BMC Genomics

RESEARCH

Open Access

De novo assembly and annotation of the singing mouse genome

Samantha K. Smith[1*], Paul W. Frazel[2†], Alireza Khodadadi-Jamayran[3†], Paul Zappile[4], Christian Marier[4], Mariam Okhovat[1,5], Stuart Brown[6,7], Michael A. Long[2], Adriana Heguy[4] and Steven M. Phelps[1]

Abstract

Background Developing genomic resources for a diverse range of species is an important step towards understanding the mechanisms underlying complex traits. Specifically, organisms that exhibit unique and accessible phenotypes-of-interest allow researchers to address questions that may be ill-suited to traditional model organisms. We sequenced the genome and transcriptome of Alston's singing mouse (*Scotinomys teguina*), an emerging model for social cognition and vocal communication. In addition to producing advertisement songs used for mate attraction and male-male competition, these rodents are diurnal, live at high-altitudes, and are obligate insectivores, providing opportunities to explore diverse physiological, ecological, and evolutionary questions.

Results Using PromethION, Illumina, and PacBio sequencing, we produced an annotated genome and transcriptome, which were validated using gene expression and functional enrichment analyses. To assess the usefulness of our assemblies, we performed single nuclei sequencing on cells of the orofacial motor cortex, a brain region implicated in song coordination, identifying 12 cell types.

Conclusions These resources will provide the opportunity to identify the molecular basis of complex traits in singing mice as well as to contribute data that can be used for large-scale comparative analyses.

Keywords Genome, Rodents, Vocal communication, Social cognition

Background

The rapid development of sequencing tools in the last 20 years has allowed interrogation of coding and non-coding sequence evolution [1–6], gene regulation [7–15], protein-genome interactions [16–19], and many other processes [20–22]. Although the initial focus of genomics was on a few model organisms, nontraditional rodents have proved particularly useful subjects because of their diverse phenotypes and the ease with which tools developed for model rodents can be applied to them. For example, many of the molecular and neural tools developed in laboratory mice and rats (viral vectors, antibodies, other reagents) are easily adapted to other rodent species, and the extensive mapping of the rodent brain provides a strong foundation for understanding the variation in the neurobiology of complex behaviors. Genomic resources are essential for work with nontraditional

†Paul W. Frazel and Alireza Khodadadi-Jamayran contributed equally to this work.

*Correspondence:
Samantha K. Smith
samksmith@utexas.edu
[1] Department of Integrative Biology, University of Texas at Austin, Austin, TX 78712, USA
[2] Department of Neuroscience and Physiology, New York University Grossman School of Medicine, New York, NY 10016, USA
[3] Applied Bioinformatics Laboratory, New York University Grossman School of Medicine, New York, NY 10016, USA
[4] Genome Technology Center, New York University Grossman School of Medicine, New York, NY 10016, USA
[5] Present Address: Oregon Health & Science University, Portland, OR, USA
[6] NYU Center for Health Informatics and Bioinformatics, New York University Grossman School of Medicine, New York, NY 10016, USA
[7] Present Address: Exxon Mobil Corporate, Houston, TX, USA

species, because they allow more detailed analysis of gene expression, sequence-driven manipulations of gene function, as well as comparative analysis of genome evolution [23].

Alston's singing mouse, *Scotinomys teguina*, produces a unique, easily quantifiable vocal display that makes it an excellent model for understanding the genomic mechanisms of complex, behavioral traits. These diurnal rodents live in the montane grasslands of central America and are obligate insectivores [24]. Singing mice are named for the long, elaborate songs they use for mate attraction and male-male competition [24–28]. Their unique natural history as well as their complex social interactions make singing mice an excellent candidate for exploring the mechanisms and evolution of traits such as circadian rhythms, diet and energy balance, the challenges of thermoregulation or high-altitude living, dynamic vocal communication, and more.

Unlike model rodents such as house mice, singing mice produce highly structured advertisement songs (Fig. 1) which make them an emerging model for social cognition and vocal communication. These songs consist of rapidly repeated, frequency-modulated notes which span ~ 16 kHz in as little as 12 ms [29]. Note amplitudes, frequencies, and repetition rates are modulated over the course of the song (Fig. 1). Singing mice have highly structured vocalizations that are rapidly exchanged with conspecifics (counter-singing) in a manner whose timescale resemble human conversational speech, a feature not found in house mouse communication [30].

In addition to variation among species [26, 29], the advertisement song also varies between individuals [31] and populations of singing mice [29]. Among individuals, both internal and external cues drive song differences. For example, androgen levels and adiposity signals such as circulating leptin are associated with differences in "song effort" measures (e.g., song length), but not spectral features (e.g., frequency bandwidth) which may be set by vocal anatomy [27, 28, 31–33]. The social environment also tunes vocal output. For example, when other males are present, male singing mice produce longer songs and rapidly turn-take during singing bouts, finely coordinating song onset, and offset [30]. Together, these unique song features and the complexity of cues that impact song provide an exemplary opportunity to understand many fundamental questions such as how internal and external cues are integrated to modulate behavior, how elaborate vocalizations evolve, and more. The development of genomic resources for singing mice will provide opportunities to explore these questions as well as contribute to comparative work.

We sequenced the singing mouse genome and transcriptome using PromethION, Illumina, and PacBio technologies. We next assembled and annotated the genome and transcriptome and examined gene expression to assess the utility of our assemblies. Finally, we extracted and sequenced cell nuclei from the orofacial motor cortex (OMC) using 10X Genomics. The OMC was chosen because it plays an important role in the social modulation of singing behavior, namely counter singing between conspecifics [30]. To identify cell-types within the OMC, we used gene expression analysis. Data associated with this project are deposited under NCBI BioProject PRJNA878522 and raw data are available on GEO (GSE212957 series) and the NCBI assembly database. The annotated genome and transcriptome data can be viewed on the UCSC genome browser (https://genome.ucsc.edu/). A well-annotated genome and transcriptome will enable future work identifying the genomic substrates of a variety of physiological and behavioral adaptations.

Methods

DNA isolation

All animal procedures were approved by the University of Texas at Austin and New York University Grossman School of Medicine IACUC. For PacBio and Illumina

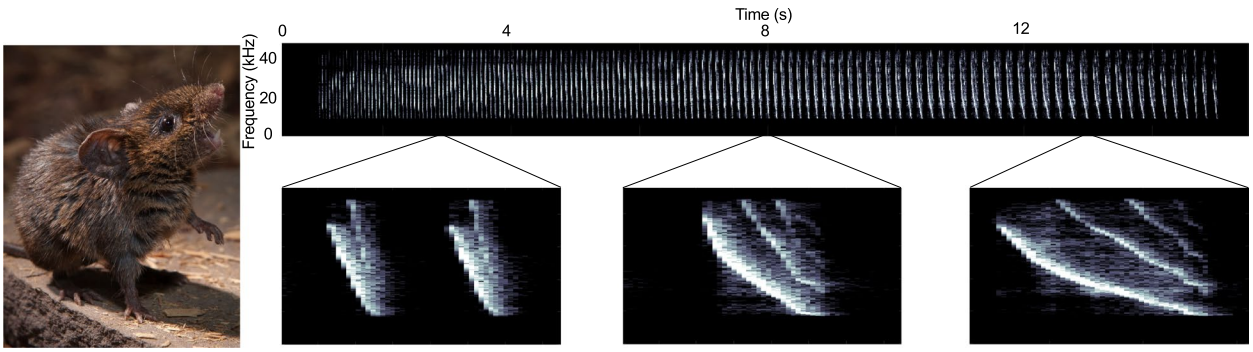

Fig. 1 (Left) A singing mouse and (right) a spectrogram of a representative advertisement song. Insets below show how frequency bandwidth, note shape, and note length change over the course of a song. Each inset is 0.15 s. Photo: Long lab

sequencing, we sacrificed 1 adult, male singing mouse. Liver and brain were dissected out and flash-frozen immediately. GDNA was then extracted from tissues using a Qiagen DNeasy kit. We visualized DNA integrity on an agarose gel and quantified DNA quality using a nanodrop. For PromethION sequencing, we sacrificed an adult, male singing mouse, extracted its liver, and immediately froze the tissue. High molecular weight DNA was extracted from liver tissue using a Genomic-tip 20/G DNA kit (QIAGEN, 10223).

DNA sequencing and genome assembly

We did library preparations and sequencing using PromethION technology (Oxford Nanopore MinION) at the New York University Langone Medical Center.

PacBio library preparation and sequencing was done at Duke University using 6–8 kb insert sizes with subreads ranging from 2 KB-3 KB (PacBio RS platform). We did Illumina library preparation and sequencing at the University of Texas Austin Core facility. Two Illumina libraries were created: a fragmentation library consisting of 170, 400, and 900 bp segments (PE Barcode + 2 × 100, 3 lanes = 510 M reads requested) and a mate-pair library with a 3 KB insert size (PE Barcode + 2 × 100, 1 lane = 170 M reads requested). Libraries were sequenced on an Illumina HiSeq 2500.

We assembled long reads from PacBio and PromethION and short reads from 10X genomics using the mixed read assembly tool MaSuRCA (v. 3.2.8) [34]. The assembled reference genome was masked for repeats using RepeatMasker (v. 1.332) [35].

RNA isolation and sequencing

RNA extraction and sequencing was done at UT Austin and NYULMC. All animals were sacrificed using isoflurane overdose. At UT Austin, forebrain, hindlimb skeletal muscle, gonads, and liver were dissected from 1 adult male singing mouse and immediately flash frozen. We extracted total RNA using a standard TRIzol method. RNA was then submitted to the UT Core facility for library preparation and Illumina sequencing. For RNAseq performed at NYU, we extracted RNA from freshly dissected tissue of two male and two female singing mice (liver and brain) using a Qiagen RNeasy Mini kit (Qiagen 74,104). We then homogenized the tissue using a rotor–stator homogenizer with disposable tips and did on-column DNAse digestion following manufacturer's instructions. An automated system performed poly-A library prep, and samples were run on a single-read Illumina HiSeq 4000 flowcell.

Transcriptome assembly, and analysis

Transcriptome assembly

We assembled a de novo and reference guided transcriptome using Trinity (v2.8.4) [36, 37]. For the guided transcript assembly all the RNA-Seq reads were mapped to the assembled reference genome using STAR mapper (v2.5.0c) [38]. Alignments were guided by a Gene Transfer Format (GTF) file. For quality control, we mapped the RNA reads to the assembled transcripts provided by Trinity [36, 37]. More than 83% of the reads mapped properly, suggesting a high-quality transcript assembly. To assess the completeness of the transcriptome and genome assembly, we used BUSCO (v. 5.4.5) [39].

Differential expression analysis

We calculated the mean read insert sizes and their standard deviations using Picard tools (v. 1.126) [40]. Read count tables were generated using HTSeq (v0.6.0) [41] and normalized based on library size factors using DEseq2 [42]. For differential expression analysis, we used BEDTools (v2.17.0) [43] and bedGraphToBigWig (v. 4; ENCODE) [44, 45] to generate read-per-million (RPM) normalized BigWig files. To compare gene expression across samples and their replicates, we used principal component analysis and Euclidean distance-based sample clustering. All downstream statistical analyses and plot generation were performed in R (v3.1.1) [46].

Functional enrichment

GO and KEGG analyses. To assess the accuracy of transcriptome assembly and annotation, GO MWU [47] and KEGG [48–50] analyses were used. The GO MWU method of gene ontology (GO) enrichment analysis uses a ranked list of genes to identify whether each GO category is significantly enriched with up- or down-regulated genes [47, 51]. We did functional enrichment analysis of GO and KEGG Reactome pathways using g:Profiler (v. e101_eg48_p14_baf17f0) with a g:SCS significance threshold of 0.05 [52]. Ordered gene lists for each tissue type included only those that had a |fold-change| of at least 2. We exported GO functional enrichment results from g:Profiler and created network pathways [53] using the EnrichmentMap application [54] in Cytoscape [55]. Maps were created with FDR Q value < 0.01 and combined coefficients > 0.375 with a combined constant of 0.5. We used an expression file of normalized fold-change values to create heatmaps of genes enriched pathways. We then used AutoAnnotate to interpret the function of groups of nodes in the network.

Genome annotation

RNA-Seq reads from Illumina and the assembled reference genome were used to create transcript-backed and prediction-based annotations. We concatenated both the guided and de novo transcriptome assembly results and cleaned them using the PASA pipeline (v. 2.5.3) [56] for UniVec vector sequences [57]. Cufflinks [58–61] was used to make a GTF file for PASA pipeline and the tdn.accs file was made using the de novo assembly. We used Stringtie2 (v. 2.2.0) [62] to make a another GTF file which was then passed to TransDecoder (v. 5.5.0, Haas, BJ. https://github.com/TransDecoder/TransDecoder) to identify coding regions within transcripts. We then used three different ab initio predictors, GlimmerHMM (v. 3.0.4) [63], GeneMarkHMM [64], and Augustus (v. 3.5.0) [65] and combined the resulting GFF3 files. Miniprot (v. 0.12) [66] and Uniref100 [67] were used for protein alignment and prediction. Finally, all output files from PASA, TransDecoder, the three ab initio predictor programs, and miniprot were passed to EVidenceModeler (v. 2.1.0) [68, 69] to generate a complete and comprehensive annotation file. The resulting GFF was converted to a GTF for downstream analyses (bulk RNA-Seq and snRNA-Seq). A cDNA fasta file was produced from the GTF and used as an input for BLAST [70]. We blasted the cDNA file against the entire Uniprot database (all organisms; [71]). BLAST results were then used to annotate the GTF file with gene symbols.

Single nuclei sequencing and analysis

Single nuclei sequencing

We tagged the orofacial motor cortex (OMC) area from one adult male singing mouse for extraction via stereotaxic injection of fluorescent dextran beads into the brain as previously described [30]. Post injection, the mouse was immediately transcardially perfused with ice-cold artificial cerebrospinal fluid (aCSF). We sectioned the brain into 250 µm sections, located the dyed area under a dissecting microscope, and removed the region with a scalpel. Extracted tissue was immediately flash frozen in liquid nitrogen and stored overnight at -80 °C. We dissociated nuclei using a modified version of the Mccarroll lab protocol [72]. FITC-tagged NeuN antibody (Sigma, MAB377) was prepared following manufacturer's instructions (Abcam, 188,285), and used to enrich for NeuN+, DAPI+nuclei on a MoFlo XDP flow cytometer (Beckman Coulter) with a 100 µm nozzle. We loaded 9000 sorted nuclei into GEMs on a 10X Genomics Chromium Controller (1st generation, 10X v3 chemistry) using 3' v3 chemistry and recovered 3500 high-quality nuclei after standard analysis (10X Genomics CellRanger pipeline v. 3.1.0) [73].

Single nuclei gene expression analysis

Single-nuclei gene expression data were generated using the 10X Genetics Chromium system, following the manufacturer's instructions for sample and library prep. We aligned raw FASTQ files to the singing mouse transcriptome and then assigned reads to individual nuclei via the 10X CellRanger pipeline. The resulting gene expression matrix was analyzed using the standard Seurat package (v. 3) [74] in RStudio (v. 4.0.2). We excluded genes with expression in < 3 nuclei from the analysis. We filtered the expression data to only keep nuclei with fewer than 11,500 genes, and fewer than 40,000 molecules detected, excluding 14 likely doublet nuclei.

Singlet nuclei gene expression data were then log-normalized using the Seurat pipeline [74] and only the top 2,000 most variable genes were selected for downstream analysis. We ran PCA on the top 2,000 variable genes using standard Seurat settings and clustered nuclei via standard commands using the first 20 principal components. A resolution value of 0.1 was used to capture large, cell-type-level clusters of similar nuclei, resulting in 12 clusters that were categorized into major cell types based on known marker genes. We did dimensional reduction via two standard methods, tSNE (t-distributed Stochastic Neighbor Embedding) [75] and UMAP (Uniform Manifold Approximation and Projection) [76, 77]. UMAP better distinguished clusters and was used for downstream analyses. Marker genes for each cluster (genes significantly enriched) were identified using the standard threshold values of > 0.25 percent of nuclei expressing the gene and > 0.25 log-fold change. We plotted the top 10 markers genes for each cluster on a heatmap and compared these with known marker genes to determine what cell types are represented by each cluster.

Results

The annotated singing mouse genome and transcriptome can be viewed at the UCSC genome browser (https://genome.ucsc.edu/).

Genome and transcriptome assembly and annotation

After assembly, the total genome size was 2.4 Gb. After scaffolding there were 7806 contigs. We assembled both a de novo and reference guided transcriptome and identified 754,907 transcripts. Assembly quality and completeness metrics can be found in Table 1.

We annotated the genome using the PASA pipeline and resulting GTF file can be accessed at the UCSC genome browser (https://genome.ucsc.edu/). We annotated genes using blast and only included annotations for genes that had at least 80% sequence similarity to the reference gene (14,989 genes included). Our scaffolds have 92.5% of the complete, single-copy BUSCOs (Table 1).

Table 1 Assembly characteristics

Assembly Characteristics	
BUSCOs	
Complete	3,212 (95.8%)
Complete and single-copy	3,102 (92.5%)
Complete and duplicated	110 (3.3%)
Fragmented	61 (1.8%)
Missing	81 (2.4%)
Genome Assembly Statistics	
Total length	2,401,463,659
Number of Scaffolds	7,806
Number of contigs	7,806
Percent gaps	0.00%
Scaffold N50	1 MB
Contigs N50	1 MB
Transcriptome Statistics	
Alignment to assembled genome	83.41%
Total trinity genes	914,330
Total trinity transcripts	1,191,461
Percent GC	44.05%
All transcripts	
Contig N50	826
Median contig length	351
Average contig	620.03
Total assembled bases	738,746,092
Longest Isoform only	
Contig N50	513
Median contig length	327
Average contig	478.26
Total assembled bases	437,286,933

Validation

We validated the quality of the transcriptome and annotations by doing gene expression analyses. Reads were normalized using DEseq2 [42] and samples were clustered by Euclidean distance (Fig. 2). As expected, samples clustered by tissue type. Principal components analysis revealed two components that distinguished tissue type (Fig. 3). PC1 separated brain gene expression from that of the liver, while PC2 distinguished brain and liver expression from that of the muscle and gonads. We then compared gene expression profiles between pairs of tissues. To validate that we accurately mapped transcripts to annotated genes, we did GO MWU [47] and KEGG [48–50] analyses. We found that the metabolic pathways KEGG term was the most significantly enriched among all annotated genes within the genome (Fig. 4). We then did GO MWU [47] on each tissue-type gene list which we ranked by fold-change. This analysis revealed enrichment of expected GO terms based on tissue type. For example, genes upregulated in the brain were enriched with terms related to synapse structure and function (Fig. 5a). To further assess whether we detect of appropriate, tissue-specific gene expression, we constructed network pathways from the brain GO enrichment results. The analysis determined 389 gene sets ("nodes") and 770 instances of overlap between gene sets ("edges") that were sorted into 146 clusters (Fig. 5b). We found that the network was annotated with functions consistent with first-principles predictions based on the focal tissue. For example, the brain functional GO network was annotated with functions such as "postsynaptic membrane component".

Single nuclei sequencing of the Orofacial Motor Cortex (OMC)

We assessed the quality of the data using cellranger (114 outliers removed, 3486 nuclei retained). For 3,486 nuclei that passed quality control, we did dimensional reductions (see Methods) and displayed them using t-distributed stochastic neighbor embedding (t-SNE; Fig. 6a) [75] and uniform manifold approximation and projection (UMAP; Fig. 6b) [76, 77]. Major brain cell types in 12 clusters were clearly identifiable based on canonical cell-type marker gene expression (Fig. 6c). We identified 5 clusters of nuclei as excitatory neurons, expressing high levels of Syt1 (synaptotagmin-1), two clusters as inhibitory neurons, which expressed high levels of Gad-2 (glutamate decarboxylase 2), one cluster of astrocytes, expressing Gfap (glial fibrillary protein), and one cluster of oligodendrocytes, which expressed Mbp (myelin basic protein) [78–80]. Normalized expression of the top ten marker genes for each brain cell type clearly distinguished the 12 nuclei clusters using t-SNE and UMAP (Fig. 6d).

Discussion

We sequenced, assembled, and annotated a genome and transcriptome for Alston's singing mouse, a model for complex social behavior and vocal communication. Transcriptome and gene annotation quality were validated using gene expression and functional enrichment analyses. Finally, we did single-nuclei sequencing of cells of the orofacial motor cortex (OMC), a region involved in vocal turn-taking in singing mice [30] and identified major cell types. The annotated genome and transcriptome will be a valuable resource that will allow characterization of the genetic basis of complex traits in singing mice as well as be useful for comparative studies more broadly.

Genome and transcriptome assembly and annotation

By using three sequencing technologies, we were able to create a high quality de novo genome assembly. Short reads, like those generated by Illumina, provided the highest base-pair-level accuracy [81–83]. Longer reads

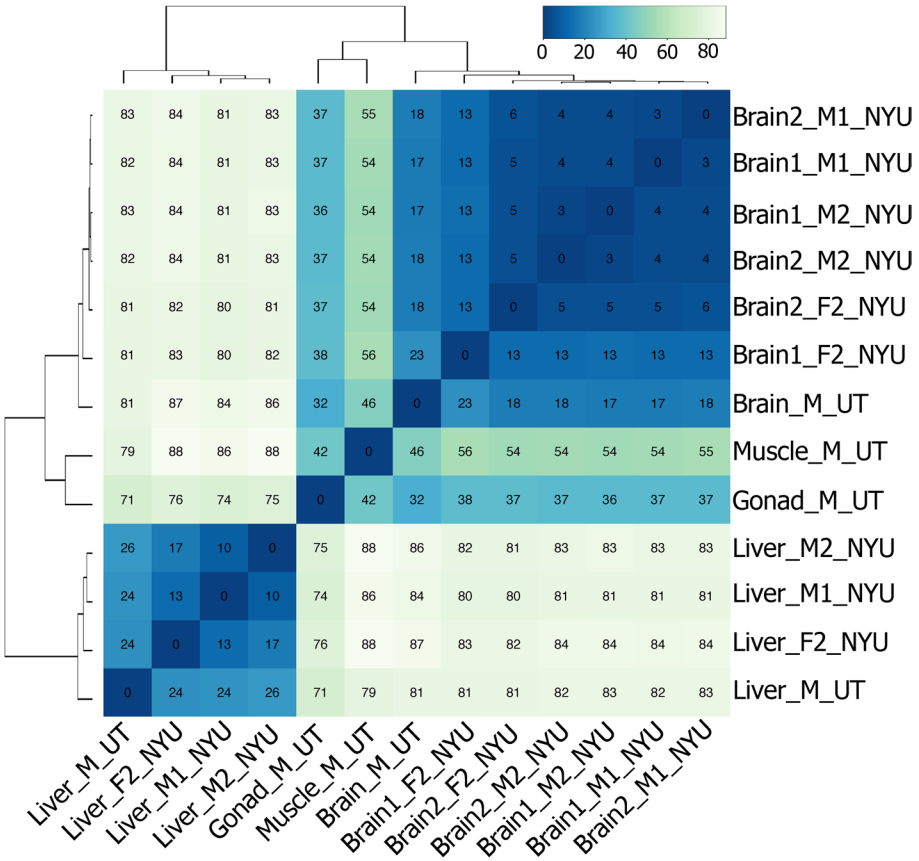

Fig. 2 Euclidean-distance-based heatmap shows that samples of the same tissue type have the most similar gene expression. Lower values (darker blue) indicate more similarity

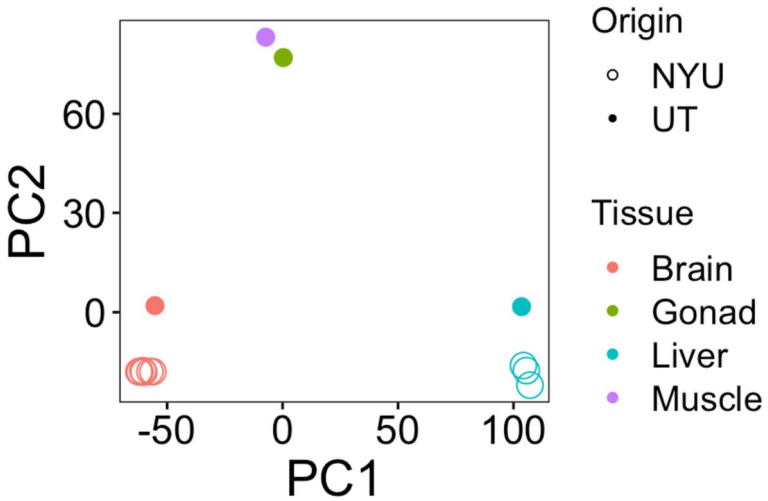

Fig. 3 Biplot of the first two principal components, which distinguish tissue types. Brain and liver gene expression drive PC1, while the differences between brain/liver gene expression and that of the gonad and liver underlie PC2. Dot color indicates tissue type and whether the circle is filled in or not indicates where the samples were collected

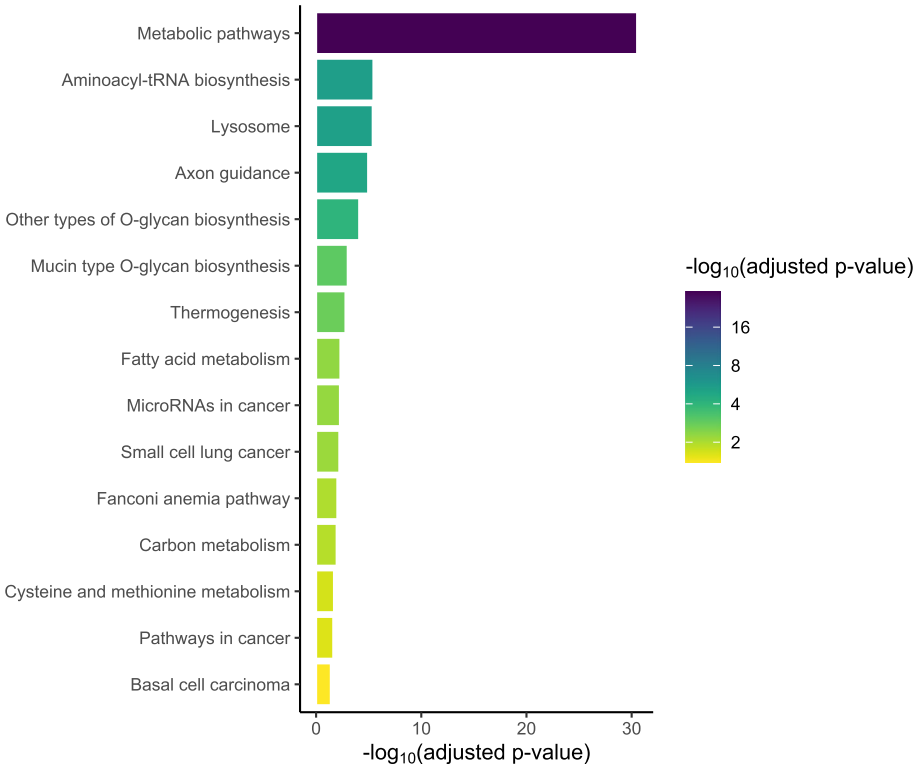

Fig. 4 Barplot of KEGG terms for all annotated genes shows that metabolic pathways are enriched in this dataset [49–51]

generated by PacBio's SMRT sequencing [84, 85], produced excellent contigs. Finally, contig scaffolding was facilitated by PromethION's nanopore sequencing [86, 87], which can sequence the longest stretches of DNA [88, 89].

These sequencing efforts culminated in a 2.4 Gb genome. The size of the singing mouse draft genome is like that of other sequenced rodents such as house mice, *Mus musculus*, (strain: C57BL/6 J, genome size: 2.5 Gb) [90] and white-footed mice, *Peromyscus leucopus* (genome size: 2.45 Gb) [91]. Using BUSCO [39], we found that our assembly captures much of the genome; our scaffolds contain 92.5% of the complete and single-copy BUSCOs (annotated collection of ubiquitous mammalian genes). However, DNA was extracted from two different singing mice for sequencing. Tissue from one individual was used for long read PromethION sequencing while the other was used to generate PacBio and Illumina reads. Using DNA from two individuals in our assembly could impact future analyses of standing genetic variation.

We assembled 754,907 transcripts into a de novo transcriptome with a contig N50 of 826. When aligned to the reference genome, 83.41% of the transcriptome mapped, indicating a quality transcriptome assembly [36, 37]. As expected, the contig N50 based on only the longest isoform per gene was lower than that of all transcripts since including all transcript isoforms can exaggerate N50 values.

We did differential expression and functional enrichment analyses to test the quality of the transcriptome and annotations. In support of our expectations for a quality assembly and annotation, we found tissue-specific gene expression profiles. Two distinct clustering methods showed that samples of the same tissues type, both technical and biological replicates, had identical (technical replicates) or very similar (biological replicates) expression patterns that differed greatly from other tissues. Functional enrichment analysis identified the putative function of differentially expressed genes across tissue type. Within the brain, we found enrichment of expected pathways such as synapse-related GO terms. A network-based approach supported these results, clustering related nodes into larger functional groups with brain-relevant annotations such as "glutamate neurotransmitter receptor." This approach is useful because GO functional categories often share many genes and the results of GO analyses can often be redundant [53, 55]. A network approach clusters by gene overlap which allows the annotation of groups of similar gene sets, rather than annotating each gene set independently. The results of these analyses suggest that our transcriptome assembly

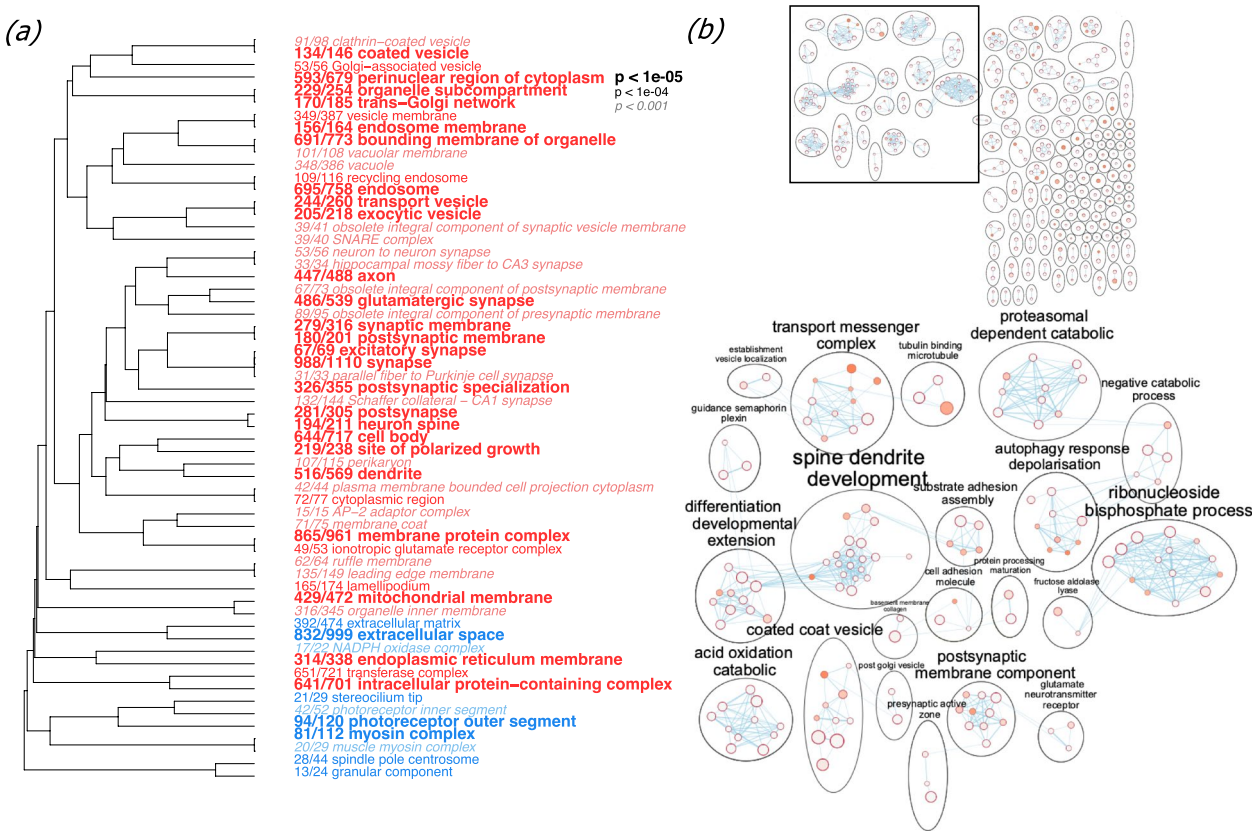

Fig. 5 **a** GO tree of enriched terms for ranked brain gene expression (GO division: cellular compartment) using Fisher's exact test. Font indicates significance and the fraction before each term shows the number of genes annotated with the GO term relative to the total number of such genes in the dataset. (**b**, top) A network plot of brain GO enrichment results was made in Cytoscape using EnrichmentMap (FDR Q value < 0.01, combined coefficients > 0.375, combined constant 0.5). Most highly connected nodes (**b**, bottom) are annotated using AutoAnnotate. Each node is a gene set and the size of each node represents the number of genes in the gene set. Edges (lines between nodes) represent overlap between gene sets and their width refer to the number of genes that are shared by the nodes. The color of each node represents enrichment scores (q-value)

is of high quality and the annotations we created are accurate. For our functional enrichment analyses, we focused on brain gene expression because we are interested in understanding how the nervous system drives complex behavior. A quality transcriptome assembly and gene annotations allowed us to examine gene expression of single nuclei to identify specific brain cell types that may contribute to behavior (see Single-nuclei sequencing of the OMC). Single-nuclei sequencing of regions implicated in song production [30, 92] can be paired with other approaches such as sequence-based interventions and epigenetic profiling to understand how the brain patterns vocal output.

The genome, transcriptome, and annotation GTF file can be viewed at the UCSC genome browser (https://genome.ucsc.edu/) and raw data accessed on GEO under series GSE212957. We identified 14,989 genes that have at least 80% sequence similarity to reference genes. This is on the order of annotation efforts in other species [90, 91, 93, 94]. We combined the PASA pipeline with multiple ab initio gene finding, protein homology, and weighted consensus gene structure tools which generated a more comprehensive genome annotation.

Single-nuclei sequencing of the OMC

Dimensional reductions of the expression profiles of 3,486 OMC nuclei revealed 12 clusters that were categorized into major cell types based on marker genes. Most nuclei fell into 7 clusters that were categorized as neuronal, which we expected, since we used NeuN antibodies to enrich for neuronal nuclei prior to sequencing. Five of these clusters were identified as containing nuclei of excitatory neurons, expressing high levels of Syt1 (synaptotagmin-1), gene encoding a $Ca2+$ sensory for neurotransmitter release [80, 95–97]. Two inhibitory interneuron clusters were identified by expression of Gad-2 (glutamate decarboxylase 2), which encodes an enzyme that catalyzes the synthesis of GABA, an inhibitory neurotransmitter [80, 98]. Despite selecting for neuronal nuclei, a few small clusters of other cell types were

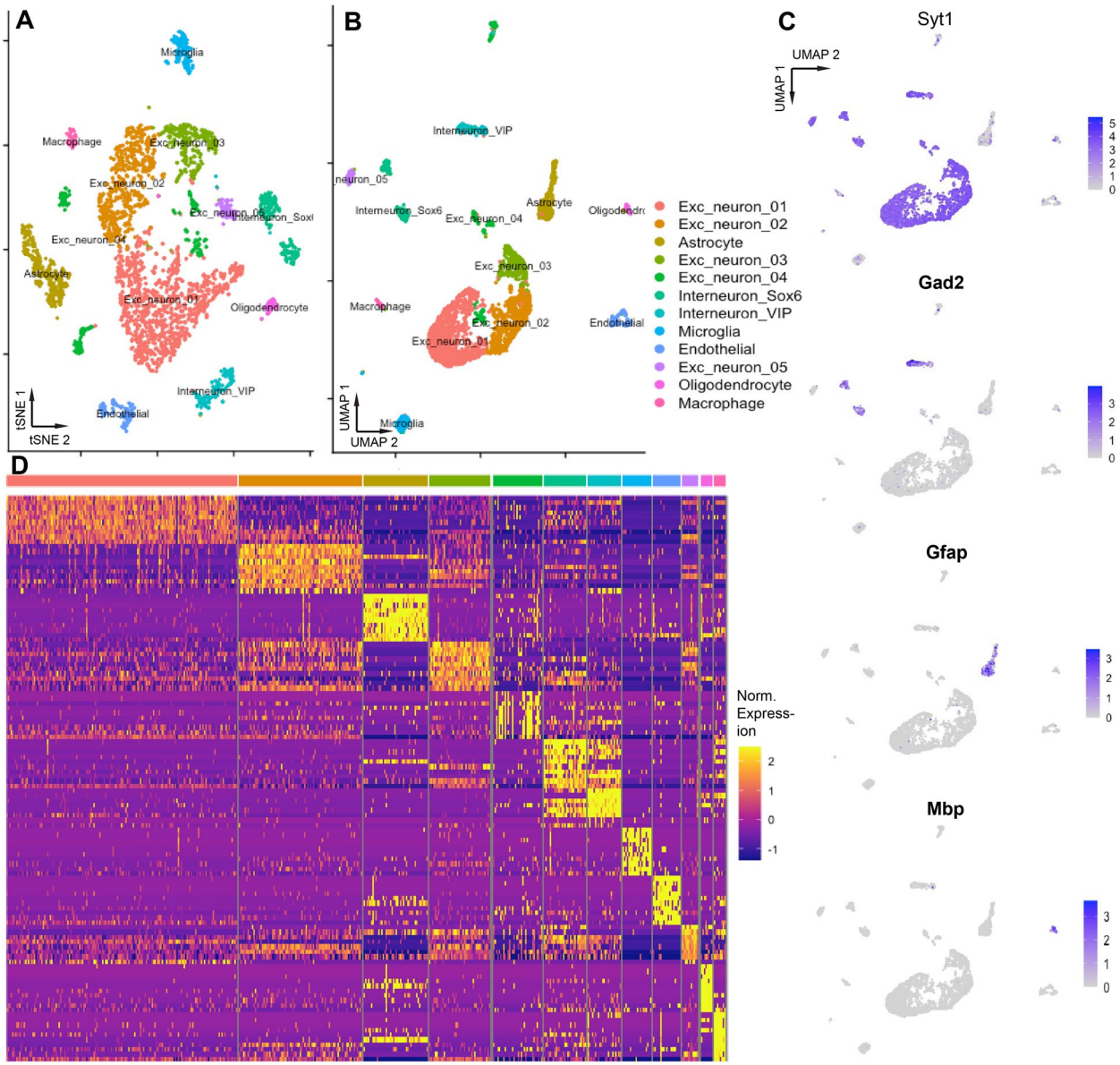

Fig. 6 The *S. teguina* genome and transcriptome enable identification of brain cell types in a single-nuclei RNAseq dataset from brain area OMC. **A** tSNE dimensional reduction of 3486 nuclei isolated from brain area OMC. **B** UMAP dimensional reduction of same data **C**. Feature plots of normalized gene expression data for key brain cell type markers (Syt1: synaptotagmin-1, neuron; Gad2: glutamate decarboxylase 2, inhibitory neuron; Gfap: glial fibrillary protein, astrocyte; Mbp: myelin binding protein, oligodendrocyte). **D** Heatmap of normalized gene expression data for the top 10 marker genes for each brain cell type identified

also identified such as astrocytes (high Gfap) and oligodendrocytes (Mbp) [80]. Clear identification of brain-cell types demonstrates the robustness of the singing mouse transcriptome and genome and demonstrates the broad applicability of 10X single cell/single nuclei technology to a nontraditional rodent species. The brain region we chose for single nuclei analysis is an important temporal regulator of the singing mouse advertisement song [30]. By combining single-cell analysis of relevant brain regions with circuit-level studies [92], we can develop hypotheses about the role of each network node and the cellular mechanisms that underlie these functions. Together, these resources allow us to examine how the nervous system directs complex behavior.

Uses of these resources

The contribution of a high-quality singing mouse genome and transcriptome increases the diversity of

available model species and improves our ability to ask mechanistic questions. Singing mice are a particularly useful model for understanding how the brain drives complex behavior due to their unique, quantifiable phenotype, their tractability in the lab, and our ability to adapt existing neurobiological tools and resources for singing mice. The addition of genomic resources provides further opportunity to use singing mice to study novel questions in social cognition. For example, to explore the genomic basis of complex traits, we could use the genomic resources we have generated to examine gene regulation (e.g., ChIP-seq: [99]; ATAC-seq: [100]), sequence evolution (e.g., tests of selection: [101–103]), and more. These data also contribute to a library of resources that can be used for larger comparative analyses. Increasing the diversity of model systems, through the addition of species that are well-suited to particular questions, is essential to understanding the mechanisms that drive complex traits.

Abbreviations

aCSF	Artificial cerebrospinal fluid
BLAST	Basic local alignment search tool
cDNA	Complementary DNA
DAPI	4′,6′-Diamidino-2-phenylindole
ENCODE	Encyclopedia of DNA elements
FITC	Fluorescein isothiocyanate
GDNA	Genomic DNA
GEM	Gel bead-in emulsion
GO MWU	Gene ontology Mann–Whitney U test
GTF	Gene transfer format
KEGG	Kyoto encyclopedia of genes and genomes
OMC	Orofacial motor cortex
PASA	Program to assemble spliced alignments
PCA	Principal components analysis
snRNASeq	Single nuclei RNA sequencing
STAR	Spliced transcripts alignment to a reference
tSNE	T-distributed stochastic neighbor embedding
UMAP	Uniform manifold approximation and projection for dimension reduction

Acknowledgements

We would like to thank the Genome Technology Center (GTC) and UT Genomic Sequencing and Analysis Facility (GSAF) for expert library preparation and sequencing. We are grateful to the Applied Bioinformatics Laboratories (ABS) for providing bioinformatics support and helping with the analysis and interpretation of the data. GTC and ABL are shared resources partially supported by the Cancer Center Support Grant P30CA016087 at the Laura and Isaac Perlmutter Cancer Center. This work has used computing resources at the NYU School of Medicine High Performance Computing (HPC) Facility and the Texas Advanced Computing Center (TACC). We also thank the anonymous reviewers whose suggestions have improved this manuscript.

Authors' contributions

SMP, AH, SB, and MAL designed the research. AH and MO generated data. PWF generated snRNAseq data and analyzed the single nuclei expression data. CM, AK-J, SB, PZ, and SKS analyzed genome and bulk transcriptome data. SKS wrote the manuscript. SKS, PWF, and AK-J prepared figures and tables. All authors read and approved the final version of the manuscript.

Funding

This work was funded by a Cancer Center Support Grant P30CA016087 (AH), PacBio Sequel National Institutes of Health Shared Instrumentation Grant 1S10OD023423-01 (AH), NIH R01 NS113071 (MAL and SMP), and NSF IOS 1457350 (SMP).

Availability of data and materials

Data were deposited under NCBI BioProject PRJNA878522. We deposited DNA, bulk RNA, and single nuclei RNA sequencing data on GEO under GSE212957 series. PromethION sequencing data are under sample GSM6564485, PacBIO data under sample GSM6564486, and Illumina DNA sequencing data under sample GSM6564487. Bulk RNA data generated at UT Austin are under samples GSM6564488, GSM6564489, GSM6564490, and GSM6564491. Bulk RNA data from NYU are under samples GSM6564492, GSM6564493, GSM6564494, GSM6564495, GSM6564496, GSM6564497, GSM6564498, GSM6564499, and GSM6564500. Single nuclei sequencing data are under sample GSM6564501. Sequencing data and annotations can be browsed at the UCSC Genome Browser (https://genome.ucsc.edu/).

Declarations

Ethics approval and consent to participate

All experimental procedures were approved by the University of Texas Austin and New York University Grossman School of Medicine IACUC. All methods were done in accordance with IACUC regulations. We confirm that we have reported this research following the ARRIVE guidelines for the reporting of animal research.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 24 May 2023 Accepted: 14 September 2023
Published online: 26 September 2023

References

1. Dhanoa JK, Sethi RS, Verma R, Arora JS, Mukhopadhyay CS. Long non-coding RNA: its evolutionary relics and biological implications in mammals: a review. J Anim Sci Technol. 2018;60(1):25.
2. Necsulea A, Kaessmann H. Evolutionary dynamics of coding and non-coding transcriptomes. Nat Rev Genet. 2014;15(11):734–48.
3. Patthy L. Genome evolution and the evolution of exon-shuffling — a review. Gene. 1999;238(1):103–14.
4. Ulitsky I. Evolution to the rescue: using comparative genomics to understand long non-coding RNAs. Nat Rev Genet. 2016;17(10):601–14.
5. Ulitsky I, Bartel DP. lincRNAs: Genomics, Evolution, and Mechanisms. Cell. 2013;154(1):26–46.
6. Zhang J, Yang J-R. Determinants of the rate of protein sequence evolution. Nat Rev Genet. 2015;16(7):409–20.
7. Fraser P, Bickmore W. Nuclear organization of the genome and the potential for gene regulation. Nature. 2007;447(7143):413–7.
8. Haraksingh RR, Snyder MP. Impacts of variation in the human genome on gene regulation. J Mol Biol. 2013;425(21):3970–7.
9. He L, Hannon GJ. MicroRNAs: small RNAs with a big role in gene regulation. Nat Rev Genet. 2004;5(7):522–31.
10. Li Y, Hu M, Shen Y. Gene regulation in the 3D genome. Hum Mol Genet. 2018;27(R2):R228–33.
11. Pennacchio LA, Rubin EM. Genomic strategies to identify mammalian regulatory sequences. Nat Rev Genet. 2001;2(2):100–9.
12. Simonis M, Klous P, Splinter E, Moshkin Y, Willemsen R, de Wit E, et al. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4C). Nat Genet. 2006;38(11):1348–54.
13. Smallwood A, Ren B. Genome organization and long-range regulation of gene expression by enhancers. Curr Opin Cell Biol. 2013;25(3):387–94.

14. Würtele H, Chartrand P. Genome-wide scanning of HoxB1-associated loci in mouse ES cells using an open-ended Chromosome Conformation Capture methodology. Chromosome Res. 2006;14(5):477–95.
15. Zhao Z, Tavoosidana G, Sjölinder M, Göndör A, Mariano P, Wang S, et al. Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. Nat Genet. 2006;38(11):1341–7.
16. Belokopytova P, Fishman V. Predicting genome architecture: challenges and solutions. Front Genet. 2021;22(11):617202.
17. Chen K, Rajewsky N. The evolution of gene regulation by transcription factors and microRNAs. Nat Rev Genet. 2007;8(2):93–103.
18. Hobert O. Gene regulation by transcription factors and microRNAs. Science. 2008;319(5871):1785–6.
19. Kim TH, Ren B. Genome-wide analysis of protein-DNA interactions. Annu Rev Genomics Hum Genet. 2006;7(1):81–102.
20. Baack EJ, Rieseberg LH. A genomic view of introgression and hybrid speciation. Curr Opin Genet Dev. 2007;17(6):513–8.
21. Cain AK, Barquist L, Goodman AL, Paulsen IT, Parkhill J, van Opijnen T. A decade of advances in transposon-insertion sequencing. Nat Rev Genet. 2020;21(9):526–40.
22. Farré M, Ruiz-Herrera A. The plasticity of genome architecture. Genes. 2020;11(12):1413.
23. Matz MV. Fantastic beasts and how to sequence them: ecological genomics for obscure model organisms. Trends Genet. 2018;34(2):121–32.
24. Hooper ET, Carleton MD. Hooper & Carleton 1976. Available from: https://deepblue.lib.umich.edu/bitstream/handle/2027.42/56395/MP151.pdf?sequence=1. [Cited 2017 Aug 16].
25. Fernández-Vargas M, Tang-Martínez Z, Phelps SM. Singing, allogrooming, and allomarking behaviour during inter- and intra-sexual encounters in the Neotropical short-tailed singing mouse (Scotinomys teguina). Behaviour. 2011;148(8):945–65.
26. Miller JR, Engstrom MD. Vocal Stereotypy and Singing Behavior in Baiomyine Mice. J Mammal. 2007;88(6):1447–65.
27. Pasch B, George AS, Hamlin HJ, Guillette LJ, Phelps SM. Androgens modulate song effort and aggression in Neotropical singing mice. Horm Behav. 2011;59(1):90–7.
28. Pasch B, George AS, Campbell P, Phelps SM. Androgen-dependent male vocal performance influences female preference in Neotropical singing mice. Anim Behav. 2011;82(2):177–83.
29. Campbell P, Pasch B, Pino JL, Crino OL, Phillips M, Phelps SM. Geographic variation in the songs of neotropical singing mice: testing the relative importance of drift and local adaptation. Evolution. 2010;64(7):1955–72.
30. Okobi DE, Banerjee A, Matheson AMM, Phelps SM, Long MA. Motor cortical control of vocal interaction in neotropical singing mice. Science. 2019;363(6430):983–8.
31. Burkhard TT, Westwick RR, Phelps SM. Adiposity signals predict vocal effort in Alston's singing mice. Proc R Soc B Biol Sci. 1877;2018(285):20180090.
32. Giglio EM, Phelps SM. Leptin regulates song effort in Neotropical singing mice (Scotinomys teguina). Anim Behav. 2020;1(167):209–19.
33. Smith SK, Burkhard TT, Phelps SM. A comparative characterization of laryngeal anatomy in the singing mouse. J Anat. Available from: http://onlinelibrary.wiley.com/doi/abs/10.1111/joa.13315. [Cited 2021 Jan 6].
34. Zimin AV, Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. The MaSuRCA genome assembler. Bioinformatics. 2013;29(21):2669–77.
35. Smit, AFA, Hubley, R & Green, P. RepeatMasker Open-4.0.2013–2015. http://www.repeatmasker.org.
36. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29(7):644–52.
37. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity. Nat Protoc. 2013;8(8). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3875132/. [Cited 2021 Jan 20].
38. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinforma Oxf Engl. 2013;29(1):15–21.
39. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31(19):3210–2.
40. Picard Tools - By Broad Institute. Available from: http://broadinstitute.github.io/picard/. [Cited 2021 Aug 30].
41. Anders S, Pyl PT, Huber W. HTSeq–a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015;31(2):166–9.
42. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.
43. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841–2.
44. Davis CA, Hitz BC, Sloan CA, Chan ET, Davidson JM, Gabdank I, et al. The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res. 2018;46(D1):D794-801.
45. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57–74.
46. R: The R Project for Statistical Computing. Available from: https://www.r-project.org/. [Cited 2021 Aug 30].
47. Wright RM, Kenkel CD, Dunn CE, Shilling EN, Bay LK, Matz MV. Intraspecific differences in molecular stress responses and coral pathobiome contribute to mortality under bacterial challenge in Acropora millepora. Sci Rep. 2017;7(1):2609.
48. Kanehisa M. Toward understanding the origin and evolution of cellular organisms. Protein Sci. 2019;28(11):1947–51.
49. Kanehisa M, Furumichi M, Sato Y, Ishiguro-Watanabe M, Tanabe M. KEGG: integrating viruses and cellular organisms. Nucleic Acids Res. Available from: http://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkaa970/5943834. [Cited 2020 Dec 6].
50. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 2000;28(1):27–30.
51. Matz MV. Rank-based gene ontology analysis with adaptive clustering. 2021. Available from: https://github.com/z0on/GO_MWU. [Cited 2021 Aug 19].
52. Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, et al. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res. 2019;47(W1):W191–8.
53. Reimand J, Isserlin R, Voisin V, Kucera M, Tannus-Lopes C, Rostamianfar A, et al. Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. Nat Protoc. 2019;14(2):482–517.
54. Merico D, Isserlin R, Stueker O, Emili A, Bader GD. Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. PLoS One. 2010;5(11):e13984.
55. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13(11):2498–504.
56. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK, Hannick LI, et al. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res. 2003;31(19):5654–66.
57. The UniVec Database. Available from: https://www-ncbi-nlm-nih-gov.ezproxy.lib.utexas.edu/tools/vecscreen/univec/. [Cited 2021 Aug 30].
58. Roberts A, Pimentel H, Trapnell C, Pachter L. Identification of novel transcripts in annotated genomes using RNA-Seq. Bioinformatics. 2011;27(17):2325–9.
59. Roberts A, Trapnell C, Donaghey J, Rinn JL, Pachter L. Improving RNA-Seq expression estimates by correcting for fragment bias. Genome Biol. 2011;12(3):R22.
60. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010;28(5):511–5.
61. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol. 2013;31(1):46–53.
62. Kovaka S, Zimin AV, Pertea GM, Razaghi R, Salzberg SL, Pertea M. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. Genome Biol. 2019;20(1):1–3.

63. Majoros WH, Pertea M, Salzberg SL. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. Bioinformatics. 2004;20(16):2878–9.
64 Lukashin AV, Borodovsky M. GeneMark. hmm: new solutions for gene finding. Nucleic Acids Res. 1998;26(4):1107–15.
65. Stanke M, Diekhans M, Baertsch R, Haussler D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics. 2008;24(5):637–44.
66 Li H. Protein-to-genome alignment with miniprot. Bioinformatics. 2023;39(1):btad014.
67. Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics. 2007;23(10):1282–8.
68. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol. 2008;9:1–22.
69. Haas BJ, Zeng Q, Pearson MD, Cuomo CA, Wortman JR. Approaches to fungal genome annotation. Mycology. 2011;2(3):118–41.
70. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–10.
71. UniProt. Available from: https://www.uniprot.org/. [Cited 2021 Aug 30].
72. Bortolin L. Extraction of nuclei from brain tissue. 2020. Available from: https://www.protocols.io/view/extraction-of-nuclei-from-brain-tissue-2srged6. [Cited 2021 Aug 30].
73. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017;8(1):14049.
74. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, et al. Comprehensive integration of single-cell data. Cell. 2019;177(7):1888-1902.e21.
75. van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res. 2008;9(11):2579–605.
76. Becht E, McInnes L, Healy J, Dutertre C-A, Kwok IWH, Ng LG, et al. Dimensionality reduction for visualizing single-cell data using UMAP. Nat Biotechnol. 2019;37(1):38–44.
77. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. ArXiv180203426 Cs Stat. 2020. Available from: http://arxiv.org/abs/1802.03426. [Cited 2021 Jul 16].
78. Almanzar N, Antony J, Baghel AS, Bakerman I, Bansal I, Barres BA, et al. A single-cell transcriptomic atlas characterizes ageing tissues in the mouse. Nature. 2020;583(7817):590–5.
79. Yao Z, van Velthoven CTJ, Nguyen TN, Goldy J, Sedeno-Cortes AE, Baftizadeh F, et al. A taxonomy of transcriptomic cell types across the isocortex and hippocampal formation. Cell. 2021;184(12):3222-3241.e26.
80. Zhang Y, Chen K, Sloan SA, Bennett ML, Scholze AR, O'Keeffe S, et al. An RNA-sequencing transcriptome and splicing database of glia, neurons, and vascular cells of the cerebral cortex. J Neurosci. 2014;34(36):11929–47.
81. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008;456(7218):53–9.
82. Bentley DR. Whole-genome re-sequencing. Curr Opin Genet Dev. 2006;16(6):545–52.
83. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016;17(6):333–51.
84. Merker JD, Wenger AM, Sneddon T, Grove M, Zappala Z, Fresard L, et al. Long-read genome sequencing identifies causal structural variation in a Mendelian disease. Genet Med. 2018;20(1):159–63.
85. Roberts RJ, Carneiro MO, Schatz MC. The advantages of SMRT sequencing. Genome Biol. 2013;14(6):405.
86. Jain M, Olsen HE, Paten B, Akeson M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biol. 2016;17(1):239.
87. Rang FJ, Kloosterman WP, de Ridder J. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. Genome Biol. 2018;19(1):90.
88. Yuan Y, Bayer PE, Batley J, Edwards D. Improvements in Genomic Technologies: Application to Crop Genomics. Trends Biotechnol. 2017;35(6):547–58.
89. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol. 2018;36(4):338–45.
90. Chinwalla AT, Cook LL, Delehaunty KD, Fewell GA, Fulton LA, Fulton RS, et al. Initial sequencing and comparative analysis of the mouse genome. Nature. 2002;420(6915):520–62.
91. Long AD, Baldwin-Brown J, Tao Y, Cook VJ, Balderrama-Gutierrez G, Corbett-Detig R, et al. The genome of Peromyscus leucopus, natural host for Lyme disease and other emerging infections. Sci Adv. 2019;5(7):eaaw6441.
92. Zheng DJ, Okobi DE, Shu R, Agrawal R, Smith SK, Long MA, et al. Mapping the vocal circuitry of Alston's singing mouse with pseudorabies virus. bioRxiv. 2021 Jul 17;2021.07.16.452718.
93. Li R, Fan W, Tian G, Zhu H, He L, Cai J, et al. The sequence and de novo assembly of the giant panda genome. Nature. 2010;463(7279):311–7.
94. Tamazian G, Simonov S, Dobrynin P, Makunin A, Logachev A, Komissarov A, et al. Annotated features of domestic cat – Felis catus genome. GigaScience. 2014;3(1). Available from: https://doi.org/10.1186/2047-217X-3-13. [Cited 2021 Sep 1].
95. Geppert M, Goda Y, Hammer RE, Li C, Rosahl TW, Stevens CF, et al. Synaptotagmin I: A major Ca2+ sensor for transmitter release at a central synapse. Cell. 1994;79(4):717–27.
96. Maximov A, Südhof TC. Autonomous Function of Synaptotagmin 1 in Triggering Synchronous Release Independent of Asynchronous Release. Neuron. 2005;48(4):547–54.
97. Yoshihara M, Littleton JT. Synaptotagmin I Functions as a Calcium Sensor to Synchronize Neurotransmitter Release. Neuron. 2002;36(5):897–908.
98. Erlander MG, Tillakaratne NJ, Feldblum S, Patel N, Tobin AJ. Two genes encode distinct glutamate decarboxylases. Neuron. 1991;7(1):91–100.
99. Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-Wide Mapping of in Vivo Protein-DNA Interactions. Science. 2007;316(5830):1497–502.
100. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods. 2013;10(12):1213–8.
101. Booker TR, Jackson BC, Keightley PD. Detecting positive selection in the genome. BMC Biol. 2017;15(1):98.
102. Fu W, Akey JM. Selection and adaptation in the human genome. Annu Rev Genomics Hum Genet. 2013;14(1):467–89.
103. Pavlidis P, Alachiotis N. A survey of methods and tools to detect recent and strong positive selection. J Biol Res-Thessalon. 2017;24(1):7.

Publisher's Note