Variance Analysis of Homeologous Genes in the Developing Allotetraploid Species *Xenopus leavis*

LeAnn Lo, Ronald Cutler, Mark Pownall, Chen Dong, Margaret Saha

Biology Department, College of William and Mary, Williamsburg, VA

I. Introduction

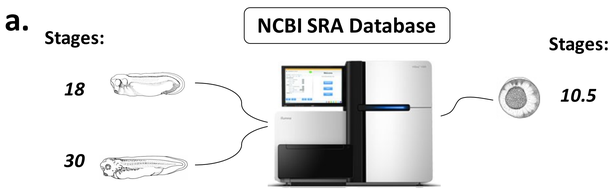
The allotetraploid Xenopus laevis genome is divided into 2 diploid subgenomes hypothesized to come from an ancient polyploidzation event, following with each species having denoted L and S (Long and Short) chromosomes, where 56% of all genes were comprised of as homeologous pairs [3]. Homeologs have been observed to have high retention rates in major signaling pathways as well as in development genes [3]. High variability in homeologs has been observed in many of these signaling pathways such as Notch [1]. Using pooled RNA-sequencing data from X. laevis embryos, we examined the variation patterns of all known homeologs against non-homeologous genes throughout the species major neural developmental stages. We discovered that homeologs are more variant than non-homeologs throughout development and that variant homeolog pairs retain this pattern. Earlier development stages also have higher variability compared to later stages in both homeologs and non-homeologs.

II. Methods

RNA-Seq data was gathered from five different experiments (primarily from the SRA database), targeting their control replicates. Data that was used was prepared with the same techniques and equipment used in Saha Lab, specifically with polyA selection, paired-end layout, and with the Illumina HiSeq 2000. Stages 10, 18, and 30 were chosen as major neural developmental stages and for having the highest availability of data, they are also the stages of interest within the research lab. Stage 10 is the beginning of the gastrula stage where the embryo has begun differentiation and begins the development of the anterior-posterior axis [3]. Stage 18 is in the middle of the neurula stage at which neurulation occurs, the folding of the neural plate into the neural tube. Stage 30 is the late tailbud stage where neurulation is complete and tail formation begins. All raw data retrieved from the SRA database as well as local lab data was processed through the pipeline depicted in 1b. Read data was extracted from the database, then aligned to a reference genome and quantified into read counts using RNA-Seq programs. All read counts were clustered to determine unwanted variation, later removed through batch effect removal and filtered to have a mean of 5 counts for each gene. Differential expression between homeologs and non-homeologs was observed by comparing their expression in the various stages and their variation (standard deviation) throughout the multiple replicates. Certain homeologous genes of interest with high and low variation between replicates were then observed.

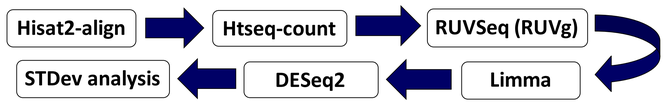
1. RNA-Seq Methods and Pooled Data

a.



	Sessions	Ding	AP Rotations	Peshkin
Stage 10.5	2	3	0	0
Stage 18	0	0	5	1
Stage 30	0	0	5	1

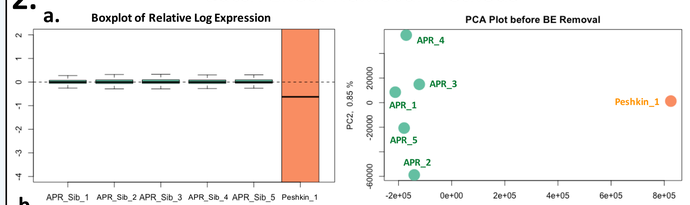
b.



Schematic of data (a) Appearances of embryos at given stage, table shows number of replicates retrieved from certain data sets. Sessions data from Sessions et al. (2016), Ding data from Ding Y et al. (2017), AP Rotations from Saha Lab, Peshkin data from Peshkin L et al. (2015) (b) Analysis pipeline where reads are aligned to reference genome, count reads per gene, removing unwanted batch effects with empirical gene controls, retrieve batch removed counts given log-expression values, retrieve normalized counts, finding variation and plotting using R.

2. Batch Effect Removal Methods

a.



b.

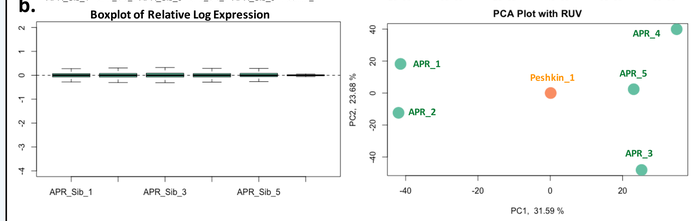
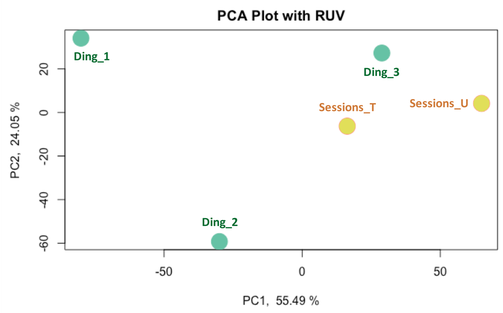


Figure 2a) Shows the initial RLE and PCA plot before any sort of normalization or batch effect removal. The RLE plot shows the log-ratio of read count to median read count, or how similar the overall gene counts are on a long scale. As shown in 2a, Peshkin, a new data set, has a widely different range of counts compared to APR, this is corrected by BE removal as these differences can be attributed to the samples coming from different sources and have minor differences in preparation. Regardless of these differences, they are still similar and can be corrected to maintain all genetic expression patterns and have the external variation removed. These two sets of data can then be compared after batch effect removal (Figure 2b).

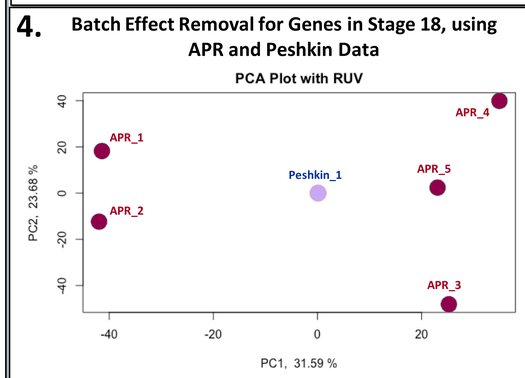
III. Results

3. Batch Effect Removal for Genes in Stage 10.5, using APR and Peshkin Data



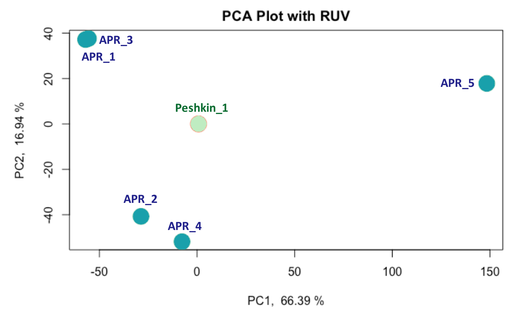
Sessions and Ding are two different sets of data with different library preparations, Ding is single-ended compared to the rest of data batches. However, this does not impact the batch effect removal, its library preparation is consider part of its "batch" that is not necessary for the overall gene expression. Clustering and batch effect modification returns meaningful data without confounds that distract from the biological significance.

4. Batch Effect Removal for Genes in Stage 18, using APR and Peshkin Data



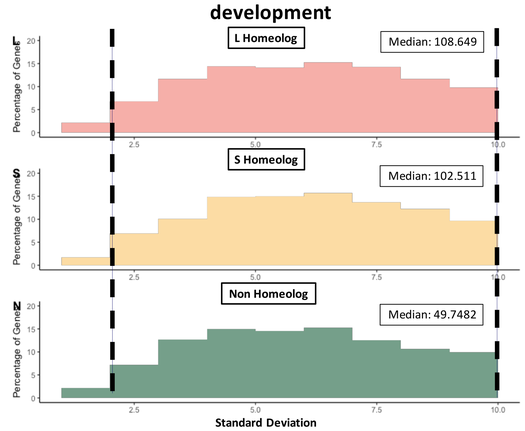
When clustering Peshkin to APR, APR does not consistently cluster together due to being from different embryos and from general genetic variation. The goal of batch effect removal is to remove external variation that is not necessary to model and observe genetic expression from similar embryos. However, it is important to retain group differences and preserve biological significance when modifying count data.

Batch Effect Removal for Genes in Stage 30, using APR and Peshkin Data



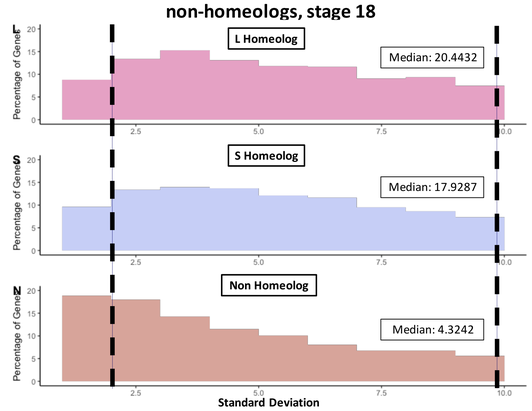
As there was only one replicate of Peshkin, it clustered towards APR which were already closely related due to being from the same experimental batch. Batch effect removal is necessary to prevent false positives when discovering differential expression and doing any gene comparisons. This is demonstrated when running DESeq, a differential gene analysis program, before any correction to provide RUVg with normalized counts, it counts nearly half of the genes in all of the samples to be differentially expressed, a very unlikely outcome.

Homeologs have the highest variance in earlier stages of development



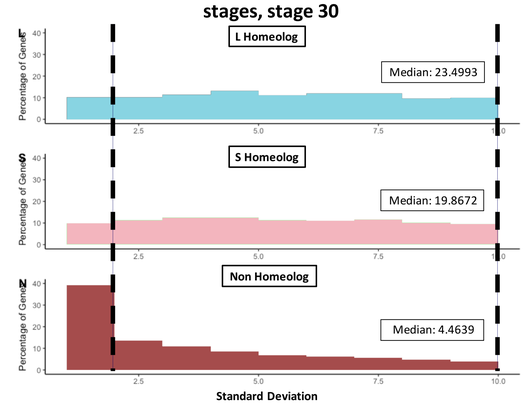
Ding and Sessions data was clustered together and their replicates' variance was calculated. The amount of variance that the homeologs have is significantly higher than any other time stage, this suggests that homeologs are more variant in earlier development stages before they become consistent in their genetic expression. Non-homeologs are less variant but closer to the homeologs in variance than in any other stage, this could be due to the early development stage as well.

5. Homeologs show consistent greater variance compared to non-homeologs, stage 18



APR and Peshkin data was clustered together and their replicates' variance was calculated. Non-homeologs have nearly twice as much consistent genes than homeologs with 20% having a standard deviation of 0. Stage 18 demonstrates a consistent higher amount of variation within homeologs compared to non-homeologs.

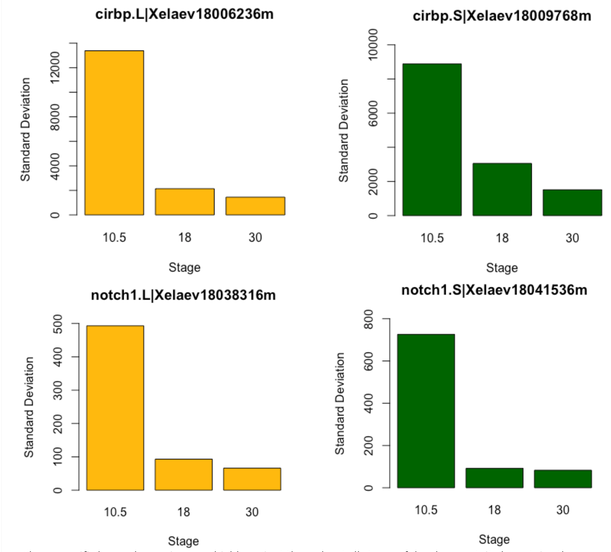
6. Homeologs retain variance throughout late development stages, stage 30



APR and Peshkin data was clustered together and their replicates' variance was calculated. The median shows that there is generally higher variance in the two homeologs (L/S) and significantly less in non-homeologs. Non-homeologs also have a higher percentage of genes with 0 standard deviation, or very consistent genes (first bar of histogram).

III. Results

7. Specific Variant Homeologous Genes



These specific homeologs pairs were highly variant throughout all stages of development, it also retains the same expression pattern of decreasing over time. It has been observed in Sessions et al. that the L homeolog has higher gene expression levels compared to the S homeolog. This is also represented in variance as more counts or reads are expressed in a gene, the more likely it is to be variant. Many of the homeologs with high variance have very high counts even after filtering all genes with low counts.

IV. Conclusions

1) Homeologs have consistently higher variance compared to non-homeologs in all of the given time frames amongst several different replicates of *X. laevis* embryos.
2) In earlier developmental stages, homeologs exhibit more variance than in any later stage, non-homeologs are also more variant in earlier stages. However, homeologs remain the most variant of the two.
3) In later developmental stages, non-homeologs begin to significantly stabilize in variance with nearly 40% having little to none variance (0 SD) at stage 30 compared to 5-10% of little variance in earlier stages.
4) Matched homeolog pairs (L and S) conserve the pattern of high variability in earlier development stages which steadily decreases as development continues.
5) The L homeolog is consistently more variable than the S homeolog in all stages, this is due to the L or long homeolog typically having more expression within the pair.
6) Some of the highly variable homeologs are also differentially expressed in other *X. laevis* experiments with early development perturbations such as anterior and posterior rotations (cirbp).

V. Future Direction

1) Gene enrichment analysis to determine function and common motifs in highly variant genes
2) Comparing homeolog isoforms and novel isoform discovery
3) Observing homeologs variance within *X. borealis* (tetraploid) RNA-Seq data, another species from the same subgenus as X. Laevis, as well as *X. andrei*, an octoploid species
4) 5KB Upstream Alignments and Comparisons of variant genes, as well as from AP axis perturbed areas, a known highly variant area or Notch, a major cell signaling pathway.
5) Modifying batch effect removal to account for additional batches by having more than two batches per stage, this will introduce more data in the current stages and have more available stages

References

1. Michiue, T. Yamamoto, Y. Yasuoka, T. Goto, T. Ikeda, K. Nagura, T. Nakayama, M. Taira, T. Kinoshita. High variability of expression profiles of homeologous genes for Wnt, Hh, Notch, and Hippo signaling pathways in Xenopus laevis *Dev. Biol.* (2017)
2. Risso D, Ngai J, Speed T and Dudoit S (2014). "Normalization of RNA-seq data using factor analysis of control genes or samples." *Nature Biotechnology*, 32(9), pp. 896–902.
3. Session, A., Yoshinobu Uno, Taejoon Kwon, et al. (2016). Genome evolution in the allotetraploid frog Xenopus laevis. *NATURE*, *538*, 26
4. Watanabe, M., Yasuoka, Y.,(2017). Conservatism and variability of gene expression profiles among homeologous transcription factors in Xenopus laevis. *Developmental Biology, 426*(2), 301-324.

Funding

1) NSF Grant 1257895 to MSS 2)NIH 1R15HD077624-01 to MSS 3) HHMI Undergraduate Science Education Program