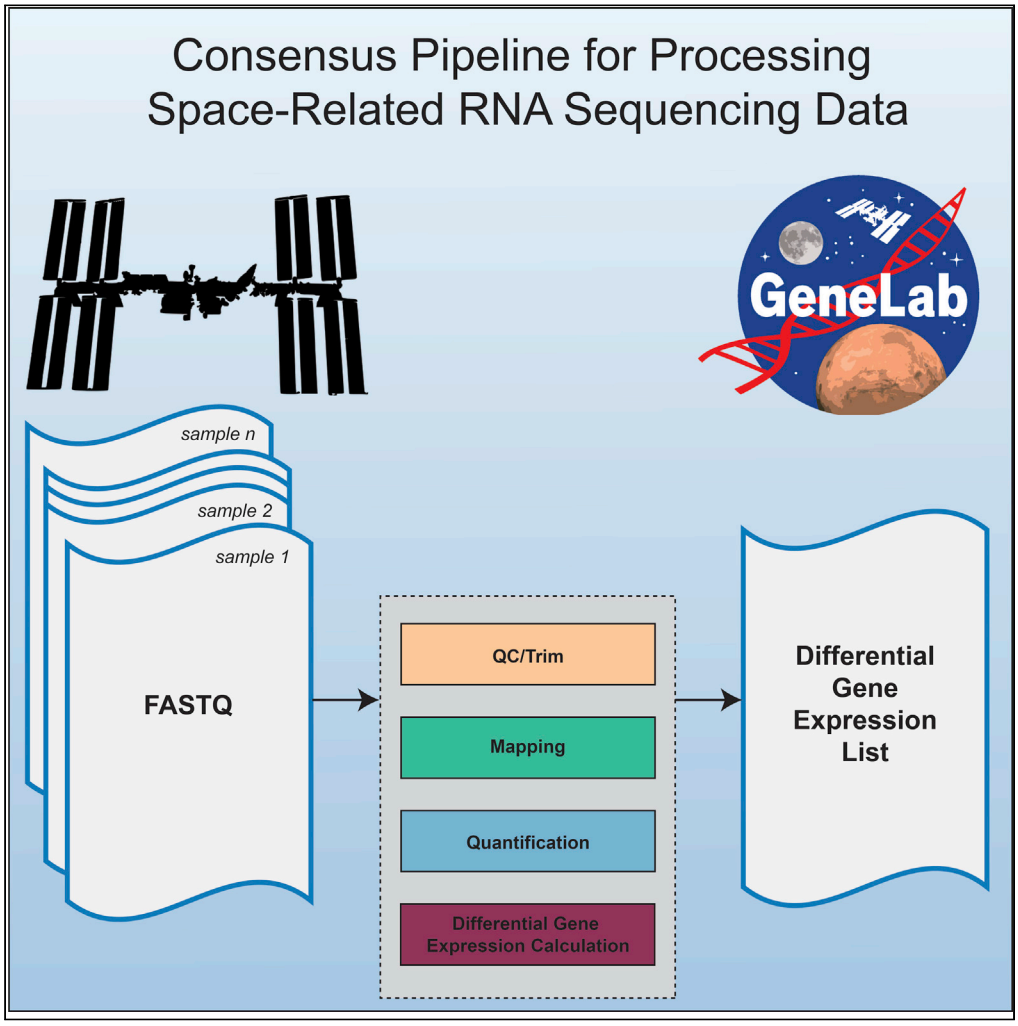
iScience

CellPress
OPEN ACCESS

Article

NASA GeneLab RNA-seq consensus pipeline: standardized processing of short-read RNA-seq data



Eliah G. Overbey, Amanda M. Saravia-Butler, Zhe Zhang, ..., Luis Zea, Sylvain V. Costes, Jonathan M. Galazka

sylvain.v.costes@nasa.gov (S.V.C.)
jonathan.m.galazka@nasa.gov (J.M.G.)

Highlights

Analysis of omics data from different spaceflight studies presents unique challenges

A standardized pipeline for RNA-seq analysis eliminates data processing variation

The GeneLab RNA-seq pipeline includes QC, trimming, mapping, quantification, and DGE

Space-relevant data processed with this pipeline are available at genelab.nasa.gov

Overbey et al., iScience 24, 102361
April 23, 2021
https://doi.org/10.1016/j.isci.2021.102361

iScience

CellPress
OPEN ACCESS

Article

NASA GeneLab RNA-seq consensus pipeline: standardized processing of short-read RNA-seq data

Eliah G. Overbey,[1,39] Amanda M. Saravia-Butler,[2,3,39] 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,[3] 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,*] and Jonathan M. Galazka[3,40,*]

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. In addition, 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., 2020). 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 et al. 2017) all affect results. Because of such challenges, groups such as ENCODE and MINSEQE have developed standardized analysis pipelines for better comparison of RNA-seq datasets (ENCODE Project Consortium et al., 2020; Functional Genomics Data Society, 2012).

[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, PA 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, 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

Continued

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 expression, metabolite pools, and metagenomes. The most prevalent data type in this repository is RNA-seq from organisms exposed to spaceflight conditions. As of August 2020, the NASA GeneLab database has over eighty datasets with RNA-sequencing data (Table S1). These datasets include *Homo sapiens* (human), *Mus musculus* (mouse), *Drosophila melanogaster* (fruit fly), *Arabidopsis thaliana* (model higher plant), *Oryzias latipes* (Japanese rice fish), *Helix lucorum* (land snail), *Brassica rapa* (Fast Plant), *Eruca vesicaria* (arugula/edible plant), *Euprymna scolopes* (Hawaiian bobtail squid), *Ceratopteris richardii* (aquatic fern), and the bacterium, *Bacillus subtilis* from experiments performed during true spaceflight on various orbital platforms such as the Space Shuttle and International Space Station (ISS), as well as spaceflight-analog studies, such as hindlimb unloading and bed rest studies (Berrios et al., 2020).

NASA's GeneLab and Ames Life Sciences Data Archive (ALSDA) projects have put forward an ambitious strategy focused on integrating data, metadata, and biospecimens to fully utilize the 40+ years of archived NASA Life Sciences data (Scott et al., 2020). One of the first steps in this effort is the ability to analyze how experimental factors common to multiple datasets impact molecular signaling. Such meta-analysis can only occur if metadata, data, and processed data are harmonized. As part of this strategy, GeneLab engaged with the scientific community and held its first Analysis Working Group (AWG) workshop in 2018. Spaceflight researchers from universities and organizations across the United States and abroad met to begin the creation of a standardized, consensus data-processing pipeline for one of the most common types of spaceflight datasets: transcription profiling via RNA-sequencing. Scientists at this workshop met to discuss the merits of various bioinformatic software tools for processing RNA-sequencing data and ultimately agreed on a single pipeline of these tools.

The main driver for developing the consensus pipeline was to present consistently processed data to the public, therefore making space-relevant multi-omics data more accessible and reusable. The overall goals were (1) to get more consistently processed data to the public; (2) to provide output data from every step of the consensus pipeline so users can download and use these "intermediate" data; (3) to support easier and more consistent analysis of space-relevant data by users including those in the NASA AWGs; and (4) to allow easier cross-analysis of experiments to identify effects that result from the spaceflight environment, independent of confounding factors. In addition, many of these data in the GeneLab database have not been previously 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 preprocessing, data processing, and differential gene expression computation/annotation (Figure 1A). These steps use tools for quality control (FastQC, MultiQC) (Andrews, 2010; Ewels et al., 2016), read trimming (TrimGalore) (Krueger 2019), mapping (STAR) (Dobin et al., 2013), quantification (RSEM) (Li and Dewey 2011), and differential gene expression calculation/annotation (DESeq2) (Love et al. 2014) (Figure 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 (Figure 2A). The pipeline begins with quality control (QC) of raw FASTQ files from a short-read Illumina sequencer using the FastQC software (Andrews, 2010) (Figure 2B). FastQC is one of the most widely used QC programs for short-read sequencing data. It provides information that 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

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 92093, 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 92093, USA

[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, USA

[26]NGM Biopharmaceuticals, South San Francisco, CA 94080, USA

[27]National Jewish Health, Center for Genes, Environment, and Health,

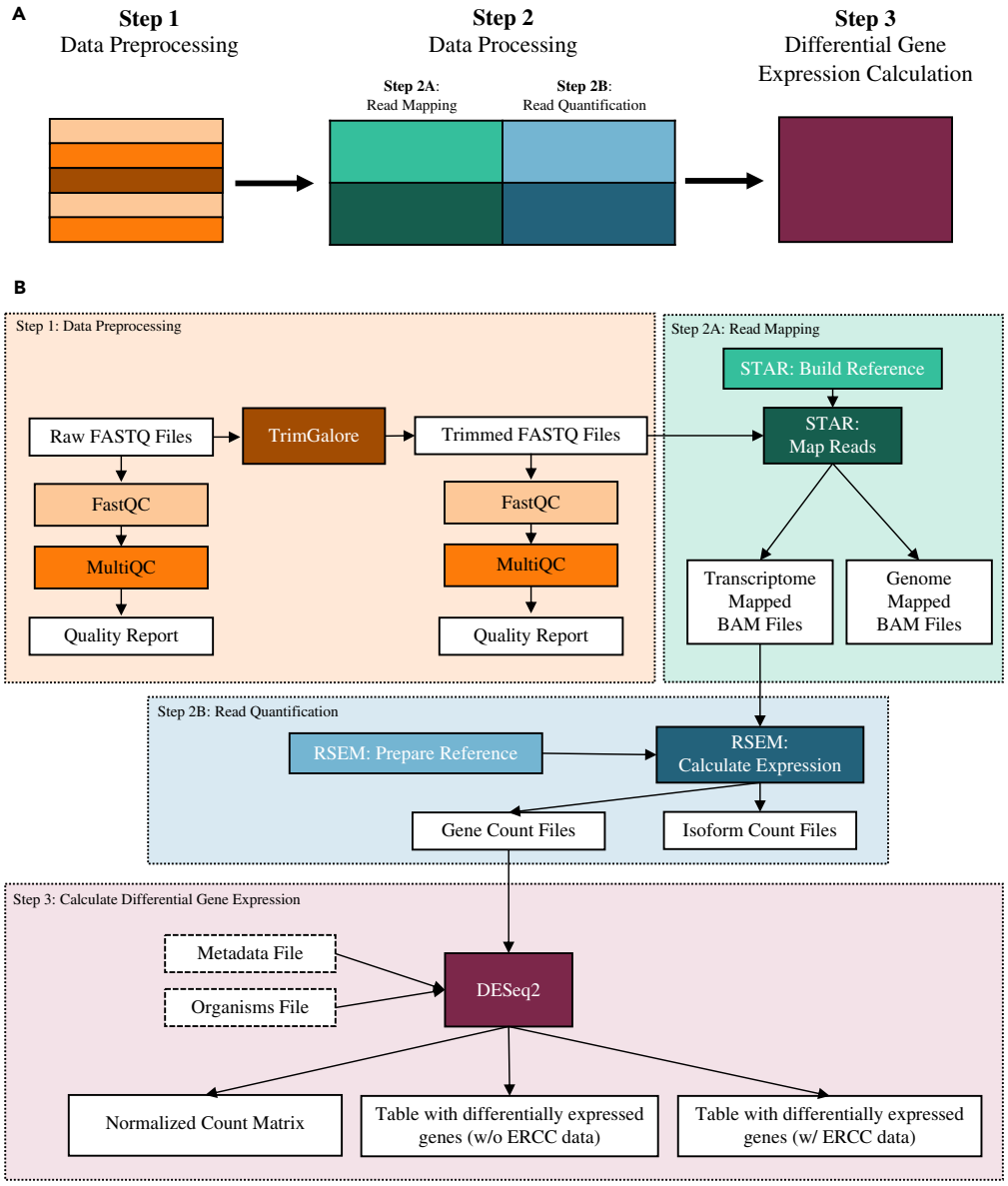


Figure 1. GeneLab RNA-seq Consensus Pipeline (RCP)

(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.

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) (Figure 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 and low-quality bases that would disrupt read mapping during the data processing pipeline step (Figure 2D). 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

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, FL 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, PA 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, 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

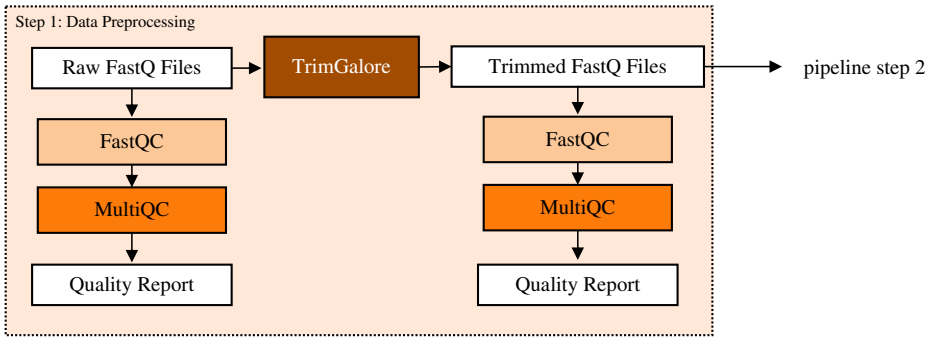
[39]These authors contributed equally

[40]Lead contact

*Correspondence: sylvain.v.costes@nasa.gov (S.V.C.), jonathan.m.galazka@nasa.gov (J.M.G.)

https://doi.org/10.1016/j.isci.2021.102361

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

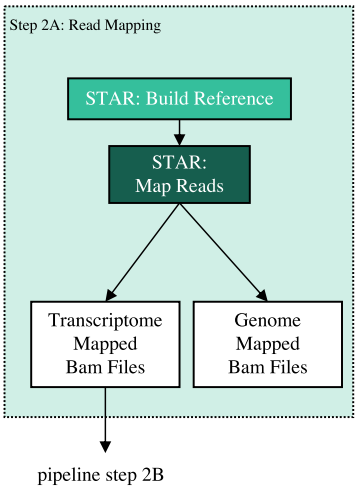
(A) Data preprocessing pipeline. FastQ files from Illumina base-calling software are quality checked using FastQC and MultiQC. Data are then trimmed using TrimGalore and are re-checked for quality; (B) flags used for FastQC program; (C) flags used for MultiQC program; (D) flags used for TrimGalore program; 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.

step, bases that are part of a sequencing adapter or of low quality are removed from each read, and reads that become too short are subsequently 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 have been preprocessed, the sequenced reads are ready for mapping and quantification.

Data processing: read mapping and sample quantification

In the data processing step (Figure 1; Step 2A), the trimmed reads are first aligned to the reference genome (Figure 3A) with STAR, a splice-aware aligner (Dobin et al., 2013). STAR must be run in two steps. The first

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.

step is to create indexed genome files (Figure 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 (Figure 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 Binary Alignment Map (BAM) format, which has a separate entry for each mapped read and states which transcript each read is 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 (Figure 1A; Step 2B, Figure 4A). For this task, the RCP uses RSEM (Li and Dewey 2011). 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 (Figure 4A). The first phase uses the reference genome and GTF files (with or without ERCC as appropriate) (Table S2) to prepare indexed genome files (Figure 4B). The second phase uses the indexed files and the mapped reads from STAR to assign counts to each gene (Figure 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 Table S2 . These reference genomes were the most recent releases at the time each STAR and RSEM indexed references were created. Although 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.

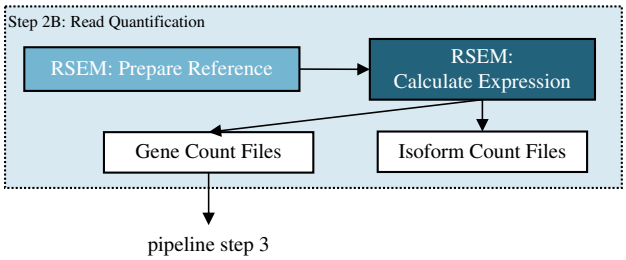
We elected to adopt a mapping-based approach rather than rapidly quantifying the reads via a k-mer-based counting algorithm, pseudo-aligners, or a quasi-mapping method that utilizes RNA-seq inference procedures such as Kallisto (Bray et al., 2016) or Salmon (Patro et al., 2017) despite their speed advantages. 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 (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 et al. 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 (Figure 1; Step 3, Figure 5A). Unlike the previous steps, a custom R script (GeneLab_DGE_wERCC.R or GeneLab_DGE_noERCC.R) (Data S1and 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 (Figure 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

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

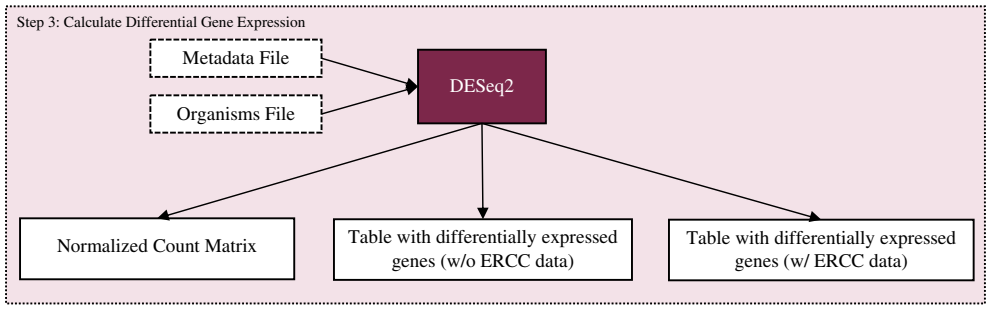
Figure 4. Data processing (pipeline step 2B): gene quantification

(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.

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 scripts in order.

The GeneLab DGE R script requires three inputs: the quantified count 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

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.

annotations to load. Because 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, Data S1) and one for those without (GeneLab_DGE_noERCC.R, Data 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 et al. 2013; Thomas 2003) for that 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 versus ground and ground versus 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 et al., 2015), 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 with ERCC spike-in, we use the GeneLab_DGE_wERCC.R script (Data 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. Because 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).

To enable DESeq2 analysis with and without considering ERCC reads, the DESeqDataSet object is used to create a DESeqDataSet object containing only ERCC reads. Because all samples must contain ERCC spike-in for ERCC-normalization, the DESeqDataSet object containing only ERCC reads is used to identify and remove any samples that do not contain ERCC reads. Next, a DESeqDataSet object containing only non-ERCC reads is created by removing rows containing ERCC reads. These data are then used for DESeq2 analysis.

For DESeq2 analysis with ERCC-normalization (Data S2), the size factor object of the non-ERCC data is replaced with group B ERCC size factors for re-scaling in the first DESeq2 step. Group B ERCC genes contain the same concentration in both mix1 and mix 2. Therefore, only group B ERCC genes are used for generating the size factors for re-scaling duirng ERCC-normalization. For DESeq2 analysis without ERCC-normalization, the DESeq2 default algorithm is applied to the DESeqDataSet object containing only non-ERCC reads. The unnormalized and DESeq2-normalized count data as well as the sample table are then outputted as CSV files. The "Unnormalized_Counts.csv," "Normalized_Counts.csv," and "ERCC_Normalized_Counts.csv" files for each RNA-seq dataset are available in the GeneLab Data Repository; the "SampleTable.csv" file is used internally for verifying and validating the processed data prior to publication.

There are two types of hypothesis tests that can be run with DESeq2, the likelihood ratio test (LRT), which is similar to an analysis of variance (ANOVA) calculation in linear regression and allows for comparison across all groups, and the Wald test, in which the estimated standard error of a log2 fold change is used to compare differences between two groups. The DGE step of the RCP performs both of these analyses. After normalization, the DESeq2 likelihood ratio test design is applied to the normalized data (both ERCC- and nonERCC-normalized data) to generate the F statistic p value, which is similar to an ANOVA p value and reveals genes that are changed in any number of combinations of all factors defined in the experiment.

To prepare for building a gene/pathway annotation database, the STRINGdb (Szklarczyk et al., 2019) and PANTHER.db (Thomas 2003) libraries are loaded, and the organisms.csv file is read and used to indicate the Bioconductor AnnotationData Package needed (Huber et al., 2015; Gentleman et al., 2004). The current gene annotation database for the organism specified at the beginning of the R script is then loaded.

Next, DGE tables containing normalized counts for each sample, pairwise DGE results, and current gene annotations as well as computer-readable DGE tables (that will be used for 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 other 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),

Table 1. Differential gene expression output table—annotations

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

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).

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 downregulated 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 Tables 1 and 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 (**) (Figure 5B). 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 (**) (Tables 1 and 2).

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.

Table 2. Differential gene expression output table—statistics

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

Truncated version of the differential_expression.csv file provided as GeneLab processed data for GLDS-251. Following the seven 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.

Prior to downstream analysis, the processed data files were filtered so that only samples that met the criteria listed above were included. Because 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, because 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 euthanasia, 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. In addition, because 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, Mendeley Data: https://doi.org/10.17632/fv3kd6h7k4.1) were then used to create Principal Component Analysis (PCA) plots (Figures 6A, 6B, S1A, and S1B), heatmaps containing the top 30 most significant FLT versus GC differentially expressed (and annotated) genes (adj. p value <0.05 and |log2FC| > 1) (Figures 6C, 6D, S1C, and S1D), and to evaluate FLT versus GC gene ontology (GO) differences using Gene Set Enrichment (GSEA) analysis (Tables 3 and S5). These results can then be further evaluated to identify similarities and differences in gene expression between these two studies and draw novel conclusions about the effects of spaceflight that are consistent across spaceflight experiments.

DISCUSSION

The differentially expressed genes calculated by the RCP can be further explored with a variety of tools designed for higher-order analysis. For example, there are tools that can look for enriched pathways, gene ontology terms, or protein and/or metabolite networks. Popular software tools among the GeneLab working group members include WebGestalt (Liao et al., 2019), STRING (Szklarczyk et al., 2019), GSEA (Subramanian et al., 2005), PIANO (Väremo et al. 2013), Reactome (Szklarczyk et al., 2019), and ToppFun (Chen et al., 2009). There is no universal consensus on which tools are the most useful for higher-order analysis (Nguyen et al., 2019). RCP users are encouraged to try multiple tools in order to analyze their data from a variety of perspectives.

The RCP has been designed to handle sequencing experiments that either lack or include the ERCC RNA spike-in mix—a set of 96 polyadenylated RNAs that can be used during differential gene expression calculation to normalize read counts across samples (Munro et al., 2014). However, the use of normalization

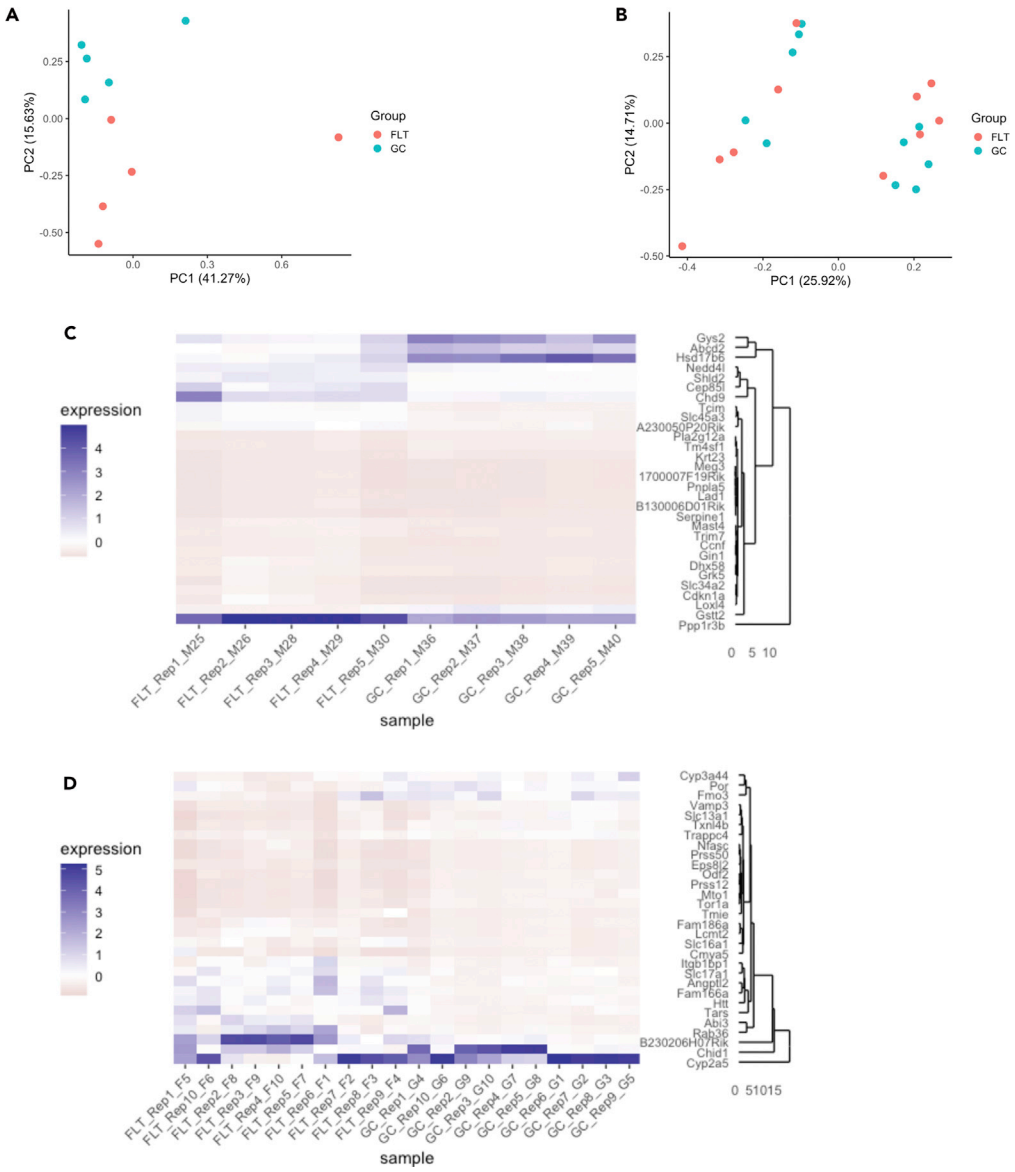


Figure 6. Global and differential gene expression in spaceflight versus ground control liver samples from GeneLab datasets

(A and 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 and 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.

according to ERCC spike-ins remains controversial among AWG members, and Munro et al. suggested its usage only for determining limit of detection of ratio (LODR), expression ratio variability, and measurement bias (Munro et al., 2014). For this reason, ERCC normalization remains optional in the GeneLab pipeline, and both kinds of DGE outputs are provided in the GeneLab database. In addition, ERCC spike-in could have two other usages. First, it allows us to evaluate whether normalization succeeded in removing

Table 3. Comparison of gene ontology in spaceflight versus ground control liver samples from GeneLab datasets

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

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.

systemic bias between libraries by using methods such as Rlog and VST when normalizing the spike-in RNAs along with all other genes. Second, most normalization methods of RNA-seq data assume that most genes are not differentially expressed toward one direction. Comparing spike-in measurements between libraries will help us to estimate the validity of this assumption.

A high number of biological replicates can increase certainty in the differentially expressed genes determined by the RCP. However, conducting experiments in spaceflight often limits the number of biological replicates that a researcher can include. Therefore, it is important 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 that 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 that 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 et al. 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 that 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 *Escherichia 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 RNA-seq 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). In addition, the pipeline cannot compensate for poor library preparation technique or inadequate sample size. Sample preservation protocols between datasets need to also be evaluated, because variations in sample preservation protocol could lead to poor correlation between studies that are otherwise identical (Lai 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 that 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 et al. 2017). This will allow PCR duplicates to be removed after mapping and before quantification.

In addition, 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 are 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. Instructions for installing packages using Conda are provided on Github (https://github.com/nasa/GeneLab_Data_Processing/tree/master/RNAseq/RNAseq_Tool_Instal). 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). Original data have been deposited to Mendeley Data: https://doi.org/10.17632/fv3kd6h7k4.1).

METHODS

All methods can be found in the accompanying transparent methods supplemental file.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2021.102361.

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, AMSB, ZZ, KSR, HF, WAdS, RB, and JMG refined this into a draft pipeline. EGO and AMSB wrote and validated the final processing scripts. EGO and AMSB wrote the original manuscript draft and generated figures. All authors reviewed and edited the final draft.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: September 8, 2020
Revised: October 30, 2020
Accepted: March 23, 2021
Published: April 23, 2021

REFERENCES

Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data. Babraham Bioinformatics (Babraham Institute).

Baruzzo, G., Hayer, K.E., Kim, E.J., Di Camillo, B., FitzGerald, G.A., and Grant, G.R. (2017). Simulation-based comprehensive benchmarking of RNA-seq aligners. Nat. Methods *14*, 135–139.

Berrios, D.C., Galazka, J., Grigorev, K., Gebre, S., and Costes, S.V. (2020). NASA GeneLab: interfaces for the exploration of space omics data. Nucleic Acids Res. https://doi.org/10.1093/nar/gkaa887.

Bray, N.L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. *34*, 525–527.

Castro-Wallace, S.L., Chiu, C.Y., John, K.K., Stahl, S.E., Rubins, K.H., McIntyre, A.B.R., Dworkin, J.P., Lupisella, M.L., Smith, D.J., Botkin, D.J., et al. (2017). Nanopore DNA sequencing and genome assembly on the International space station. Sci. Rep. *7*, 18022.

Chen, J., Bardes, E.E., Aronow, B.J., and Jegga, A.G. (2009). ToppGene suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Res. *37*, W305–W311.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M.W., Gaffney, D.J., Elo, L.L., Zhang, X., et al. (2016). A survey of best practices for RNA-seq data analysis. Genome Biol. *17*, 13.

Costa-Silva, J., Domingues, D., and Lopes, F.M. (2017). RNA-seq differential expression analysis: an extended review and a software tool. PLoS One *12*, e0190152.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

ENCODE Project Consortium, Snyder, M.P., Gingeras, T.R., Moore, J.E., Weng, Z., Gerstein, M.B., Ren, B., Hardison, R.C., Stamatoyannopoulos, J.A., Graveley, B.R., et al. (2020). Perspectives on ENCODE. Nature *583*, 693–698.

Ewels, P., Magnusson, M., Lundin, S., and Käller, M. (2016). MultiQC: Summarize analysis results for multiple tools and samples in a single report. Bioinformatics *32*, 3047–3048.

Functional Genomics Data Society (2012). MINSEQE: Minimum Information about a high-throughput SEQuencing Experiment (version 1.0). http://fged.org/projects/minseqe/.

Gentleman, R.C., Carey, V.J., Bates, D.M., Ben, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., et al. (2004). Bioconductor: open software development for computational Biology and bioinformatics. Genome Biol. *5*, R80.

Huber, W., Carey, V.J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B.S., Bravo, H.C., Davis, S., Gatto, L., Girke, T., et al. (2015). Orchestrating high-throughput genomic analysis with bioconductor. Nat. Methods *12*, 115–121.

Jain, M., Olsen, H.E., Paten, B., and Akeson, M. (2016). Erratum to: the Oxford nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biol. *17*, 256.

Jiang, L., Schlesinger, F., Davis, C.A., Zhang, Y., Li, R., Salit, M., Gingeras, T.R., and Oliver, B. (2011). Synthetic spike-in standards for RNA-seq experiments. Genome Research. https://doi.org/10.1101/gr.121095.111.

Krueger, F. (2019). Trim Galore: a wrapper around cutadapt and FastQC to consistently apply adapter and quality trimming to FastQ files, with extra functionality for RRBS data (version 0.6.5). https://github.com/FelixKrueger/TrimGalore.

Lai Polo, S.-H., Saravia-Butler, A.M., Boyko, V., Dinh, M.T., Chen, Y.-C., Fogle, H., Reinsch, S.S., Ray, S., Chakravarty, K., Marcu, O., et al. (2020). RNAseq analysis of rodent spaceflight experiments is confounded by sample collection techniques. iScience. https://doi.org/10.1101/2020.07.18.209775.

Langmead, B., Cole, T., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25.

Law, C.W., Alhamdoosh, M., Su, S., Dong, X., Tian, L., Smyth, G.K., and Ritchie, M.E. (2016). RNA-seq analysis is easy as 1-2-3 with Limma, Glimma and edgeR. F1000Res. *5*, https://doi.org/10.12688/f1000research.9005.3.

Liao, Y., Wang, J., Jaehnig, E.J., Shi, Z., and Zhang, B. (2019). WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. Nucleic Acids Res. *47*, W199–W205.

Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. BMC Bioinformatics *12*, 323.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics. https://doi.org/10.1093/bioinformatics/btp324.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. J. *17*, 10–12.

McIntyre, L.M., Lopiano, K.K., Morse, A.M., Amin, V., Oberg, A.L., Young, L.J., and Nuzhdin, S.V. (2011). RNA-seq: technical variability and sampling. BMC Genomics *12*, 293.

Mi, H., Muruganujan, A., and Thomas, P.D. (2013). PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. Nucleic Acids Res. *41*, D377–D386.

Munro, S.A., Lund, S.P., Pine, P.S., Binder, H., Clevert, D.-A., Conesa, A., Dopazo, J., Fasold, M., Hochreiter, S., Hong, H., et al. (2014). Assessing technical performance in differential gene expression experiments with external spike-in RNA control ratio mixtures. Nat. Commun. *5*, 5125.

Nguyen, T.-M., Shafi, A., Nguyen, T., and Draghici, S. (2019). Identifying significantly impacted pathways: a comprehensive review and assessment. Genome Biol. *20*, 203.

Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. Nat. Methods *14*, 417–419.

Raplee, I.D., Evsikov, A.V., and Marín de Evsikova, C. (2019). Aligning the aligners: comparison of RNA sequencing data alignment and gene

expression quantification tools for clinical breast cancer research. J. Personalized Med. *9*, https://doi.org/10.3390/jpm9020018.

Risso, D., Ngai, J., Speed, T.P., and Dudoit, S. (2014). Normalization of RNA-seq data using factor Analysis of control genes or samples. Nat. Biotechnol. *32*, 896–902.

Risso, D., Schwartz, K., Sherlock, G., and Dudoit, S. (2011). GC-content normalization for RNA-seq data. BMC Bioinformatics *12*, 480.

Roberts, R.J., Carneiro, M.O., and Schatz, M.C. (2013). The advantages of SMRT sequencing. Genome Biol. *14*, 405.

Rocca-Serra, P., Brandizi, M., Maguire, E., Sklyar, N., Taylor, C., Begley, K., Field, D., Harris, S., Hide, W., Hofmann, O., et al. (2010). ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level. Bioinformatics. https://doi.org/10.1093/bioinformatics/btq415.

Rutter, L., Barker, R., Bezdan, D., Cope, H., Costes, S.V., Degoricija, L., Fisch, K.M., Gabitto, M., Gebre, S., Giacomello, S., et al. (2020). A new era for space Life science: International standards for space omics processing (ISSOP). Patterns *1*, https://doi.org/10.1016/j.patter.2020.100148.

Sansone, S.A., Rocca-Serra, P., Field, D., Maguire, E., Taylor, C., Hofmann, O., Fang, H., Neumann, S., Tong, W., Amaral-Zettler, L., et al. (2012). Toward interoperable bioscience data. Nat. Genet. *44*, 121–126.

Schaarschmidt, S., Fischer, A., Zuther, E., and Hincha, D.K. (2020). Evaluation of seven different RNA-seq alignment tools based on experimental data from the model plant Arabidopsis thaliana. Int. J. Mol. Sci. *21*, https://doi.org/10.3390/ijms21051720.

Schurch, N.J., Schofield, P., Gierliński, M., Cole, C., Sherstnev, A., Singh, V., Wrobel, N., Gharbi, K., Simpson, G., Owen-Hughes, T., et al. (2016). How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? RNA. https://doi.org/10.1261/rna.053959.115.

Scott, R.T., Grigorev, K., Mackintosh, G., Gebre, S.G., Mason, C.E., Del Alto, M.E., and Costes, S.V. (2020). Advancing the integration of Biosciences data sharing to further enable space exploration*33* (Cell Rep.).

Smith, T., Heger, A., and Sudbery, I. (2017). UMI-tools: modeling sequencing errors in unique molecular identifiers to improve quantification accuracy. Genome Res. *27*, 491–499.

Soneson, C., Love, M., and Robinson, M. (2015). Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences, [version 2; Peer Review: 2 Approved]. F1000Res. *4*, 1521.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set Enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U S A *102*, 15545–15550.

Szklarczyk, D., Gable, A.L., Lyon, D., and Junge, A. (2019). STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids. https://academic.oup.com/nar/article-abstract/47/D1/D607/5198476.

Teng, M., Love, M.I., Davis, C.A., Djebali, S., Dobin, A., Graveley, B.R., Li, S., Mason, C.E., Olson, S., Pervouchine, D., et al. (2016). A benchmark for RNA-seq quantification pipelines. Genome Biol. *17*, 74.

Thomas, P.D. (2003). PANTHER: a library of protein families and subfamilies indexed by function. Genome Res. https://doi.org/10.1101/gr.772403.

Väremo, L., Nielsen, J., and Nookaew, I. (2013). Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods. Nucleic Acids Res. *41*, 4378–4391.

Williams, C.R., Baccarella, A., Parrish, J.Z., and Kim, C.C. (2016). Trimming of sequence reads alters RNA-seq gene expression estimates. BMC Bioinformatics *17*, 103.

Wu, D.C., Yao, J., Ho, K.S., Lambowitz, A.M., and Wilke, C.O. (2018). Limitations of alignment-free tools in total RNA-seq quantification. BMC Genomics *19*, 510.

Yang, C., Wu, P.-Y., Tong, L., Phan, J.H., and Wang, M.D. (2015). The impact of RNA-seq aligners on gene expression estimation. ACM BCM *2015*, 462–471.

iScience, Volume *24*

Supplemental information

NASA GeneLab RNA-seq consensus pipeline: standardized processing of short-read RNA-seq data

Eliah G. Overbey, Amanda M. Saravia-Butler, Zhe Zhang, Komal S. Rathi, Homer Fogle, Willian A. da Silveira, Richard J. Barker, Joseph J. Bass, Afshin Beheshti, Daniel C. Berrios, Elizabeth A. Blaber, Egle Cekanaviciute, Helio A. Costa, Laurence B. Davin, Kathleen M. Fisch, Samrawit G. Gebre, Matthew Geniza, Rachel Gilbert, Simon Gilroy, Gary Hardiman, Raúl Herranz, Yared H. Kidane, Colin P.S. Kruse, Michael D. Lee, Ted Liefeld, Norman G. Lewis, J. Tyson McDonald, Robert Meller, Tejaswini Mishra, Imara Y. Perera, Shayoni Ray, Sigrid S. Reinsch, Sara Brin Rosenthal, Michael Strong, Nathaniel J. Szewczyk, Candice G.T. Tahimic, Deanne M. Taylor, Joshua P. Vandenbrink, Alicia Villacampa, Silvio Weging, Chris Wolverton, Sarah E. Wyatt, Luis Zea, Sylvain V. Costes, and Jonathan M. Galazka

Supplemental Information

Transparent Methods

The tools used in the consensus pipeline are documented in Supplemental Table 4: Pipeline Tools and Links [Table S4: "Pipeline Tools and Links, Related to Transparent Methods"]. 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). There are currently over 80 RNA-seq datasets available [Table S1: "GeneLab RNA-Seq Datasets, Related to Transparent Methods"].

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: "Sample Names, Related to Figure 6"]. 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).

Supplemental Figures

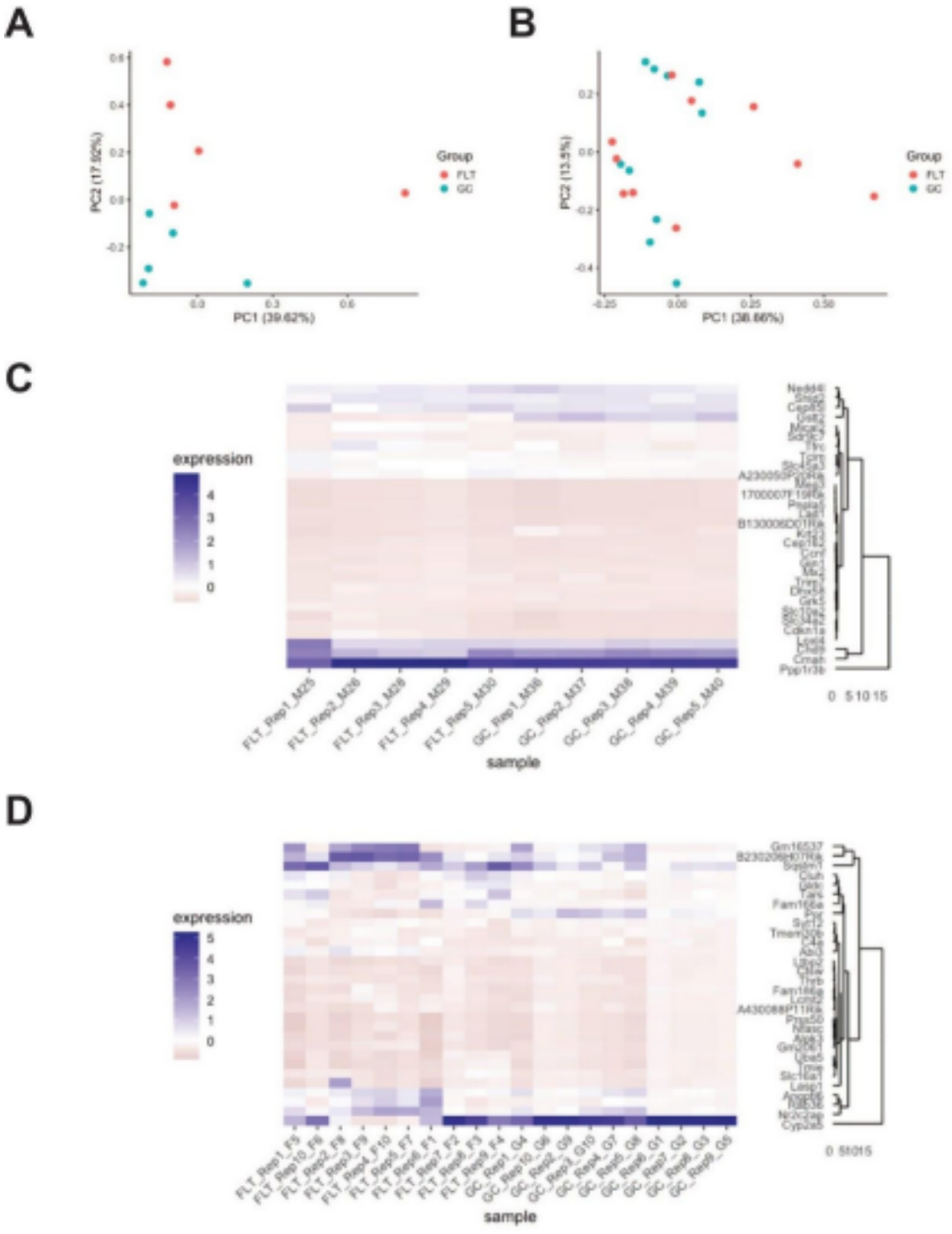

A
PC2 (17.92%)
PC1 (39.62%)
Group
FLT
GC
B
PC2 (13.5%)
PC1 (38.66%)
Group
FLT
GC
C
expression
sample
FLT_Rep1_M25
FLT_Rep2_M26
FLT_Rep3_M28
FLT_Rep4_M29
FLT_Rep5_M30
GC_Rep1_M36
GC_Rep2_M37
GC_Rep3_M38
GC_Rep4_M39
GC_Rep5_M40
D
expression
sample

Figure S1 (Related to Figure 6). Global and differential gene expression in ERCC-normalized 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 ERCC-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 ERCC-normalized differential expression tables for each respective dataset on the NASA GeneLab Data Repository. 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.

Supplemental Tables

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	109, 13	0, 11	0, 0
GLDS-245	166, 0	81, 0	1, 0

Table S6 (Related to Table 3). Comparison of gene ontology in ERCC-normalized 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 ERCC-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.