Title

NASA GeneLab RNA-Seq Consensus Pipeline: Standardized Processing of Short-Read RNA-Seq Data

Authors/Affiliations

Eliah G. Overbey[‡,1], Amanda M. Saravia-Butler[‡,2,3], Zhe Zhang[4], Komal S. Rathi[4], Homer Fogle[5,3], Willian A. da Silveira[6], Richard J. Barker[7], Joseph J. Bass[8], Afshin Beheshti[37,38], Daniel C. Berrios[39], Elizabeth A. Blaber[9], Egle Cekanaviciute[3], Helio A. Costa[10], Laurence B. Davin[11], Kathleen M. Fisch[12], Samrawit G. Gebre[3,37], Matthew Geniza[13], Rachel Gilbert[14], Simon Gilroy[7], Gary Hardiman[6,15], Raúl Herranz[16], Yared H. Kidane[17], Colin P.S. Kruse[18], Michael D. Lee[19,20], Ted Liefeld[21], Norman G. Lewis[11], J. Tyson McDonald[22], Robert Meller[23], Tejaswini Mishra[24], Imara Y. Perera[25], Shayoni Ray[26], Sigrid S. Reinsch[3], Sara Brin Rosenthal[12], Michael Strong[27], Nathaniel J Szewczyk[28], Candice G.T. Tahimic[29], Deanne M. Taylor[30], Joshua P. Vandenbrink[31], Alicia Villacampa[16], Silvio Weging[32], Chris Wolverton[33], Sarah E. Wyatt[34,35], Luis Zea[36], Sylvain V. Costes[*,3], Jonathan M. Galazka[*,3]

1: Department of Genome Sciences, University of Washington, Seattle, WA, 98195, USA
2: Logyx, LLC, Mountain View, CA 94043, USA
3: Space Biosciences Division, NASA Ames Research Center, Moffett Field, CA 94035, USA
4: Department of Biomedical and Health Informatics, The Children's Hospital of Philadelphia, University of Pennsylvania, Philadelphia, 19104, USA
5: The Bionetics Corporation, NASA Ames Research Center; Moffett Field, CA, 94035, USA
6: Institute for Global Food Security (IGFS) & School of Biological Sciences Queen's University Belfast, UK
7: Department of Botany, University of Wisconsin, Madison, WI, 53706, USA
8: MRC Versus Arthritis Centre for Musculoskeletal Ageing Research, Royal Derby Hospital, University of Nottingham & National Institute for Health Research Nottingham Biomedical Research Centre, Derby, DE22 3DT, UK

9: Center for Biotechnology and Interdisciplinary Studies, Department of Biomedical Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180, USA

10: Departments of Pathology, and of Biomedical Data Science, Stanford University School of Medicine, Stanford, CA, 94305, USA

11: Institute of Biological Chemistry, Washington State University, Pullman, WA, 99164, USA

12: Center for Computational Biology & Bioinformatics, Department of Medicine, University of California, San Diego, La Jolla, CA, 92037, USA

13: Phylos Bioscience, Portland, OR, 97214, USA

14: NASA Postdoctoral Program, Universities Space Research Association, NASA Ames Research Center, Moffett Field, CA, 94035, USA

15: Medical University of South Carolina, Charleston, SC, USA

16: Centro de Investigaciones Biológicas Margarita Salas (CSIC), Ramiro de Maeztu 9, 28040, Madrid, Spain

17: Center for Pediatric Bone Biology and Translational Research, Texas Scottish Rite Hospital for Children, 2222 Welborn St., Dallas, TX, 75219, USA

18: Los Alamos National Laboratory, Bioscience Division, Los Alamos, NM, 87545, USA

19: Exobiology Branch, NASA Ames Research Center, Mountain View, CA, 94035, USA

20: Blue Marble Space Institute of Science, Seattle, WA, 98154, USA

21: Department of Medicine, University of California San Diego, San Diego, CA, USA, 92093

22: Department of Radiation Medicine, Georgetown University Medical Center, Washington, DC, 20007, USA

23: Department of Neurobiology and Pharmacology, Morehouse School of Medicine, Atlanta, GA, 30310, USA

24: Department of Genetics, Stanford University School of Medicine, Stanford, CA, 94305, USA

25: Department of Plant and Microbial Biology, North Carolina State University, Raleigh NC, 27695

26: NGM Biopharmaceuticals, South San Francisco, CA, 94080, USA

27: National Jewish Health, Center for Genes, Environment, and Health, 1400 Jackson Street, Denver, CO, 80206, USA

28: Ohio Musculoskeletal and Neurological Institute and Department of Biomedical Sciences, Ohio University, Athens, OH, 43147, USA

29: Department of Biology, University of North Florida, Jacksonville, Florida, 32224, USA
30: Department of Biomedical and Health Informatics, Children's Hospital of Philadelphia and the Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, 19104, USA
31: Department of Biology, Louisiana Tech University, Ruston, LA, 71272, USA
32: Institute of Computer Science, Martin-Luther University Halle-Wittenberg, Von-Seckendorff-Platz 1, Halle, 06120, Germany
33: Department of Botany and Microbiology, Ohio Wesleyan University, Delaware, OH, USA
34: Department of Environmental and Plant Biology, Ohio University, Athens, OH, 45701, USA
35: Interdisciplinary Program in Molecular and Cellular Biology, Ohio University, Athens, OH, 45701, USA
36: BioServe Space Technologies, Aerospace Engineering Sciences Department, University of Colorado Boulder. 80303 USA
37: KBR, NASA Ames Research Center; Moffett Field, CA, 94035, USA
38: Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, 02142, USA
39: USRA/NASA Ames Research Center, Moffett Field, CA, 94035, USA

‡Co-first authors
*Corresponding authors
*Correspondence: sylvain.v.costes@nasa.gov, jonathan.m.galazka@nasa.gov

Summary

With the development of transcriptomic technologies, we are able to quantify precise changes in gene expression profiles from astronauts and other organisms exposed to spaceflight. Members of NASA GeneLab and GeneLab-associated analysis working groups (AWGs) have developed a

consensus pipeline for analyzing short-read RNA-sequencing data from spaceflight-associated experiments. The pipeline includes quality control, read trimming, mapping, and gene quantification steps, culminating in the detection of differentially expressed genes. This data analysis pipeline and the results of its execution using data submitted to GeneLab are now all publicly available through the GeneLab database. We present here the full details and rationale for the construction of this pipeline in order to promote transparency, reproducibility and reusability of pipeline data, to provide a template for data processing of future spaceflight-relevant datasets, and to encourage cross-analysis of data from other databases with the data available in GeneLab.

Introduction

Opportunities to perform biological studies in space are rare due to high costs and a limited number of funding sources, rocket launches, and spaceflight crew hours for experimental procedures. Additionally, spaceflight research is decentralized and distributed across numerous laboratories in the United States and abroad. As a result, studies performed in different laboratories often utilize different organisms, strains, cell lines, and experimental procedures. Adding to this complexity are variance in spaceflight factors and/or confounders within each study, such as degree of radiation exposure, experiment duration, CO_2 concentration, light cycle, and water availability, all of which can have effects on an organism's health and gene expression profiles during spaceflight (Rutter et al. n.d.). In order to optimize the integration of data from this diverse array of spaceflight experiments, it is paramount that variations in data processing are minimized.

There is presently no consensus on how best to analyze RNA-seq data and the impact of analysis tool selection on results is an active field of research. Indeed, selections of trimming parameters (Williams et al. 2016), read aligner (Yang et al. 2015), quantification tool (Teng et al. 2016), and differential expression detection algorithm (Costa-Silva, Domingues, and Lopes 2017) all affect results. Because of such challenges, groups like ENCODE and MINSEQE have developed standardized analysis pipelines for better comparison of RNA-seq datasets (ENCODE Project Consortium et al. 2020; "FGED: MINSEQE" n.d.).

The NASA GeneLab database (https://genelab-data.ndc.nasa.gov/genelab/projects) was created as a central repository for spaceflight-related omics-data. The repository includes data from experiments that profile transcription (RNA-seq, microarray), DNA/RNA methylation, protein

115 expression, metabolite pools, and metagenomes. The most prevalent data type in this repository is
116 RNA-seq from organisms exposed to spaceflight conditions. As of August 2020, the NASA
117 GeneLab database has over eighty datasets with RNA-sequencing data [Table S1]. These datasets
118 include *Homo sapiens* (human), *Mus musculus* (mouse), *Drosophila melanogaster* (fruit fly),
119 *Arabidopsis thaliana* (model higher plant), *Oryzias latipes* (Japanese rice fish), *Helix lucorum*
120 (land snail), *Brassica rapa* (Fast Plant®), *Eruca vesicaria* (arugula/edible plant), *Euprymna*
121 *scolopes* (Hawaiian bobtail squid), *Ceratopteris richardii* (aquatic fern), and the bacterium,
122 *Bacillus subtilis* from experiments performed during true spaceflight on various orbital platforms
123 such as the Space Shuttle and International Space Station (ISS), as well as spaceflight-analog
124 studies, such as hindlimb unloading and bed rest studies (Berrios et al., n.d.).
125 NASA's GeneLab and Ames Life Sciences Data Archive (ALSDA) projects have put
126 forward an ambitious strategy focused on integrating data, metadata, and biospecimens to fully
127 utilize the 40+ years of archived NASA Life Sciences data (Scott et al. 2020). One of the first steps
128 in this effort is the ability to analyze how experimental factors common to multiple datasets impact
129 molecular signaling. Such meta-analysis can only occur if metadata, data, and processed data are
130 harmonized. As part of this strategy, GeneLab engaged with the scientific community and held its
131 first Analysis Working Group (AWG) workshop in 2018. Spaceflight researchers from universities
132 and organizations across the United States and abroad met to begin the creation of a standardized,
133 consensus data-processing pipeline for one of the most common types of spaceflight datasets:
134 transcription profiling via RNA-sequencing. Scientists at this workshop met to discuss the merits
135 of various bioinformatic software tools for processing RNA-sequencing data, and ultimately
136 agreed on a single pipeline of these tools.
137 The main driver for developing the consensus pipeline was to present consistently
138 processed data to the public, therefore making space-relevant multi-omics data more accessible
139 and reusable. The overall goals were: 1) To get more consistently processed data to the public; 2)
140 To provide output data from every step of the consensus pipeline so users can download and use
141 these "intermediate" data; 3) To support easier and more consistent analysis of space-relevant data
142 by users including those in the NASA AWGs; and 4) To allow easier cross-analysis of experiments
143 to identify effects that result from the spaceflight environment, independent of confounding
144 factors. In addition, many of these data in the GeneLab database have not been previously
145 analyzed, as their generation was relatively recent. Therefore, providing new and processed

datasets to the public allows biologists and others to more easily interpret these data, and contributes significantly to our collective knowledge of the effects of spaceflight on terrestrial organisms.

Here we present the RNA-seq consensus pipeline (RCP) developed by the GeneLab AWG along with the rationale behind the tool settings and options selected. The RCP includes three distinct steps: data pre-processing, data processing, and differential gene expression computation/annotation [Fig 1A]. These steps use tools for quality control (FastQC, MultiQC) (Andrews and Others 2010; Ewels et al. 2016), read trimming (TrimGalore) (Krueger 2019), mapping (STAR) (Dobin et al. 2013), quantification (RSEM) (B. Li and Dewey 2011), and differential gene expression calculation/annotation (DESeq2) (Love, Huber, and Anders 2014) [Fig 1B]. The RCP has been integrated into the GeneLab database and files produced by the RCP for each RNA-seq dataset hosted in GeneLab are and will continue to be publicly available for download.

Results

Data Pre-processing: Quality Control and Trimming

There are three distinct steps to the RCP, the first of which is data preprocessing [Fig 2A]. The pipeline begins with quality control (QC) of raw FASTQ files from a short-read Illumina sequencer using the FastQC software (Andrews and Others 2010) [Fig 2B]. FastQC is one of the most widely used QC programs for short-read sequencing data. It provides information which can be used to assess sample and sequencing quality, including base statistics, per base sequencing quality, per sequence quality scores, per base sequence content, per base GC content, per sequence GC content, per base N content, sequence length distributions, sequence duplication levels, overrepresented sequences and k-mer content.

The FastQC program is run on each individual sample file. However, reviewing the FastQC results for each sample file can be tedious and time consuming. Experiments typically have many sample files (biological and/or technical replicates) for multiple experimental conditions (spaceflight, ground control, etc). For this reason, we also use the MultiQC package (Ewels et al. 2016) [Fig 2C] to create a summary statistics report that includes the same quality control result categories from FastQC across all experiment samples.

After performing quality control on the raw FASTQ data, reads are trimmed using TrimGalore (Krueger 2019) to remove sequencing adapters that would disrupt read mapping during the data processing pipeline step [Fig 2D]. Quality trimming is not performed as this has been shown to decrease the accuracy of quantification results (Williams et al. 2016). TrimGalore is a wrapper program that uses the cutadapt program (Martin 2011) for read trimming. TrimGalore was selected for the RCP due to its simplified command line interface, thorough output of trimming metrics, and ability to automatically detect adapters. In this step, bases that are part of a sequencing adapter are removed from each read and reads that become too short will subsequently be removed. After trimming, the quality control programs, FastQC and MultiQC, are again run on the trimmed FASTQ files for viewing the quality control metrics of the reads that will be used for data processing. Once the data has been preprocessed, the sequenced reads are ready for mapping and quantification.

Data Processing: Read Mapping and Sample Quantification

In the data processing step [Fig 1; Step 2A], the trimmed reads are first aligned to the reference genome [Fig 3A] with STAR, a splice-aware aligner (Dobin et al. 2013). STAR must be run in two steps. The first step is to create indexed genome files [Fig 3B]. These files are used to assist read mapping and only need to be generated once for each reference genome file. This step requires reference FASTA and GTF files [Table S2]. Some datasets include the External RNA Control Consortium (ERCC) spike-in control - a pool of 96 synthetic RNAs with various lengths and GC content covering a 2^{20} concentration range (Jiang et al. 2011). If ERCC spike-ins were included, the spike-in FASTA and GTF files are appended to the reference FASTA and GTF files, respectively. The second step of STAR mapping is to use the indexed reference genome and the trimmed reads from the preprocessing step in order to map the reads to the genome and the transcriptome [Fig 3C]. STAR will also produce genome mapped data, which can optionally be used to find reads that map outside of annotated reference transcripts. STAR mapping output data are in BAM format, which has a separate entry for each mapped read and states which transcript each read mapped to. In order to improve the detection and quantification of splice sites, STAR is run in “two-pass mode”. Here, splice sites are detected in the initial mapping to the reference and used to build a new reference that includes these splice sites. Reads are then re-mapped to this

dynamically generated reference to improve the quantification of splice isoforms (Dobin et al. 2013). Users are provided with these results (as per sample SJ.out files) for further analysis of differential splicing.

The second part of processing is quantifying the number of reads mapped to each annotated transcript and gene [Fig 1A; Step 2B]. For this task, the RCP uses RSEM (B. Li and Dewey 2011) [Fig. 4A]. The main reasons for using RSEM are its ability to account for reads that map to multiple transcripts and distinguish gene isoforms. In short-read sequencing experiments it is likely that some number of reads will map to multiple regions in the genome. RSEM computes maximum likelihood abundance estimates to split the read count across multiple genes. Similar to STAR, RSEM is run in two distinct phases. The first phase uses the reference genome and GTF files (with or without ERCC as appropriate) [Table S2] to prepare indexed genome files [Fig 4B]. The second phase uses the indexed files and the mapped reads from STAR to assign counts to each gene [Fig 4C]. There are two output files generated for each sample: counts assigned to genes and counts assigned to isoforms. Gene counts are used to calculate differential gene expression. Isoform counts are also generated as an option to look at differential isoform expression but are not used during differential gene expression calculation in the RCP. Once the RSEM count files are generated, the data are used to compute differentially expressed genes. A list of the reference genomes used in the GeneLab pipeline is available in Supplementary Table 2 [Table S2]. These reference genomes were the most recent releases at the time each STAR and RSEM indexed references were created. While it is possible to run STAR mapping through the RSEM toolkit, we elected not to do this because the alignment parameters used in this case are from ENCODE's STAR-RSEM pipeline and are not customizable. Thus, we would have been precluded from using the precise mapping parameters agreed to by the GeneLab AWG.

We elected to adopt a mapping-based approach rather than rapidly quantifying the reads via a k-mer-based counting algorithm, pseudo-aligners, or a quasi-mapping method that utilizes RNA-seq inference procedures such as Kallisto (Bray et al. 2016) or Salmon (Patro et al. 2017) despite their speed advantages. This is because alignment-free quantification tools do not accurately quantify low-abundant and small RNAs especially when biological variation is present (Wu et al. 2018). Furthermore, alignment of reads allows for additional analyses beyond transcript and gene quantification such as measurement of gene body coverage and detection of novel transcripts.

There are several alignment-based mapping tools available and each has advantages and disadvantages. An alignment tool that is sensitive to splice-isoforms is critical to accurately identify how expression of splice-isoforms is affected by the spaceflight environment. DNA-specific aligners such as BWA (H. Li and Durbin 2009) and Bowtie (Langmead et al. 2009) cannot handle intron-sized gaps and thus an RNA-seq specific aligner is needed (Baruzzo et al. 2017). In addition to splice-awareness, when selecting an aligner the following criteria were also considered: ability to input both single- and paired-end reads, handle strand-specific data, applicability to a variety of different model organisms with both low- and high-complexity genomic regions, efficient runtime and memory usage, ability to identify chimeric reads, high sensitivity, low rate of false discovery, and ability to output both genome and transcriptome alignments. Several studies have been conducted to compare the wide variety of available RNA-seq specific alignment tools, and of these, the STAR aligner consistently performs better than, or on par with the tools tested for the indicated criteria (Baruzzo et al. 2017; Schaarschmidt et al. 2020; Raplee, Evsikov, and Marín de Evsikova 2019).

Differential Gene Expression Calculations and Addition of Gene Annotations

Once reads have been mapped and quantified, differential expression analysis is performed using the DESeq2 R package [Fig 1; Step 3, Fig 5A]. Unlike the previous steps, a custom R script GeneLab_DGE_noERCC.R or (GeneLab_DGE_wERCC.R) [Scripts S1, S2] is used to run DESeq2, to create both unnormalized and normalized counts tables, and to generate a differential gene expression (DGE) output table containing normalized counts for each sample, DGE results, and gene annotations [Fig 5B]. The GeneLab DGE R script also creates computer-readable tables that are used by the GeneLab visualization portal to generate various plots so users can easily view and begin interpreting the processed data. These scripts are provided in the NASA GeneLab_Data_Processing Github repository (https://github.com/nasa/GeneLab_Data_Processing). In the following sections we describe each step of these sections in order.

The GeneLab DGE R script requires three inputs: the quantified counts data from the previous (RSEM) step, sample metadata from the Investigation, Study, and Assay (ISA) tables in the ISA.zip file (provided in the GeneLab repository with each dataset) (Sansone et al. 2012; Rocca-Serra et al. 2010), and the organisms.csv file [Table S3], which is used to specify the

organism used in the study and relevant gene annotations to load. Since samples from some GeneLab RNA-seq datasets contain ERCC spike-in and others do not, there are two versions of the GeneLab DGE R script, one for datasets with ERCC spike-in (GeneLab_DGE_wERCC.R, Script S1) and one for those without (GeneLab_DGE_noERCC.R, Script S2). Prior to running either script, paths to directories containing the input data and the output data location must be defined. Each script starts by defining the organism used in the study, which should be consistent with the name in the organisms.csv file so that it matches the abbreviations used in the PANTHER database (Mi, Muruganujan, and Thomas 2013; Thomas 2003) for each organism. Next, the metadata from the ISA.zip file are imported and formatted for use with the DESeq2 package. During metadata formatting, groups for comparison are defined based on experimental factors and a sample table is created to specify the group to which each sample belongs. Next, a contrasts matrix is generated, which specifies the groups that will be compared during DGE analysis; each group is compared with every other group in a pairwise manner in both directions (i.e. spaceflight vs. ground and ground vs. spaceflight). This approach provides the user with the results for all possible group comparisons, allowing each user to select the most relevant comparisons for their particular scientific questions. After metadata formatting, the RSEM gene count data files from each sample are listed and re-ordered (to match the order the samples appear in the metadata), then imported with the R package, tximport (Soneson, Love, and Robinson, n.d.), and sample names are assigned. Prior to running DESeq2, a value of 1 is added to genes with lengths of zero, which is necessary to make a DESeqDataSet object. A DESeqDataSet object is then created using the formatted metadata and the count data that was imported with tximport.

For datasets that contain samples without ERCC spike-in, we use the GeneLab_DGE_noERCC.R script [Script S1]. To reduce the possibility of skewing the data during DESeq2 normalization (McIntyre et al. 2011; Risso et al. 2011; Conesa et al. 2016; Law et al. 2016), all genes that have a sum of less than 10 counts across all samples are removed. The cutoff value of 10 is a best practice recommended by the DESeq2 tutorial on Bioconductor. These filtered data are then prepared for normalization and DGE analysis with DESeq2. Since there is no consensus for whether or not ERCC-normalization improves the accuracy of the results (Risso et al. 2014), the GeneLab project and its AWG members decided to perform the DGE analysis both with and without ERCC-normalization (for datasets with samples containing ERCC spike-in).

298 To enable DESeq2 analysis with and without considering ERCC reads, the DESeqDataSet
299 object is used to create a DESeqDataSet object containing only ERCC reads. Since all samples
300 must contain ERCC spike-in for ERCC-normalization, the DESeqDataSet object containing only
301 ERCC reads is used to identify and remove any samples that do not contain ERCC reads. Next, a
302 DESeqDataSet object containing only non-ERCC reads is created by removing rows containing
303 ERCC reads. These data are then used for DESeq2 analysis.
304 For DESeq2 analysis with ERCC-normalization (Script S2), the size factor object of the
305 non-ERCC data is replaced with ERCC size factors for re-scaling in the first DESeq2 step. For
306 DESeq2 analysis without ERCC-normalization, the DESeq2 default algorithm is applied to the
307 DESeqDataSet object containing only non-ERCC reads. The unnormalized and DESeq2-
308 normalized counts data as well as the sample table are then outputted as CSV files. The
309 'Unnormalized_Counts.csv', 'Normalized_Counts.csv', and 'ERCC_Normalized_Counts.csv'
310 files for each RNA-seq dataset are available in the GeneLab Data Repository; the
311 'SampleTable.csv' file is used internally for verifying and validating the processed data prior to
312 publication.
313 There are two types of hypothesis tests that can be run with DESeq2, the likelihood ratio
314 test (LRT), which is similar to an analysis of variance (ANOVA) calculation in linear regression
315 and allows for comparison across all groups, and the Wald test, in which the estimated standard
316 error of a log2 fold change is used to compare differences between two groups. The DGE step of
317 the RCP performs both of these analyses. After normalization, the DESeq2 likelihood ratio test
318 design is applied to the normalized data (both ERCC- and nonERCC-normalized data) to generate
319 the F statistic p-value, which is similar to an ANOVA p-value and reveals genes that are changed
320 in any number of combinations of all factors defined in the experiment.
321 To prepare for building a gene/pathway annotation database, the STRINGdb (Szklarczyk
322 et al. 2019) and PANTHER.db (Thomas 2003) libraries are loaded and the organisms.csv file is
323 read and used to indicate the Bioconductor AnnotationData Package needed (Huber et al. 2015;
324 Gentleman et al. 2004). The current gene annotation database for the organism specified at the
325 beginning of the R script is then loaded.
326 Next, DGE tables containing normalized counts for each sample, pairwise DGE results,
327 and current gene annotations as well as computer-readable DGE tables (that will be used for
328 visualization) are created first with nonERCC-normalized data and then with ERCC-normalized

data. For pairwise DGE analysis, first normalized count data are used to create two output tables, one that is used to create the human-readable DGE output table provided to users with processed data for each dataset, and the respective computer-readable DGE output table that contains additional columns and is used to visualize the data. Next, normalized count data are iterated through Wald Tests to generate pairwise comparisons of all groups based on the contrasts matrix that was generated during metadata formatting. The pairwise DGE analysis results are then added as columns to both DGE output tables.

Then an annotation database is built by first defining the "keytype", which indicates the primary type of annotation used (for most GeneLab datasets this is ENSEMBL). The keytype is then used to map to annotations in the organism-specific Bioconductor AnnotationData Package, and the following annotation columns are added to the annotation database: SYMBOL, GENENAME, ENSEMBL (if not the primary), REFSEQ, and ENTREZID. STRING and GOSLIM annotation columns are also added to the annotation database using the STRINGdb and PANTHER.db R packages, respectively. All of the aforementioned annotation columns are added to the annotation database to enable users to perform downstream analyses without having to map gene IDs themselves. Once the annotation database is complete, additional calculations are performed on the normalized count data before assembling the final DGE output tables.

Means and standard deviations of normalized count data for each gene across all samples, and for samples within each respective group, are calculated and added as columns to the DGE output tables. A column containing the F statistic p-value, calculated previously, is also added to the DGE output tables. The following columns are added only to the computer-readable DGE output table (used for visualization): a column to indicate whether each gene (or pathway) is up- or down-regulated for each pairwise comparison, a column to indicate genes that are differentially expressed using a p-value cutoff of ≤ 0.1 and another column using a p-value cutoff of ≤ 0.05, a column indicating the log2 of the p-value for each pairwise comparison and another column indicating the log2 of the adjusted p-value, both of which are used to create Volcano plots. After all columns are added to the DGE tables, both the human- and computer-readable DGE tables are combined with the current annotation database to create the complete human- and computer-readable DGE tables. An example of the complete human readable DGE tables provided with processed RNAseq datasets in the GeneLab Data Repository is shown in Table 1 and Table 2. Principal component analysis (PCA) is also performed on the normalized count data and used to

create PCA plots for the GeneLab data visualization portal. DGE analysis of datasets without ERCC spike-in is performed exactly the same way as the nonERCC-normalized approach described above, except that no ERCC reads have to be removed from the DESeqDataSet object prior to DESeq analysis.

Both the GeneLab_DGE_wERCC.R and the GeneLab_DGE_noERCC.R scripts produce the following output files: Unnormalized_Counts.csv (*), Normalized_Counts.csv (*), SampleTable.csv (#), contrasts.csv (*), differential_expression.csv (*), visualization_output_table.csv (**), visualization_PCA_table.csv (**) [Fig 5B, Table 1, Table 2]. The GeneLab_DGE_wERCC.R script will also produce the following additional output files: ERCC_rawCounts_unfiltered.csv (#), ERCC_rawCounts_filtered.csv (#), ERCCnorm_contrasts.csv (*), ERCC_Normalized_Counts.csv (*), ERCCnorm_differential_expression.csv (*), visualization_output_table_ERCCnorm.csv (**), visualization_PCA_table_ERCCnorm.csv (**) [Fig 5B, Table 1, Table 2]. All the tools used in the consensus pipeline described above are documented in Supplemental Table 4: Pipeline Tools and Links [Table S4].

A Use Case for Data Processed with the RCP

To showcase the value of using a consensus pipeline and publishing the processed data from each step of the pipeline, downstream analyses were performed using processed data from select samples from RNAseq datasets hosted on GeneLab. One of the advantages of providing expression data of all samples in each dataset as well as all possible pairwise DGE comparisons is to allow users the flexibility to pick and choose which samples and which comparisons they would like to focus on. Thus, when selecting samples for downstream analysis, we exercised this flexibility and searched the GeneLab Data Repository for datasets/samples that met a specific set of criteria. These criteria were as follows: 1) datasets that evaluated the same tissue (liver) from the same mouse strain (C57BL/6) and sex (female), 2) only samples derived from animals flown in space and respective ground control samples, 3) studies that used the same preservation protocol (liver samples extracted from frozen carcasses post-mission) and library preparation method (ribo-depletion), and 4) samples that contained ERCC spike-in to evaluate outputs with and without ERCC normalization. Select samples from two GeneLab datasets, GLDS-168 and GLDS-245 met these criteria and processed data including the Normalized_Counts.csv,

differential_expression.csv, ERCC_Normalized_Counts.csv, and the ERCCnorm_differential_expression.csv files from these two datasets were used for downstream analyses.

Prior to downstream analysis, the processed data files were filtered so that only samples that met the criteria listed above were included. Since GLDS-168 contains samples from both the Rodent Research 1 (RR-1) and RR-3 missions and only the RR-1 mission met our first criteria of using the C57BL/6 mouse strain, RR-3 samples were removed from the process data files. GLDS-168 processed data files were subsequently filtered to remove all samples except spaceflight (FLT) and respective ground control (GC) samples to meet the second criteria listed above. Lastly, since GLDS-168 contains a set of FLT and GC samples that were spiked with ERCC and another set in which ERCC was not added, the later set of samples were removed to meet the fourth criteria. GLDS-245 contains liver samples from the RR-6 mission, which included a set of animals that were returned to earth alive after ~30 days of spaceflight and another set of animals that remained in space (aboard the ISS) for a total of ~60 days before being sacrificed aboard the ISS (note that there were respective control samples for each set of spaceflight animals described). The former set of animals had their livers dissected immediately after euthansia whereas livers from the latter set of animals were frozen in situ and dissected from frozen carcasses after return to earth. Thus, only the later (ISS-terminal) set of FLT and respective GC samples met criteria 2 and 3, so the GLDS-245 processed data files were filtered to remove all other samples. A complete list of all samples from GLDS-168 and GLDS-245 that were included in this analysis are provided in Supplemental Table S5 [Table S5]. Additionally, since the downstream analyses focused on the differences between FLT and GC samples in these two datasets, all other comparisons were removed from the differential_expression.csv and ERCCnorm_differential_expression.csv files prior to analysis.

The filtered processed data files (available in Mendeley Data, http://dx.doi.org/10.17632/fv3kd6h7k4.1) were then used to create Principal Component Analysis (PCA) plots [Fig 6A, 6B and Fig S1A, S1B], heatmaps containing the top 30 most significant FLT vs. GC differentially expressed (and annotated) genes (adj. p value < 0.05 and |log2FC| > 1) [Fig 6C, 6D and Fig S1C, S1D], and to evaluate FLT vs. GC gene ontology (GO) differences using Gene Set Enrichment (GSEA) analysis [Table 3, Table S6]. These results can then be further evaluated to identify similarities and differences in gene expression between these two studies and

422 draw novel conclusions about the effects of spaceflight that are consistent across spaceflight
423 experiments.

424 ***Discussion***

425 The differentially expressed genes calculated by the RCP can be further explored with a
426 variety of tools designed for higher-order analysis. For example, there are tools which can look for
427 enriched pathways, gene ontology terms, or protein and/or metabolite networks. Popular software
428 tools among the GeneLab working group members include WebGestalt (Liao et al. 2019),
429 STRING (Szklarczyk et al. 2019), GSEA (Subramanian et al. 2005), PIANO (Väremo, Nielsen,
430 and Nookaew 2013), Reactome (Szklarczyk et al. 2019), and ToppFun (Chen et al. 2009). There
431 is no universal consensus on which tools are the most useful for higher-order analysis (Nguyen et
432 al. 2019). RCP users are encouraged to try multiple tools in order to analyze their data from a
433 variety of perspectives.
434 The RCP has been designed to handle sequencing experiments that either lack or include
435 the ERCC RNA spike-in mix - a set of 96 polyadenylated RNAs that can be used during differential
436 gene expression calculation to normalize read counts across samples (Munro et al. 2014).
437 However, the use of normalization according to ERCC spike-ins remains controversial among
438 AWG members, and Munro *et al.* suggested its usage only for determining limit of detection of
439 ratio (LODR), expression ratio variability and measurement bias (Munro et al. 2014). For this
440 reason, ERCC normalization remains optional in the GeneLab pipeline and both kinds of DGE
441 outputs are provided in the GeneLab database. Additionally, ERCC spike-in could have two other
442 usages. First, it allows us to evaluate whether normalization succeeded in removing systemic bias
443 between libraries by using methods such as Rlog and VST when normalizing the spike-in RNAs
444 along with all other genes. Second, most normalization methods of RNA-seq data assume that
445 most genes are not differentially expressed towards one direction. Comparing spike-in
446 measurements between libraries will help us to estimate the validity of this assumption.
447 A high number of biological replicates can increase certainty in the differentially
448 expressed genes determined by the RCP. However, conducting experiments in spaceflight often
449 limits the number of biological replicates that a researcher can include. Therefore, it is important
450 to note that at least three biological replicates are required for the pipeline, specifically for DESeq2,

to perform its statistical methods. However, at least six replicates are suggested in order to minimize the false discovery rate (FDR) (Schurch et al. 2016). Finally, RNA-seq datasets hosted on GeneLab that do not contain biological replicates are only processed up until unnormalized (raw) counts are obtained, the step right before DESeq2 is used for DGE calculation.

More advanced RCP users might have additional data inquiries that fall beyond the scope of this pipeline. For this reason, there are two parts of the pipeline that include additional output that are not used in our differential gene expression computation. The first is in the output from STAR, mapping output is also provided in genomic coordinates. This is useful for obtaining reads that are mapped outside of the reference transcriptome. For example, this may be used to find novel genes, transcripts, or exons that have not yet been annotated by consortiums. The second part of the pipeline with alternative output files is RSEM. This also provides transcript-level counts which can be used to investigate differential isoform expression. Moreover, intermediate files are provided as outputs to allow users to use components of the pipeline that they find useful.

The GeneLab database also includes other types of transcriptomic data. As discussed in this article, the RCP is not used for microarray data which are fundamentally different, and the AWG is still debating the best approach for cross-dataset comparisons between microarrays. GeneLab also accepts data from long read experiments, such as those produced by Pacific Biosciences' (PacBio) single-molecule real-time (SMRT) sequencing (Roberts, Carneiro, and Schatz 2013) and Oxford Nanopore Technologies' (ONT) nanopore sequencing (Jain et al. 2016). Long-read data would be processed with similar steps to the RCP but will require tools specifically designed for the intricacies of long-read data, such as reads that contain multiple splice junctions and reads which currently have a higher base-calling error-rate. Currently, long-reads are typically used for DNA sequencing and were recently highlighted on board of the ISS using ONT for de novo assembly of the *E. coli* genome from raw reads (Castro-Wallace et al. 2017). However, even though throughput and accuracy remain far inferior to short-reads, long-reads offer some advantages for RNAseq as well, with less ambiguity for genes and isoforms detection, much faster mapping, potential identification of genes not yet known from reference genomes and eventually less bias in DGE.

To conclude, the RCP is specifically designed for RNA-seq data from short-read sequencers and has been developed in order to encourage and facilitate analysis of spaceflight multi-omic data. The creation of the RCP by a large community of scientists (GeneLab AWG:

https://genelab.nasa.gov/awg) and the sharing of pipeline details in a peer-reviewed article provide analysis transparency and enable data reproducibility.

Limitations of the Study

The results of this study are limited to short-read RNA-seq and are not applicable to other transcriptomic profiling methods (e.g. microarray, long-read RNA-seq). Additionally, the pipeline cannot compensate for poor library preparation technique or inadequate sample size. Sample preservation protocols between datasets need to also be evaluated, since variations in sample preservation protocol could lead to poor correlation between studies that are otherwise identical (Polo et al. 2020). The number of sequenced reads may also be a limiting factor in the usefulness and accuracy of the differentially expressed genes calculated by DESeq2 and, similarly, during splice isoform analysis.

Note that this article does not discuss strategies and pipelines regarding older transcriptomics data in GeneLab (i.e. more than 100 microarray datasets), as it is much more challenging to provide meta-analysis with microarrays, which are prone to strong batch effects and gene lists which are platform dependent. Future efforts of GeneLab and the AWG will address microarray pipelines.

In the future, we will add functionality to process unique molecular identifiers (UMIs) that can identify PCR duplicates using tools such as UMI-tools (Smith, Heger, and Sudbery 2017). This will allow PCR duplicates to be removed after mapping and before quantification.

Additionally, transcriptomic data will be integrated with proteomic and metabolomics data; this will help further understand the significance of gene expression changes to metabolic "fitness" in the spaceflight environment.

Resource Availability

Lead Contact: Jonathan M. Galazka

Materials Availability: No unique reagents were generated in this study.

Data and Code Availability: Spaceflight-relevant RNA-seq data is located in the GeneLab database (https://genelab-data.ndc.nasa.gov/genelab/projects). All software packages are open source and are linked in the methods section. Custom R scripts for DESeq2 are included as supplemental information and are available in the Github repository GeneLab_Data_Processing (https://github.com/nasa/GeneLab_Data_Processing). Raw data utilized to generate plots are available on Mendeley Data (http://dx.doi.org/10.17632/fv3kd6h7k4.1).

Methods

The tools used in the consensus pipeline are documented in Supplemental Table 4: Pipeline Tools and Links [Table S4]. Due to NASA security requirements, all software is updated monthly with security patching. Therefore, tool versions used to process each RNA-seq dataset hosted on the GeneLab Data Repository are provided in the RNA-seq protocol section and are also available along with exact processing scripts in the GeneLab Data Processing GitHub Repository (https://github.com/nasa/GeneLab_Data_Processing/tree/master/RNAseq/GLDS_Processing_Scripts). Specific commands, options, and flags for each tool used in the RCP are reported in the figures of the main text. Note that some packages listed here are dependencies of the packages used in the RCP. More information about such dependencies can be found in the tool documentation.

This pipeline has been run on short-read RNA-seq data in the GeneLab database (https://genelab-data.ndc.nasa.gov/genelab/projects) and is applied to new submissions to the database. Any updates to the software used in the pipeline will be noted in the Github repository GeneLab_Data_Processing (https://github.com/nasa/GeneLab_Data_Processing).

Processed RNAseq data from GLDS-168 and GLDS-245 select samples were used to provide an example of the downstream analyses that can be done using data processed with the consensus pipeline presented here. Normalized counts and ERCC-normalized counts from the following GLDS-168 and GLDS-245 samples were used to generate the PCA plots shown in Figure 6A & 6B and Supplemental Figure 1A & 1B, respectively. Samples from GLDS-168 and GLDS-245 that were used in this study are listed in Supplemental Table 5 [Table S5]. Differential gene expression (DGE) data from FLT versus GC samples using (non-ERCC) normalized counts

and ERCC-normalized counts data for each respective dataset were used to generate the heatmaps shown in Figure 6C & 6D and Supplemental Figure 1C & 1D, respectively. DGE data were filtered using an adjusted p value cutoff of < 0.05 and |log2FC| cutoff of > 1. The gene expression data were then sorted based on adjusted p values and the top 30 most differentially expressed and annotated genes were used to generate heatmaps with ggplot2 version 3.3.2 (Wickham, Navarro, and Pedersen 2016). Note that for visualization purposes, sample names were shortened.

Pairwise gene set enrichment analysis (GSEA) was performed on the (non-ERCC) normalized counts (Table 3) and ERCC-normalized counts [Table S6] from select samples in GLDS-168 and GLDS-245 using the C5: Gene Ontology (GO) gene set (MSigDB v7.2) as described (Subramanian et al. 2005). All comparisons were performed using the phenotype permutation. The ranked lists of genes were defined by the signal-to-noise metric and the statistical significance were determined by 1000 permutations of the gene set. FDR <= 0.25 were considered significant for comparisons according to the authors' recommendation.

The data used to generate all PCA plots, heatmaps, and GSEA shown are provided on Mendeley (http://dx.doi.org/10.17632/fv3kd6h7k4.1).

Acknowledgments

This work was funded in part by the NASA Space Biology program within the NASA Science Mission Directorate's (SMD) Biological and Physical Sciences (BPS) Division, NASA award numbers NNX15AG56G, 80NSSC19K0132, the Biotechnology and Biological Sciences Research Council (grant number BB/ N015894/1), the MRC Versus Arthritis Centre for Musculoskeletal Ageing Research (grant numbers MR/P021220/1 and MR/R502364/1), the Spanish Research Agency (AEI grant number RTI2018-099309-B-I00, co-funded by EU-ERDF) and the National Institute for Health Research Nottingham Biomedical Research Centre. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

Author Contributions

All authors developed the initial analysis scheme at the 2019 GeneLab AWG workshop. EGO, ZZ, KSR, HF, WAdS, RB and JMG refined this into a draft pipeline. EGO and AMSB wrote and

567 validated the final processing scripts. EGO and AMSB wrote the original manuscript draft and
568 generated figures. All authors reviewed and edited the final draft.
569

570 ## ***Declaration of Interests***

571 The authors declare no competing interests.

572 ## ***Figure and Scheme Legends***

573 **Figure 1: GeneLab RNA-seq Consensus Pipeline (RCP). A:** The three broad steps of the RCP.
574 The RCP handles: 1) Data preprocessing to trim sequencing adapters and to provide quality control
575 metrics; 2) Data processing to map reads to the reference genome and quantify the number of read
576 counts per gene; and 3) Differential gene expression calculation, which will provide a list of
577 differentially expressed genes that can be sorted by adjusted p-value and log fold-change. **B:** The
578 full RCP annotated with tools, input files, and output files.
579

580 **Figure 2: Data preprocessing (pipeline step 1): Quality control and trimming. A:** Data
581 Preprocessing pipeline. FastQ files from Illumina base-calling software are quality checked using
582 FastQC and MultiQC. Data is then trimmed using TrimGalore and are re-checked for quality; **B:**
583 Flags used for FastQC program; **C:** Flags used for MultiQC program; **D:** Flags used for
584 TrimGalore program; trimmed reads (*fastq.gz) are then used as input data for FastQC (B)
585 followed by MultiQC (C) to generate trimmed read quality metrics. Tool versions used to process
586 each dataset are included in the RNA-seq processing protocol in the GLDS Repository.
587

588 **Figure 3: Data processing (pipeline step 2A): Read mapping. A:** Data processing pipeline.
589 Trimmed reads are mapped to their reference genome and transcriptome with STAR. Gene counts
590 are then quantified with RSEM; **B:** Flags used for generating the indexed STAR reference files;
591 **C:** Flags used for mapping reads with STAR. Tool versions used to process each dataset are
592 included in the RNA-seq processing protocol in the GLDS Repository.
593

Figure 4: Data processing (pipeline step 2B): Gene quantification. A: Data processing pipeline. Mapping results from STAR are quantified by RSEM; **B:** Parameters for RSEM indexed reference files generation; **C:** Parameters for quantifying gene and isoform counts with RSEM. Tool versions used to process each dataset are included in the RNA-seq processing protocol in the GLDS Repository.

Figure 5: Differential gene expression calculation (pipeline step 3). A: Data processing pipeline. The R program DESeq2 is run in order to determine which genes are differentially expressed between experimental conditions using gene count files from RSEM. **B:** Output files generated. The table columns distinguish which script produces each output. The columns distinguish how those output files are used.

Figure 6. Global and differential gene expression in spaceflight versus ground control liver samples from GeneLab datasets. A, B: Principal component analysis of global gene expression in spaceflight (FLT) and respective ground control (GC) liver samples from the A) Rodent Research 1 (RR-1) NASA Validation mission (GLDS-168) and B) RR-6 ISS-terminal mission (GLDS-245). Plots were generated using data in the normalized counts tables for each respective dataset on the NASA GeneLab Data Repository. **C, D:** Heatmaps showing the top 30 differentially expressed genes in spaceflight (FLT) versus ground control (GC) liver samples from the C) Rodent Research 1 (RR-1) NASA Validation mission (GLDS-168) and D) RR-6 ISS-terminal mission (GLDS-245). Heatmaps were generated using data in the differential expression tables for each respective dataset on the NASA GeneLab Data Repository and are colored by relative expression. Adj. p-value < 0.05 and |log2FC| > 1. All samples included were derived from frozen carcasses post-mission and utilized the ribo-depletion library preparation method.

Table 1. Differential gene expression output table - Annotations. Truncated version of the differential_expression.csv file provided as GeneLab processed data for GLDS-251. The first 7 columns of the differential gene expression output table contain gene IDs and annotations (for remainder of columns, refer to Table 2).

624 **Table 2. Differential gene expression output table - Statistics**. Truncated version of the
625 differential_expression.csv file provided as GeneLab processed data for GLDS-251. Following the
626 7 columns of gene IDs and annotations (Table 1) are normalized gene expression data for each
627 sample (**Norm. expr. (sample A)**) then results from all possible pairwise comparisons, including
628 log2 fold change (**Log2fc (comparison A)**), p values (**P.value (comparison A)**), and adjusted p
629 values (**Adj.p.value (comparison A)**) calculated from the Wald Tests. Next are the average gene
630 expression (**Mean (all samples)**) and standard deviation (**Stdev (all samples)**) of all samples
631 followed by the F-statistic p value generated from the likelihood ratio test (LRT.p.value), and the
632 last set of columns are the average gene expressions (Group.Mean) and standard deviations
633 (Group.Stdev) of samples within each group.
634
635 **Table 3. Comparison of gene ontology in spaceflight versus ground control liver samples**
636 **from GeneLab datasets.** The number of enriched gene ontology (GO) terms identified by Gene
637 Set Enrichment Analysis (GSEA, phenotype permutation) was evaluated in spaceflight (FLT)
638 versus ground control (GC) liver samples from the Rodent Research 1 (RR-1) NASA Validation
639 mission (GLDS-168), and RR-6 ISS-terminal mission (GLDS-245). For GO terms, the number on
640 the left corresponds to GO terms enriched in FLT samples and the number on the right corresponds
641 to GO terms enriched in GC samples. These data were generated using the normalized counts for
642 each respective dataset on the NASA GeneLab Data Repository. All samples included were
643 derived from frozen carcasses post-mission and utilized the ribo-depletion library preparation
644 method. GLDS-168, FLT n=5 and GC n=5; GLDS-245, FLT n=10 and GC n=10. p values and
645 FDR values are indicated.
646
647
648

References

Andrews, Simon, and Others. 2010. "FastQC: A Quality Control Tool for High Throughput Sequence Data." Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom.

Baruzzo, Giacomo, Katharina E. Hayer, Eun Ji Kim, Barbara Di Camillo, Garret A. FitzGerald, and Gregory R. Grant. 2017. "Simulation-Based Comprehensive Benchmarking of RNA-Seq Aligners." *Nature Methods* 14 (2): 135–39.

Berrios, Daniel, Jonathan Galazka, Kirill Gorev, Samrawit Gebre, and Sylvain Costes. n.d. "Interfaces for the Exploration of Space Omics Data." *Nucleic Acids Research*.

Bray, Nicolas L., Harold Pimentel, Páll Melsted, and Lior Pachter. 2016. "Near-Optimal Probabilistic RNA-Seq Quantification." *Nature Biotechnology* 34 (5): 525–27.

Castro-Wallace, Sarah L., Charles Y. Chiu, Kristen K. John, Sarah E. Stahl, Kathleen H. Rubins, Alexa B. R. McIntyre, Jason P. Dworkin, et al. 2017. "Nanopore DNA Sequencing and Genome Assembly on the International Space Station." *Scientific Reports* 7 (1): 18022.

Chen, Jing, Eric E. Bardes, Bruce J. Aronow, and Anil G. Jegga. 2009. "ToppGene Suite for Gene List Enrichment Analysis and Candidate Gene Prioritization." *Nucleic Acids Research* 37 (Web Server issue): W305–11.

Conesa, Ana, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michał Wojciech Szcześniak, et al. 2016. "A Survey of Best Practices for RNA-Seq Data Analysis." *Genome Biology* 17 (January): 13.

Costa-Silva, Juliana, Douglas Domingues, and Fabricio Martins Lopes. 2017. "RNA-Seq Differential Expression Analysis: An Extended Review and a Software Tool." *PloS One* 12 (12): e0190152.

Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. 2013. "STAR: Ultrafast Universal RNA-Seq Aligner." *Bioinformatics* 29 (1): 15–21.

ENCODE Project Consortium, Michael P. Snyder, Thomas R. Gingeras, Jill E. Moore, Zhiping Weng, Mark B. Gerstein, Bing Ren, et al. 2020. "Perspectives on ENCODE." *Nature* 583 (7818): 693–98.

Ewels, Philip, Måns Magnusson, Sverker Lundin, and Max Käller. 2016. "MultiQC: Summarize

679 Analysis Results for Multiple Tools and Samples in a Single Report." *Bioinformatics* 32
680 (19): 3047–48.
681 "FGED: MINSEQE." n.d. Accessed September 4, 2020. http://fged.org/projects/minseqe/.
682 Gentleman, Robert C., Vincent J. Carey, Douglas M. Bates, Ben Bolstad, Marcel Dettling,
683 Sandrine Dudoit, Byron Ellis, et al. 2004. "Bioconductor: Open Software Development for
684 Computational Biology and Bioinformatics." *Genome Biology* 5 (10): R80.
685 Huber, Wolfgang, Vincent J. Carey, Robert Gentleman, Simon Anders, Marc Carlson, Benilton
686 S. Carvalho, Hector Corrada Bravo, et al. 2015. "Orchestrating High-Throughput Genomic
687 Analysis with Bioconductor." *Nature Methods* 12 (2): 115–21.
688 Jain, Miten, Hugh E. Olsen, Benedict Paten, and Mark Akeson. 2016. "Erratum to: The Oxford
689 Nanopore MinION: Delivery of Nanopore Sequencing to the Genomics Community."
690 *Genome Biology* 17 (1): 256.
691 Jiang, L., F. Schlesinger, C. A. Davis, Y. Zhang, R. Li, M. Salit, T. R. Gingeras, and B. Oliver.
692 2011. "Synthetic Spike-in Standards for RNA-Seq Experiments." *Genome Research*.
693 https://doi.org/10.1101/gr.121095.111.
694 Krueger, Felix. 2019. *Trim Galore: A Wrapper around Cutadapt and FastQC to Consistently*
695 *Apply Adapter and Quality Trimming to FastQ Files, with Extra Functionality for RRBS*
696 *Data* (version Version 0.6.5). https://github.com/FelixKrueger/TrimGalore.
697 Langmead, Ben, Cole Trapnell, Mihai Pop, and Steven L. Salzberg. 2009. "Ultrafast and
698 Memory-Efficient Alignment of Short DNA Sequences to the Human Genome." *Genome*
699 *Biology* 10 (3): R25.
700 Law, Charity W., Monther Alhamdoosh, Shian Su, Xueyi Dong, Luyi Tian, Gordon K. Smyth,
701 and Matthew E. Ritchie. 2016. "RNA-Seq Analysis Is Easy as 1-2-3 with Limma, Glimma
702 and edgeR." *F1000Research* 5 (June). https://doi.org/10.12688/f1000research.9005.3.
703 Liao, Yuxing, Jing Wang, Eric J. Jaehnig, Zhiao Shi, and Bing Zhang. 2019. "WebGestalt 2019:
704 Gene Set Analysis Toolkit with Revamped UIs and APIs." *Nucleic Acids Research* 47
705 (W1): W199–205.
706 Li, Bo, and Colin N. Dewey. 2011. "RSEM: Accurate Transcript Quantification from RNA-Seq
707 Data with or without a Reference Genome." *BMC Bioinformatics* 12 (August): 323.
708 Li, H., and R. Durbin. 2009. "Fast and Accurate Short Read Alignment with Burrows-Wheeler
709 Transform." *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btp324.

710 Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. "Moderated Estimation of Fold
711 Change and Dispersion for RNA-Seq Data with DESeq2." *Genome Biology* 15 (12): 550.
712 Martin, Marcel. 2011. "Cutadapt Removes Adapter Sequences from High-Throughput
713 Sequencing Reads." *EMBnet.journal* 17 (1): 10–12.
714 McIntyre, Lauren M., Kenneth K. Lopiano, Alison M. Morse, Victor Amin, Ann L. Oberg, Linda
715 J. Young, and Sergey V. Nuzhdin. 2011. "RNA-Seq: Technical Variability and Sampling."
716 *BMC Genomics* 12 (June): 293.
717 Mi, Huaiyu, Anushya Muruganujan, and Paul D. Thomas. 2013. "PANTHER in 2013: Modeling
718 the Evolution of Gene Function, and Other Gene Attributes, in the Context of Phylogenetic
719 Trees." *Nucleic Acids Research* 41 (Database issue): D377–86.
720 Munro, Sarah A., Steven P. Lund, P. Scott Pine, Hans Binder, Djork-Arné Clevert, Ana Conesa,
721 Joaquin Dopazo, et al. 2014. "Assessing Technical Performance in Differential Gene
722 Expression Experiments with External Spike-in RNA Control Ratio Mixtures." *Nature*
723 *Communications* 5 (September): 5125.
724 Nguyen, Tuan-Minh, Adib Shafi, Tin Nguyen, and Sorin Draghici. 2019. "Identifying
725 Significantly Impacted Pathways: A Comprehensive Review and Assessment." *Genome*
726 *Biology* 20 (1): 203.
727 Patro, Rob, Geet Duggal, Michael I. Love, Rafael A. Irizarry, and Carl Kingsford. 2017.
728 "Salmon Provides Fast and Bias-Aware Quantification of Transcript Expression." *Nature*
729 *Methods* 14 (4): 417–19.
730 Polo, San-Huei Lai, Amanda M. Saravia-Butler, Valery Boyko, Marie T. Dinh, Yi-Chun Chen,
731 Homer Fogle, Sigrid S. Reinsch, et al. 2020. "RNAseq Analysis of Rodent Spaceflight
732 Experiments Is Confounded by Sample Collection Techniques."
733 https://doi.org/10.1101/2020.07.18.209775.
734 Raplee, Isaac D., Alexei V. Evsikov, and Caralina Marín de Evsikova. 2019. "Aligning the
735 Aligners: Comparison of RNA Sequencing Data Alignment and Gene Expression
736 Quantification Tools for Clinical Breast Cancer Research." *Journal of Personalized*
737 *Medicine* 9 (2). https://doi.org/10.3390/jpm9020018.
738 Risso, Davide, John Ngai, Terence P. Speed, and Sandrine Dudoit. 2014. "Normalization of
739 RNA-Seq Data Using Factor Analysis of Control Genes or Samples." *Nature Biotechnology*
740 32 (9): 896–902.

741 Risso, Davide, Katja Schwartz, Gavin Sherlock, and Sandrine Dudoit. 2011. "GC-Content
742 Normalization for RNA-Seq Data." *BMC Bioinformatics* 12 (December): 480.
743 Roberts, Richard J., Mauricio O. Carneiro, and Michael C. Schatz. 2013. "The Advantages of
744 SMRT Sequencing." *Genome Biology* 14 (7): 405.
745 Rocca-Serra, P., M. Brandizi, E. Maguire, N. Sklyar, C. Taylor, K. Begley, D. Field, et al. 2010.
746 "ISA Software Suite: Supporting Standards-Compliant Experimental Annotation and
747 Enabling Curation at the Community Level." *Bioinformatics*.
748 https://doi.org/10.1093/bioinformatics/btq415.
749 Rutter, Lindsay, Richard Barker, Daniela Bezdan, Henry Cope, Sylvain V. Costes, Lovorka
750 Degoricija, Kathleen M. Fisch, et al. n.d. "A New Era for Space Life Science: International
751 Standards for Space Omics Processing (ISSOP)." *Cell*. Accessed August 26, 2020.
752 https://drive.google.com/drive/u/0/folders/1BQ72FMlHZ_GR777KaZ6nCL3uTe4Q07x4.
753 Sansone, Susanna-Assunta, Philippe Rocca-Serra, Dawn Field, Eamonn Maguire, Chris Taylor,
754 Oliver Hofmann, Hong Fang, et al. 2012. "Toward Interoperable Bioscience Data." *Nature*
755 *Genetics* 44 (2): 121–26.
756 Schaarschmidt, Stephanie, Axel Fischer, Ellen Zuther, and Dirk K. Hincha. 2020. "Evaluation of
757 Seven Different RNA-Seq Alignment Tools Based on Experimental Data from the Model
758 Plant Arabidopsis Thaliana." *International Journal of Molecular Sciences* 21 (5).
759 https://doi.org/10.3390/ijms21051720.
760 Schurch, Nicholas J., Pietá Schofield, Marek Gierliński, Christian Cole, Alexander Sherstnev,
761 Vijender Singh, Nicola Wrobel, et al. 2016. "How Many Biological Replicates Are Needed
762 in an RNA-Seq Experiment and Which Differential Expression Tool Should You Use?"
763 *RNA*. https://doi.org/10.1261/rna.053959.115.
764 Scott, Ryan T., Kirill Grigorev, Graham Mackintosh, Samrawit G. Gebre, Christopher E. Mason,
765 Martha E. Del Alto, and Sylvain V. Costes. 2020. "Advancing the Integration of
766 Biosciences Data Sharing to Further Enable Space Exploration." *Cell Reports*.
767 Smith, Tom, Andreas Heger, and Ian Sudbery. 2017. "UMI-Tools: Modeling Sequencing Errors
768 in Unique Molecular Identifiers to Improve Quantification Accuracy." *Genome Research* 27
769 (3): 491–99.
770 Soneson, C., M. Love, and M. Robinson. n.d. "Differential Analyses for RNA-Seq: Transcript-
771 Level Estimates Improve Gene-Level Inferences [version 2; Peer Review: 2 Approved]."

772 F1000Research.
773 Subramanian, Aravind, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert,
774 Michael A. Gillette, Amanda Paulovich, et al. 2005. "Gene Set Enrichment Analysis: A
775 Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles."
776 *Proceedings of the National Academy of Sciences of the United States of America* 102 (43):
777 15545–50.
778 Szklarczyk, D., A. L. Gable, D. Lyon, and A. Junge. 2019. "STRING v11: Protein–protein
779 Association Networks with Increased Coverage, Supporting Functional Discovery in
780 Genome-Wide Experimental Datasets." *Nucleic Acids*.
781 https://academic.oup.com/nar/article-abstract/47/D1/D607/5198476.
782 Teng, Mingxiang, Michael I. Love, Carrie A. Davis, Sarah Djebali, Alexander Dobin, Brenton R.
783 Graveley, Sheng Li, et al. 2016. "A Benchmark for RNA-Seq Quantification Pipelines."
784 *Genome Biology* 17 (April): 74.
785 Thomas, P. D. 2003. "PANTHER: A Library of Protein Families and Subfamilies Indexed by
786 Function." *Genome Research*. https://doi.org/10.1101/gr.772403.
787 Väremo, Leif, Jens Nielsen, and Intawat Nookaew. 2013. "Enriching the Gene Set Analysis of
788 Genome-Wide Data by Incorporating Directionality of Gene Expression and Combining
789 Statistical Hypotheses and Methods." *Nucleic Acids Research* 41 (8): 4378–91.
790 Wickham, Hadley, Danielle Navarro, and Thomas Lin Pedersen. 2016. *ggplot2: Elegant*
791 *Graphics for Data Analysis*. Springer-Verlag New York.
792 Williams, Claire R., Alyssa Baccarella, Jay Z. Parrish, and Charles C. Kim. 2016. "Trimming of
793 Sequence Reads Alters RNA-Seq Gene Expression Estimates." *BMC Bioinformatics* 17
794 (February): 103.
795 Wu, Douglas C., Jun Yao, Kevin S. Ho, Alan M. Lambowitz, and Claus O. Wilke. 2018.
796 "Limitations of Alignment-Free Tools in Total RNA-Seq Quantification." *BMC Genomics*
797 19 (1): 510.
798 Yang, Cheng, Po-Yen Wu, Li Tong, John H. Phan, and May D. Wang. 2015. "The Impact of
799 RNA-Seq Aligners on Gene Expression Estimation." ACM-BThe ACM Conference on
800 Bioinformatics, Computational Biology and Biomedicine. ACM Conference on
801 Bioinformatics, Computational Biology and Biomedicine (September): 462–71.

802

Figures

NASA GeneLab RNA-seq Consensus Pipeline: Standardized Processing of Short-Read RNA-Seq Data

Overbey et al. (2020)

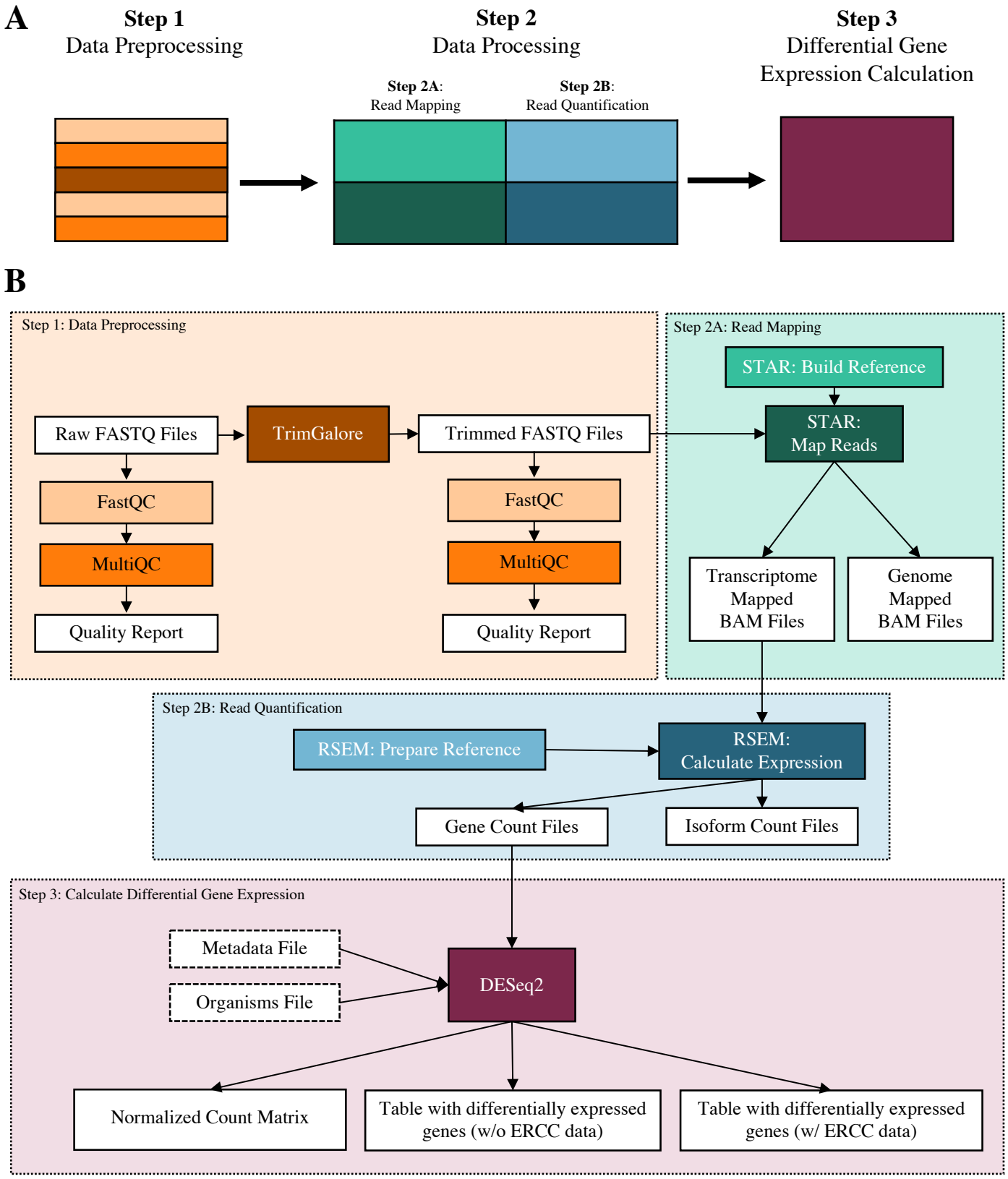
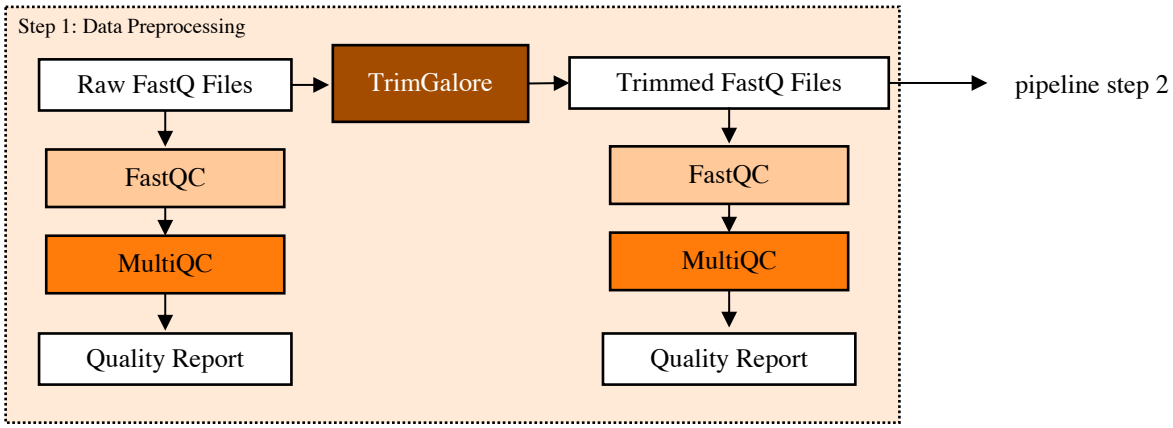


Figure 1: GeneLab RNA-seq Consensus Pipeline (RCP). A: The three broad steps of the RCP. The RCP handles: 1) Data preprocessing to trim sequencing adapters and to provide quality control metrics; 2) Data processing to map reads to the reference genome and quantify the number of read counts per gene; and 3) Differential gene expression calculation, which will provide a list of differentially expressed genes that can be sorted by adjusted p-value and log fold-change. **B:** The full RCP annotated with tools, input files, and output files.

A



B

FastQC	
Parameters	fastqc -o /path/to/output/directory \ -t number_of_threads \ /path/to/input/files
Input data files	fastq.gz
Output data files	fastqc.html (FastQC report) fastqc.zip (FastQC raw data)

C

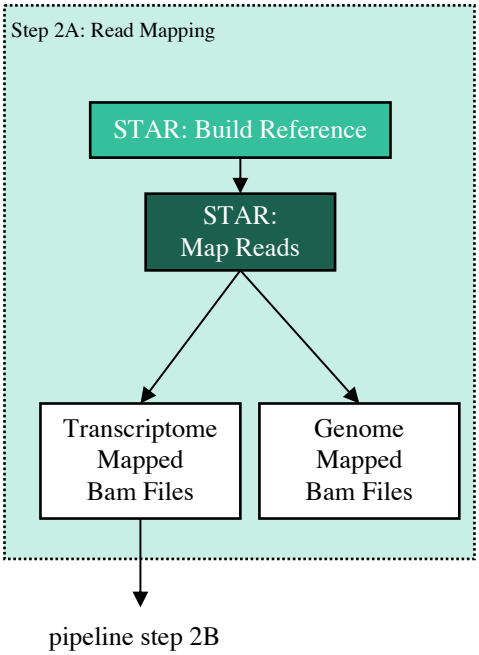
MultiQC	
Parameters	multiqc -o /path/to/output/directory \ /path/to/fastqc/output/files
Input data files	fastqc.html (FastQC report) fastqc.zip (FastQC raw data)
Output data files	multiqc_report.html (multiqc report) multiqc_data (directory containing multiqc raw data)

D

TrimGalore	
Parameters	trim_galore --gzip \ --path_to_cutadapt /path/to/cutadapt \ --phred33 \ --illumina \ # if adapters are not illumina, replace with adapters used --output_dir /path/to/TrimGalore/output/directory \ --paired \ # only for PE studies /path/to/forward/reads /path/to/reverse/reads # if SE, replace the last line with only /path/to/forward/reads
Input data files	*fastq.gz (raw reads)
Output data files	*fastq.gz (trimmed reads) *trimming_report.txt (trimming report)

Figure 2: Data preprocessing (pipeline step 1): Quality control and trimming. A: Data Preprocessing pipeline. FastQ files from Illumina base-calling software are quality checked using FastQC and MultiQC. Data is then trimmed using TrimGalore and are re-checked for quality; **B:** Flags used for FastQC program; **C:** Flags used for MultiQC program; **D:** Flags used for TrimGalore program; trimmed reads (*fastq.gz) are then used as input data for FastQC (B) followed by MultiQC (C) to generate trimmed read quality metrics. Tool versions used to process each dataset are included in the RNA-seq processing protocol in the GLDS Repository.

A



B

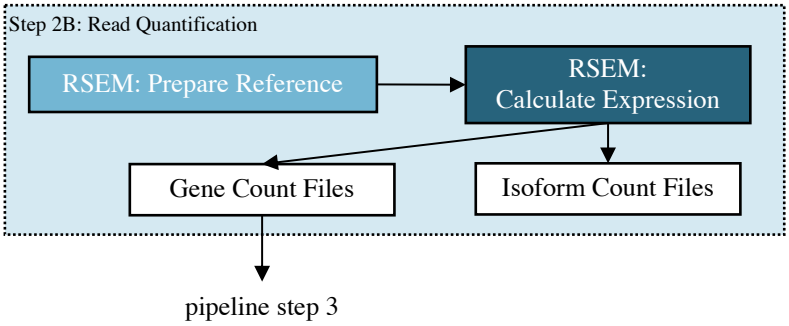
STAR	Build Reference
Parameters	STAR --runThreadN NumberOfThreads \ # Number of available cores on server node --runMode genomeGenerate \ --limitGenomeGenerateRAM 35000000000 \ # min needed for mouse ref --genomeDir /path/to/STAR/genome/directory \ --genomeFastaFiles /path/to/genome/fasta/file \ --sjdbGTFfile /path/to/annotation/gtf/file \ --sjdbOverhang ReadLength-1
Input data files	.fasta (genome sequences) .gtf (genome annotation) Note that for analyzing samples with ERCC spike-in, the genome fasta and gtf files should have the ERCC fasta and gtf files appended them.
Output data files	The following will be generated in the /STAR/genome/directory: • chrLength.txt • chrNameLength.txt • chrName.txt • chrStart.txt • exonGeTrInfo.tab • exonInfo.tab • geneInfo.tab • Genome • genomeParameters.txt • SA • SAindex • sjdbInfo.txt • sjdbList.fromGTF.out.tab • sjdbList.out.tab • transcriptInfo.tab

C

STAR	Map Reads	
Parameters	STAR --twopassMode Basic \ --limitBAMsortRAM available_memory_in_bytes \ --genomeDir /path/to/STAR/genome/directory \ --outSAMunmapped Within \ --outFilterType BySJout \ --outSAMattributes NH HI AS NM MD MC \ --outFilterMultimapNmax 20 \ --outFilterMismatchNmax 999 \ --outFilterMismatchNoverReadLmax 0.04 \ --alignIntronMin 20 \ --alignIntronMax 1000000 \ --alignMatesGapMax 1000000 \ # only needed for PE studies --alignSJoverhangMin 8 \ --alignSJDBoverhangMin 1 \ --sjdbScore 1 \ --readFilesCommand zcat \ --runThreadN NumberOfThreads \ --outSAMtype BAM SortedByCoordinate \ --quantMode TranscriptomeSAM \ --outSAMheaderHD @HD VN:1.4 SO:coordinate \ --outFileNamePrefix /path/to/STAR-output/directory/<sample_name> \ --readFilesIn /path/to/trimmed_forward_reads \ /path/to/trimmed_reverse_reads # only needed for PE studies	
Input data files	STAR index directory *fastq.gz (trimmed reads)	
Output data files	Files *Aligned.sortedByCoord.out.bam (sorted mapping to genome) *Aligned.toTranscriptome.out.bam (sorted mapping to transcriptome) *Log.final.out (reads mapped, etc) *Log.out *Log.progress.out *SJ.out.tab	Directories *_STARgenome sjdbInfo.txt sjdbList.out.tab *_STARpass1 Log.final.out SJ.out.tab *_STARtmp directory containing subdirectories that are empty – this was the location for temp files that were automatically removed after successful completion

Figure 3: Data processing (pipeline step 2A): Read mapping. A: Data processing pipeline. Trimmed reads are mapped to their reference genome and transcriptome with STAR. Gene counts are then quantified with RSEM; **B:** Flags used for generating the indexed STAR reference files; **C:** Flags used for mapping reads with STAR. Tool versions used to process each dataset are included in the RNA-seq processing protocol in the GLDS Repository.

A



B

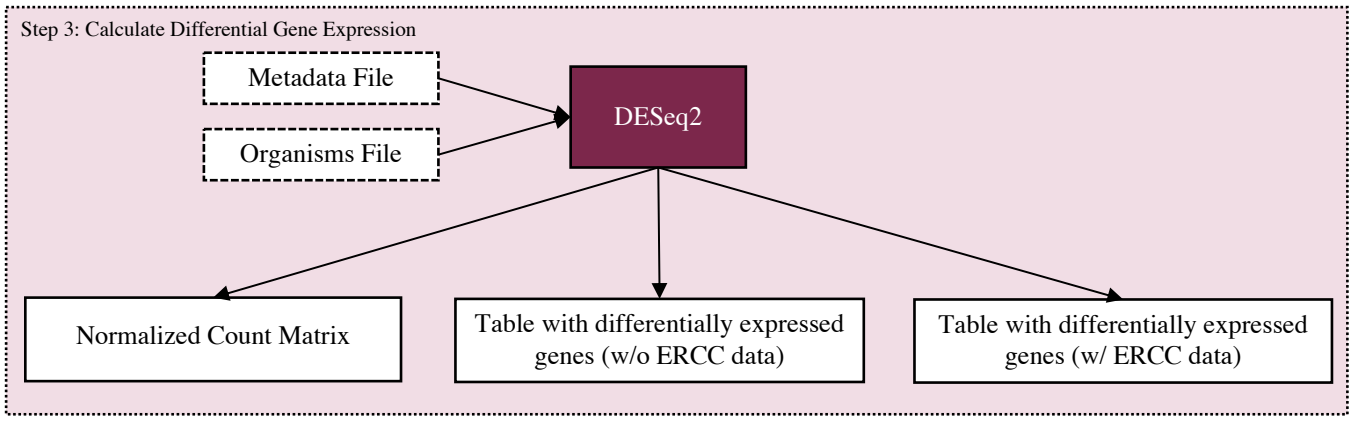
RSEM	Prepare Reference
Parameters	rsem-prepare-reference --gtf /path/to/annotation/gtf/file \ /path/to/genome/fasta/file \ /path/to/RSEM/genome/directory/RSEM_ref_prefix
Input data files	.fasta (genome sequence) .gtf (genome annotation) Note that for analyzing samples with ERCC spike-in, the genome FASTA and GTF files should have the ERCC FASTA and GTF files appended to them.
Output data files	The following will be generated in the /RSEM/genome/directory RSEM_ref_prefix.chrlist RSEM_ref_prefix.grp RSEM_ref_prefix.idx.fa RSEM_ref_prefix.n2g.idx.fa RSEM_ref_prefix.seq RSEM_ref_prefix.ti RSEM_ref_prefix.transcripts.fa

C

RSEM	Calculate Expression
Parameters	rsem-calculate-expression --num-threads NumberOfThreads \ --alignments \ --bam \ --paired-end \ # only for PE studies --seed 12345 \ --estimate-rspd \ --no-bam-output \ --strandedness reverse \ # For Illumina TruSeq stranded protocols, reads are derived from the reverse strand /path/to/*Aligned.toTranscriptome.out.bam \ /path/to/RSEM/genome/directory/RSEM_ref_prefix \ /path/to/RSEM/output/directory
Input data files	*Aligned.toTranscriptome.out.bam RSEM indexed reference
Output data files	*genes.results (counts per gene) *isoforms.results (counts per isoform) *stat (directory containing the following stats files) *cnt *model *theta

Figure 4: Data processing (pipeline step 2B): Gene quantification. A: Data processing pipeline. Mapping results from STAR are quantified by RSEM; **B:** Parameters for RSEM indexed reference files generation; **C:** Parameters for quantifying gene and isoform counts with RSEM. Tool versions used to process each dataset are included in the RNA-seq processing protocol in the GLDS Repository.

A



B

Output Type	**Output File Name** (GeneLab_DGE_noERCC.R and GeneLab_DGE_wERCC.R)	**Output File Name** (GeneLab_DGE_wERCC.R only)
Available with RNA-seq processed data in the GLDS Repository (*)	Unnormalized_Counts.csv Normalized_Counts.csv contrasts.csv differential_expression.csv	ERCCnorm_contrasts.csv ERCC_Normalized_Counts.csv ERCCnorm_differential_expression.csv
Used to generate interactive plots from RNA-seq processed data in the GLDS visualization portal (**)	visualization_output_table.csv visualization_PCA_table.csv	visualization_output_table_ERCCnorm.csv visualization_PCA_table_ERCCnorm.csv
Used for internal QC and/or V&V (#)	SampleTable.csv	ERCC_rawCounts_unfiltered.csv ERCC_rawCounts_filtered.csv

Figure 5: Differential gene expression calculation (pipeline step 3). A: Data processing pipeline. The R program DESeq2 is run in order to determine which genes are differentially expressed between experimental conditions using gene count files from RSEM. **B:** Output files generated. The table columns distinguish which script produces each output. The columns distinguish how those output files are used.

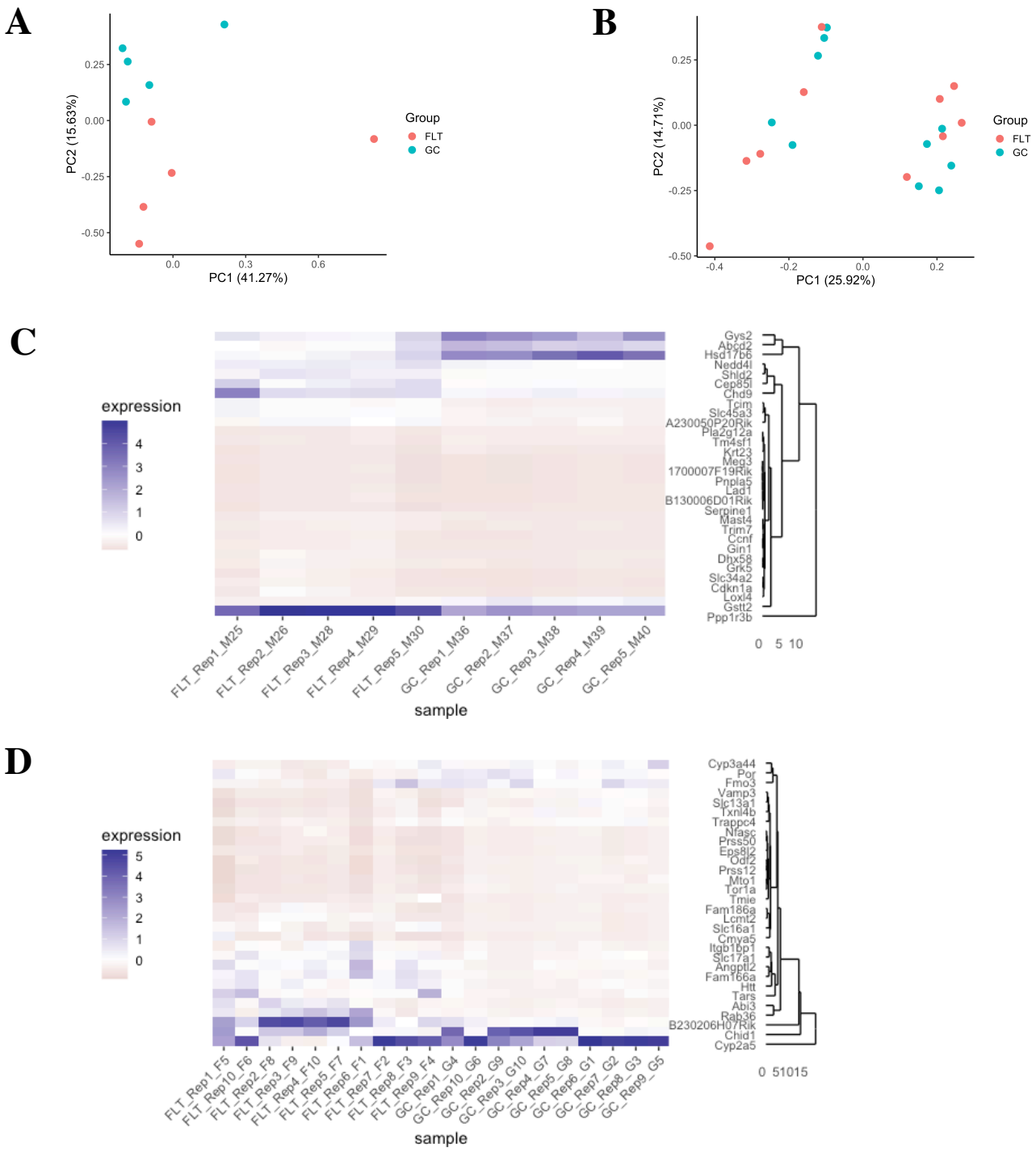


Figure 6. Global and differential gene expression in spaceflight versus ground control liver samples from GeneLab datasets. A, B: Principal component analysis of global gene expression in spaceflight (FLT) and respective ground control (GC) liver samples from the A) Rodent Research 1 (RR-1) NASA Validation mission (GLDS-168) and B) RR-6 ISS-terminal mission (GLDS-245). Plots were generated using data in the normalized counts tables for each respective dataset on the NASA GeneLab Data Repository. **C, D:** Heatmaps showing the top 30 differentially expressed genes in spaceflight (FLT) versus ground control (GC) liver samples from the C) Rodent Research 1 (RR-1) NASA Validation mission (GLDS-168) and D) RR-6 ISS-terminal mission (GLDS-245). Heatmaps were generated using data in the differential expression tables for each respective dataset on the NASA GeneLab Data Repository and is colored by relative expression. Adj. p-value < 0.05 and |log2FC| > 1. All samples included were derived from frozen carcasses post-mission and utilized the ribo-depletion library preparation method.

TAIR	SYMBOL	GENENAME	REFSEQ	ENTREZID	STRING_id	GOSLIM_IDS
AT1G01010	ANAC001	NA	NM_099983	839580	3702.AT1G01010.1	NA
AT1G01020	ARV1	NA	NM_001035846	839569	3702.AT1G01020.1	GO:0005622, GO:0005737, …
AT1G01030	NGA3	NA	NM_001331244	839321	3702.AT1G01030.1	NA
AT1G01040	ASU1	Encodes a Dicer homolog...	NM_001197952	839574	3702.AT1G01040.2	NA

Table 1. Differential gene expression output table - Annotations. Truncated version of the differential_expression.csv file provided as GeneLab processed data for GLDS-251. The first 7 columns of the differential gene expression output table contain gene IDs and annotations (for remainder of columns, refer to Table 2).

Norm. expr. (sample A)	Log2fc (comparison A)	P.value (comparison A)	Adj.p.value (comparison A)	Mean (all samples)	Stdev (all samples)	LRT p.value	Mean (group A)	Stdev (group A)
263.864	-0.078	0.648	0.848	198.735	31.756	0.484	225.550	36.759
200.493	0.341	0.033	0.198	147.061	19.197	0.740	174.839	24.073
19.040	0.691	0.137	NA	11.035	3.121	NA	15.706	2.889
644.811	0.126	0.366	0.655	669.586	68.327	1.000	688.123	76.969

Table 2. Differential gene expression output table - Statistics. Truncated version of the differential_expression.csv file provided as GeneLab processed data for GLDS-251. Following the 7 columns of gene IDs and annotations (Table 1) are normalized gene expression data for each sample (**Norm. expr. (sample A)**) then results from all possible pairwise comparisons, including log2 fold change (**Log2fc (comparison A)**), p values (**P.value (comparison A)**), and adjusted p values (**Adj.p.value (comparison A)**) calculated from the Wald Tests. Next are the average gene expression (**Mean (all samples)**) and standard deviation (**Stdev (all samples)**) of all samples followed by the F-statistic p value generated from the likelihood ratio test (LRT.p.value), and the last set of columns are the average gene expressions (Group.Mean) and standard deviations (Group.Stdev) of samples within each group.

GeneLab Dataset	# Enriched GO terms (NOM p<0.01)	# Enriched GO terms (NOM p<0.01 & FDR<0.5)	# Enriched GO terms (NOM p<0.01 & FDR<0.25)
GLDS-168	71, 135	0, 132	0, 0
GLDS-245	21, 24	2, 6	1, 0

Table 3. Comparison of gene ontology in spaceflight versus ground control liver samples from GeneLab datasets. The number of enriched gene ontology (GO) terms identified by Gene Set Enrichment Analysis (GSEA, phenotype permutation) was evaluated in spaceflight (FLT) versus ground control (GC) liver samples from the Rodent Research 1 (RR-1) NASA Validation mission (GLDS-168), and RR-6 ISS-terminal mission (GLDS-245). For GO terms, the number on the left corresponds to GO terms enriched in FLT samples and the number on the right corresponds to GO terms enriched in GC samples. These data were generated using the normalized counts for each respective dataset on the NASA GeneLab Data Repository. All samples included were derived from frozen carcasses post-mission and utilized the ribo-depletion library preparation method. GLDS-168, FLT n=5 and GC n=5; GLDS-245, FLT n=10 and GC n=10. p values and FDR values are indicated.