <-> G), or of one-ring pyrimidines (C <-> T)
Transversions	exchange of one-ring and two-ring structures: purine for pyrimidine
ts/tv	transition/transversion ratio. Transitions are expected to occur twice as frequently as transversions. In protein coding regions, this ratio is typically higher, often a little above 3. The higher ratio occurs because, especially when they occur in the third base of a codon, transversions are much more likely to change the encoded amino acid.
D_n/D_s	the ratio of missense (non-synonymous) variants to synonymous variants
dN/dS	D_n/D_s normalized to the number of potential synonymous and non-synonymous sites



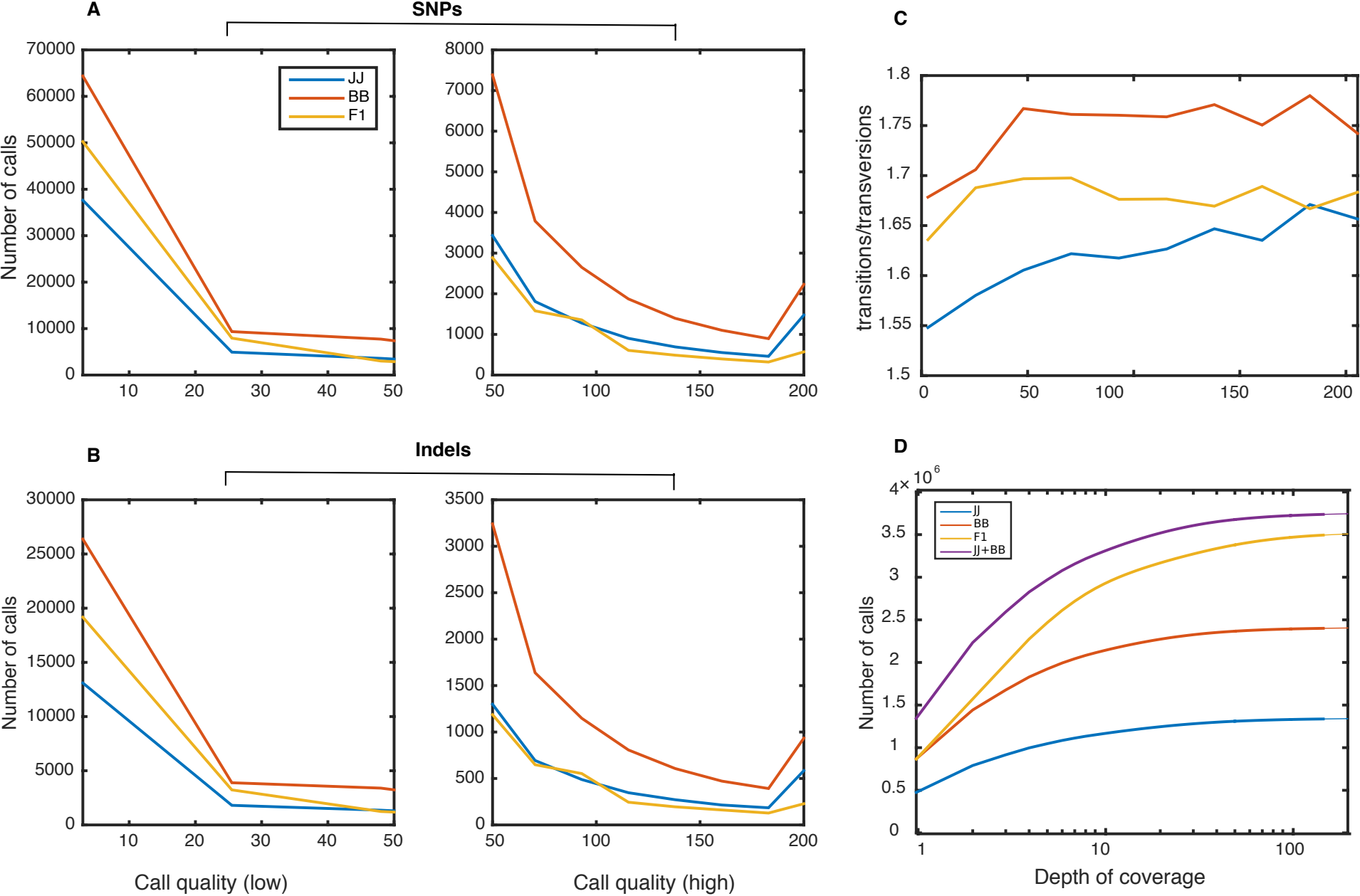
B

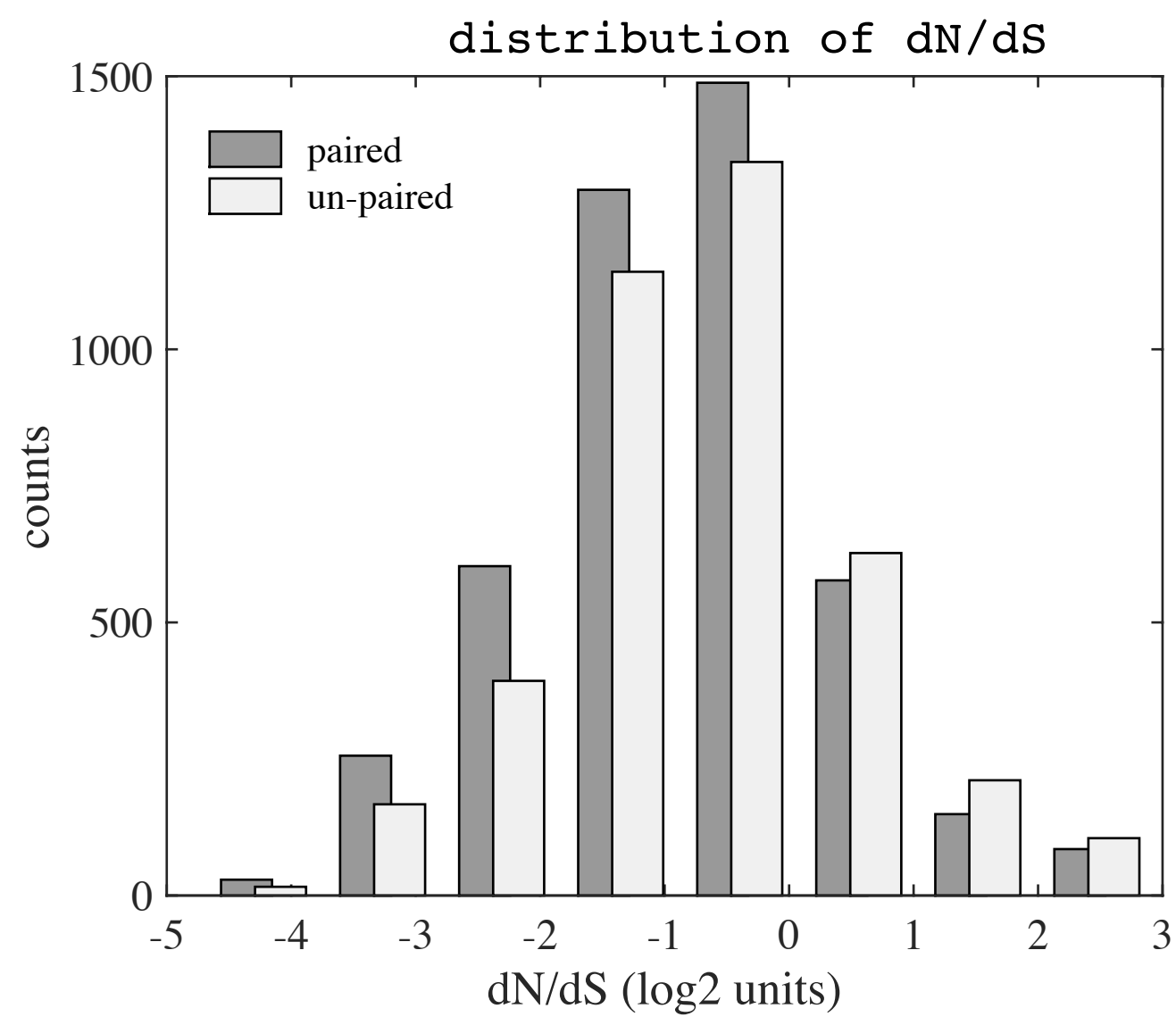
F1

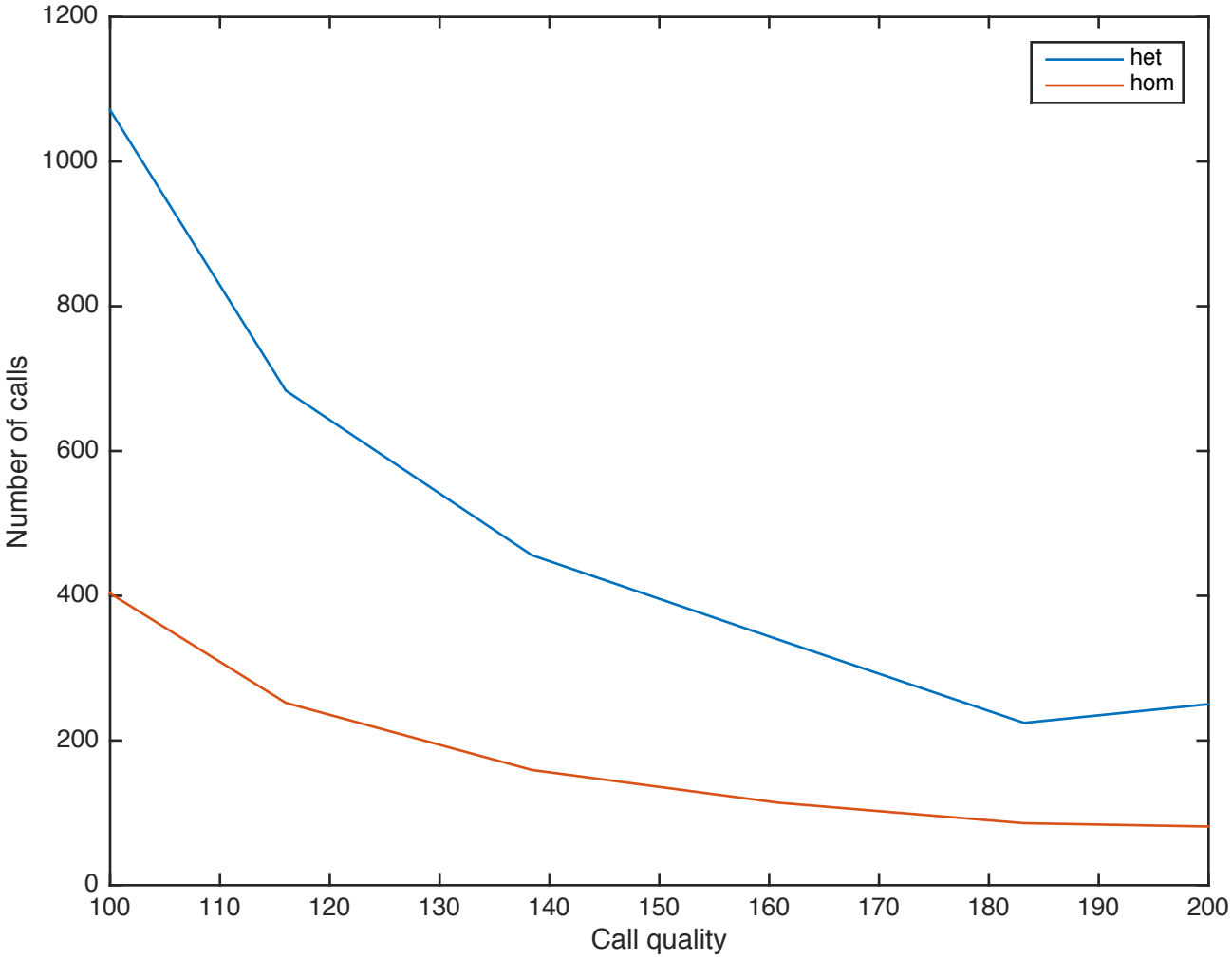




J


A
SNPs
Number of calls
JJ
BB
F1
B
Indels
Call quality (low)
Call quality (high)
C
transitions/transversions
D
Number of calls
JJ
BB
F1
JJ+BB
Depth of coverage


distribution of dN/dS
paired
un-paired
counts
1500
1000
500
0
-5
-4
-3
-2
-1
0
1
2
3
dN/dS (log2 units)


het
hom
Number of calls
Call quality
0
200
400
600
800
1000
1200
100
110
120
130
140
150
160
170
180
190
200

Highlights:

- A catalogue of coding variants in two strains of Xenopus laevis including genomic.
- A functional annotation of identified coding mutations is provided.
- The ratio of non-synonymous to synonymous mutations suggests subfunctionalization.