**Supplemental Materials Accompanying**

**Gallant et al. “The Genome and Adult Somatic Transcriptome of the Mormyrid Electric Fish *Paramormyrops kingsleyae”***

*This Document Includes:*

**Supplemental Figures:**

Fig. S1 AED Score Distribution

Fig. S2: High Stringency Heatmap

Fig. S3: Low Stringency Heatmap

**Supplemental Methods:**

*Sample Preparation*

*Genome Sequencing and Assembly*

*Genome Annotation*

*Functional Annotation*

*Sample Preparation*

*Transcriptome Assembly*

*Transcriptome Quantification*

*Analysis of gene-expression profiles*

*Identification of housekeeping genes*

*Analysis of Differential Expression*

**Supplemental Analyses:**

*Examination of the Head and Flank Skin Transcriptomes*

**Supplemental References**

***Description of Supplemental Files***

File S1: List of Short Read Libraries (XLS)

File S2: List of Differentially Expressed Genes Across Replicates (XLS)

File S3: List of Housekeeping Genes (XLS)

File S4: Identities of SM\_EO DE Genes (XLS)

File S5: GO-Term (XLS)

***Supplemental Figures***

Figure S1: Distribution of annotation Edit Distance (AED) scores for all annotated genes from MAKER standard annotation set for PKINGS\_0.1 assembly.

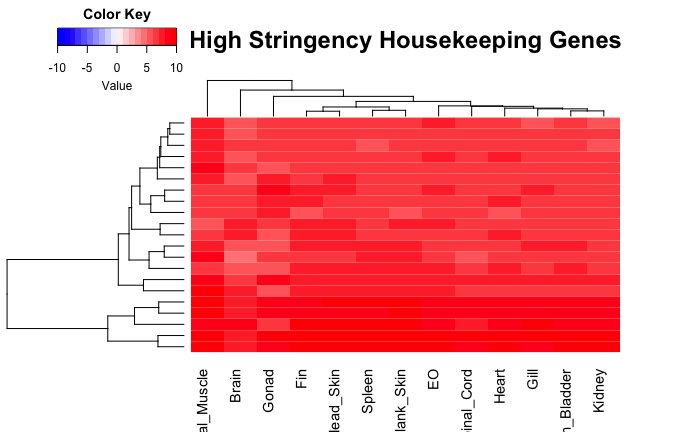


Figure S2: Log2-scaled heatmap of average expression values of 21 housekeeping genes identified using ‘high stringency’ criteria.

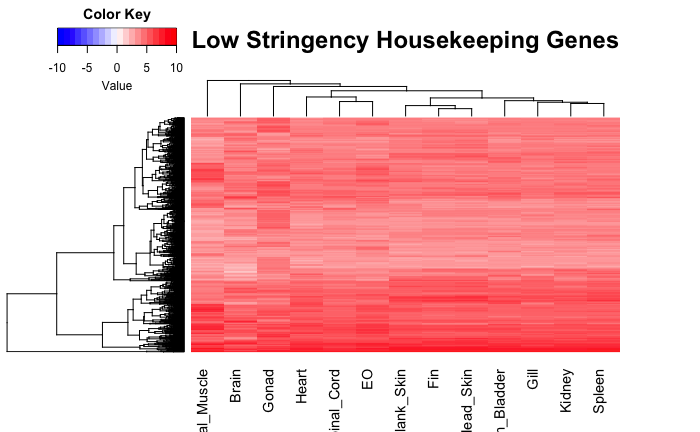


Figure S3: Log2-scaled heatmap of average expression values of 1134 housekeeping genes identified using ‘low stringency’ criteria.

***Supplemental Methods***

**Genome Sequencing, Assembly and Annotation**

*Sample Preparation*

For Illumina Genome sequencing, genomic DNA (gDNA) was isolated from fin clips made from a juvenile *Paramormyrops kingsleyae* collected in 2009 from Bambomo Creek (PKING1; EOD was P0-absent; see Gallant et al. 2011; Gallant et al. *in prep.*)using a Qiagen DNAeasy kit. Extracted gDNA was verified for concentration and quality using a Nanodrop and Qubit prior to submission for sequencing.

*Genome Sequencing and Assembly*

Based on previous estimates of mormyrid genome sizes (Hinegardener and Rosen 1972), we estimated the size of *P. kingsleyae* genome to be about 1 Gb, and thus utilized this estimate to generate coverage models for sequencing. Our collection of DNA libraries built for sequencing on the Illumina HiSeq2000 consisted of short-insert (200 bp) to allow for read overlap, and mate-pair libraries (insert sizes: 3, 8, and 10 kbp) all reads were 100 bp in length. The total resulting sequence coverage was approximately 90x: 40x of overlapping pairs (200bp), 40x of 3 kb, and 10x 8–10 kb.

The combined sequence reads were assembled with the ALLPATHS-LG software (Gnerre et al. 2011) using default parameter settings**.** Remaining small assembly gaps (<1 kb) within scaffolds were closed where possible by remapping overlapping read-pairs with a scaled modification of the ICORN2 software (<http://icorn.sourceforge.net)>. All contaminating contigs were identified by previously defined NCBI filters (alignments to non-fish species at the highest BLAST score obtained), and all contigs < 200bp were removed prior to final assembly submission.

*Genome Annotation*

To annotate the *P. kingsleyae* genome for coding and non-coding transcripts, MAKER version 2.31.9 (Cantarel, et al. 2008) was run on the *P. kingsleyae* genome assembly using the following input data: (1) **A custom repeat database**, constructed using RepeatModler (v.1.08) using the PKINGS\_0.1 genome assembly as input, with default parameters. The results of this database was merged with Repbase entries for additional fish-specific repeats. Repeat regions and additional areas of the genome consisting of low complexity were soft-masked by MAKER from seeding evidence in these regions. (2) **Protein Sources:** A custom protein database was constructed using the NCBI proteomes of *H. sapiens*, *X. tropicalis* and *D. rerio* and (3) a **transcriptome assembly**: PASA-merged genome-guided and *de-novo* assemblies of the P. kingsleyae transcriptomes, constructed as described below.

In the first stage of annotation, MAKER was run with EST2GENOME and PROTEIN2GENOME modes enabled, drawing on protein and transcriptome evidence to produce a rough approximation of gene models to train *ab initio* gene prediction algorithms.

The predicted genes resulting from this were then filtered for length and quality, and these genes were provided to the SNAP *ab initio* gene predictor for training to generate a hidden Markov model (HMM). In the second stage of annotation, MAKER was run with EST2GENOME and PROTEIN2GENOME modes disabled, now utilizing the SNAP HMM resulting from stage 1 to predict genes. The predicted genes resulting from this were then filtered for length and quality, and again provided to SNAP ab intio intiogene predictor to refine the HMM. In the third stage, MAKER was again run with EST2GENOME and PROTEIN2GENOME modes disabled, now utilizing the SNAP HMM resulting from stage 2 to predict genes. The predicted genes resulting from this were then filtered for length and quality. The resulting genes were then used to train a second *ab initio* gene predictor, AUGUSTUS. Genes were randomly split into two groups, training (n=150 genes) and test (n=150 genes) to validate the resulting HMM. Following completion of the AUGUSTUS training, sensitivity and specificity of the HMM were assessed (0.83 and 0.515 respectively). In the last stage, MAKER was again run with both the Stage 3 SNAP HMM and the Stage 3 AUGUSTUS HMM, while allowing for alternative splicing to produce a final set of predicted genes (‘maker max’ dataset).

*Functional Annotation*

Predicted transcripts were automatically translated to proteins by MAKER. Each was searched for IPR domains using InterProscan (release), and aligned using BLASTX to the manually curated set of proteins in the UniProtKB/Swissprot (release X). Putative functional names/symbols were assigned based on the best BLAST hit for each protein. Functional annotation data from additional database (PFAM, KEGG, GO, etc.) was collected for each match from the SWISSPROT database for downstream functional interrogation. Second, each predicated protein was subjected to InterProScan analysis using the PfamA-XX database, which reports predicted proteins with recognizable protein family motifs using hidden markov models. The two pieces of functional annotation data was applied to the ‘maker max’ dataset to refine the predicted gene list into the ‘maker standard’ dataset (Campbell, et al. 2014). The resulting ‘maker standard’ dataset is comprised of 27,677 genes, comprised of proteins with either/and: 1) an AED score > X, 2) a recognized Pfam protein domain, or 3) homology to a known protein in the Swiss-Prot database.

**Transcriptome Assembly and Sequencing**

*Sample Preparation*

For Illumina RNA-seq, tissue samples were taken from a multiple juvenile *Paramormyrops kingsleyae* collected in 2009 from Bambomo Creek (PKING1, PKING2, and PKING3; EOD for all specimens was P0-absent). Tissues sampled were: brain, kidney, electric organ, skeletal muscle, the skin along the flank (unenriched for electroreceptors), the skin of the face (enriched for electroreceptors), spinal cord, gonadal tissue, gill, spleen, pectoral fin, and swim bladder. Each tissue sample was dissected freshly into RNAlater, stored at 4˚C overnight, and then transferred to -20˚C until RNA isolation. RNA-isolation was performed using an RNA-easy Mini Kit for lipid tissues (Qiagen). Quality and concentration were assessed using a spectrophotometer and Qubit. RNA integrity was assessed using a BioAnalyzer (Agilent), all RNA samples sequenced had a RIN score >8 . Sequencing libraries were prepared using Illumina TruSeq Stranded mRNA Library Prep Kit LT.

*Transcriptome Assembly*

Raw sequences from all tissues were examined using the FastQC/MultiQC pipeline prior to quality trimming. Sequences were trimmed using Trimmomatic using the following settings: TruSeq3-PE.fa:2:30:10 LEADING:5 TRAILING:5 SLIDINGWINDOW:4:5 MINLEN:25. Sequenced and trimmed sequences from PKINGS2 were digitally normalized and assembled using the Trinity RNA-seq pipeline (2.4.0), using both standard de-novo and genome-guided procedures. The resulting transcriptomes were merged using PASA into a final ‘master transcriptome’ that was used for genome annotation as described above using default parameters.

*Transcriptome Quantification*

To reduce potential problems resolving ambiguous relationships between the annotated genes and the 800,000 predicted transcripts, quantification of genes was based on the MAKER predicted transcript annotations. Trimmed RNA-seq reads were aligned to the predicted transcripts resulting from gene annotation using bowtie2 and quantified using RSEM at the gene level (counts of all predicted transcripts from a gene were summed by RSEM per each gene locus) using the Trinity (2.4.0) pipeline scripts. RSEM generates two sets of matrices: raw counts of reads per gene, and transcripts per million (TPM) values for each gene. Using edgeR (Robinson, et al. 2010) the TPM was normalized using the TMM method for direct comparison of gene expression between libraries.

To minimize the potential for erroneous conclusions based on sample swapping, contamination, etc. we examined within and between sample correlations using the “PtR” software included in the Trinity release. Between-replicate correlations were >.90 for all tissues examined. Next, between sample correlations were examined to ensure that similar tissue samples were highly correlated. This was examined in two ways. First, a threshold of >10 counts per million (CPM) was applied to each gene, and then pairwise correlation between every replicate of every tissue was calculated. Sample distances were then calculated, and hierarchically clustered. Second, dimensional reduction using principal components analysis was applied to the thresholded expression matrix. Outputs were visually scanned to ensure that samples clustered tightly with between-replicates and relatively distantly between samples.

This ad-hoc QC procedure led to the rejection of three samples: both replicates of spinal cord (reason: weak correlation between samples, stronger correlation with other samples) and one sample of spleen from PKING3 (reason: PKING3 did not cluster with PKING1+2, and more closely with unrelated sample). We removed these samples from all downstream analyses.

*Analysis of gene-expression profiles*

Genes with TPM values of 0 across all samples were removed from our analysis, resulting in analysis of 26,348 genes (or 95.1% of Maker-standard predicted genes). We considered expression within a sample to be greater than 1 TPM. We computed the number of genes expressed in each of the tissues surveyed as a percent of the 26,348 genes. We ranked gene expression in each tissue from most to least abundant +SD based on TMM-normalized TPM in tables X-Y.

*Identification of housekeeping genes*

Following criteria established in mouse and human (Eisenberg and Levanon 2013; Li, et al. 2017), we defined housekeeping genes in mormyrid electric fish based on the following criteria: 1) highly expressed in all biological samples (TPM > 5) 2) low variability in expression within tissues (std. deviation across replicates < 1 TPM) and 3) no more than 2-fold differential expression in any single tissue over other tissues. This identified 1132 housekeeping genes. In addition to this set of “low-stringency” genes, we also identified a set of 21 genes that were very highly expressed (> 50 TPM in all tissues) with very low variability (std. deviation < 0.75 TPM) with no greater than 2-fold differential expression of the single tissues examined.

*Analysis of Differential Expression*

We examined patterns of differential expression for each of the 26,348 predicted genes with evidence of expression between all pairs of tissues using edgeR in ‘classic mode’. For all analyses, we considered a gene to be differentially expressed that had a fold change of 2 or greater between tissues, with a false-discovery rate corrected p-value < 0.001. We determined that 13,054 unique genes exhibited differential expression in at least one tissue based on these criteria.

***Supplemental Analyses:***

*Examination of the Head and Flank Skin Transcriptomes*

We found 186 differentially expressed genes (FDR-corrected p < 0.001 CPM > 2) between head skin (enriched for electroreceptors) and flank skin (relatively few electroreceptors). One of these genes was upregulated in head skin and 185 were upregulated in flank skin. The only gene upregulated in head skin corresponded to a protein of unknown homology. Evidence for the existence of this gene was supported strongly by evidence through annotation (AED=0.32), however homology to other proteins or signatures of Pfam domains were detected. Next, differentially expressed genes were examined for significantly (FDR < 0.05) over- and under-represented GO terms in each tissue type. There were 34 GO-terms significantly enriched among those genes upregulated in flank skin. The results unexpectedly showed a number of GO-terms that are normally associated with skeletal muscle, suggesting possible contamination through the dissection process with small muscle fibers directly attached to the flank skin tissue. This suggests that comparison of head and skin transcriptomes may not be an effective means of comparing electroreceptor-specific gene expression. Potential future studies might attempt dissection of individual electroreceptor organs using manual approaches or laser capture microdissection.

***Supplemental References***

Campbell MS, Holt C, Moore B, Yandell M 2014. Genome Annotation and Curation Using MAKER and MAKER-P. Curr Protoc Bioinformatics 48: 4 11 11-39. doi: 10.1002/0471250953.bi0411s48

Cantarel BL, et al. 2008. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res 18: 188-196. doi: 10.1101/gr.6743907

Eisenberg E, Levanon EY 2013. Human housekeeping genes, revisited. Trends Genet 29: 569-574. doi: 10.1016/j.tig.2013.05.010

Hinegardener R, Rosen DE 1972. Cellular DNA content and the evolution of teleostean fishes. American Naturalist 106: 621-644.

Li B, et al. 2017. A Comprehensive Mouse Transcriptomic BodyMap across 17 Tissues by RNA-seq. Sci Rep 7: 4200. doi: 10.1038/s41598-017-04520-z

Robinson MD, McCarthy DJ, Smyth GK 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26: 139-140. doi: 10.1093/bioinformatics/btp616

***Description of Supplemental Files***

**File S1: List of Short Read Libraries (XLS)** – Describes all short-read libraries generated for this study. All short-read libraries were deposited to the NCBI with the BioProject Accession Number PRJNA417677.

**File S2: List of Differentially Expressed Genes Across Replicates (XLS)** – An annotated list of all differentially expressed genes in at least one tissue, together with average and standard deviation of expression values in each tissue sampled.

**File S3: List of Housekeeping Genes (XLS)** – A listing of all housekeeping genes with average expression values in all tissues sampled for both ‘high stringency’ and ‘low stringency’ criteria (see Supplemental Methods, above).

**File S4: Identities of SM\_EO DE Genes (XLS)** – An annotated list of genes differentially expressed between skeletal muscle (SM) and electric organ (EO. Log fold-change values are reported between SM and EO together with false-discovery rate corrected p-values are reported, along with per-library expression values for each tissue.

**File S5: Identities of Enriched and Depleted GO-Terms in Skeletal Muscle and Electric Organ (XLS**) – A listing of all enriched and depleted GO-terms for skeletal muscle and electric organs, together with false-discovery rate corrected p-values for each term reported.