EMBRY-RIDDLE
Aeronautical University
SCHOLARLY COMMONS

Publications

1-12-2021

Knowledge Network Embedding of Transcriptomic Data from Spaceflown Mice Uncovers Signs and Symptoms Associated with Terrestrial Diseases

Amber M. Paul
NASA Ames Research Center, Universities Space Research Association, paula6@erau.edu

Charlotte A. Nelson
University of California, San Francisco

Ana Uriarte Acuna
NASA Ames Research Center

Ryan T. Scott
NASA Ames Research Center

Atul J. Butte
University of California, San Francisco

See next page for additional authors

Follow this and additional works at: https://commons.erau.edu/publication

 Part of the Biomedical Informatics Commons, and the Data Science Commons

Scholarly Commons Citation

Nelson, C.A.; Acuna, A.U.; Paul, A.M.; Scott, R.T.; Butte, A.J.; Cekanaviciute, E.; Baranzini, S.E.; Costes, S.V. Knowledge Network Embedding of Transcriptomic Data from Spaceflown Mice Uncovers Signs and Symptoms Associated with Terrestrial Diseases. Life 2021, 11, 42. https://doi.org/10.3390/life11010042

This Article is brought to you for free and open access by Scholarly Commons. It has been accepted for inclusion in Publications by an authorized administrator of Scholarly Commons. For more information, please contact commons@erau.edu.

Authors

Amber M. Paul, Charlotte A. Nelson, Ana Uriarte Acuna, Ryan T. Scott, Atul J. Butte, Egle Cekanaviciute, and Sergio E. Baranzini

This article is available at Scholarly Commons: https://commons.erau.edu/publication/1974

life

MDPI

Article

Knowledge Network Embedding of Transcriptomic Data from Spaceflown Mice Uncovers Signs and Symptoms Associated with Terrestrial Diseases

Charlotte A. Nelson [1], Ana Uriarte Acuna [2,3], Amber M. Paul [2,4], Ryan T. Scott [2,3], Atul J. Butte [5,6], Egle Cekanaviciute [2], Sergio E. Baranzini [1,5,7,*] and Sylvain V. Costes [2,*]

1 Integrated Program in Quantitative Biology, University of California San Francisco, San Francisco, CA 94143, USA; Charlotte.Nelson@ucsf.edu
2 Space Biosciences Division, NASA Ames Research Center, Moffett Field, CA 94035, USA; ana.e.uriarteacuna@nasa.gov (A.U.A.); amber.m.paul@nasa.gov (A.M.P.); ryan.t.scott@nasa.gov (R.T.S.); egle.cekanaviciute@nasa.gov (E.C.)
3 KBR, NASA Ames Research Center, Moffett Field, CA 94035, USA
4 NASA Postdoctoral Program, Universities Space Research Association (USRA), Mountain View, CA 94043, USA
5 Bakar Computational Health Sciences Institute, University of California San Francisco, San Francisco, CA 94143, USA; atul.butte@ucsf.edu
6 Department of Pediatrics, University of California San Francisco, San Francisco, CA 94143, USA
7 Weill Institute for Neuroscience, Department of Neurology, University of California San Francisco, San Francisco, CA 94143, USA
* Correspondence: Sergio.Baranzini@ucsf.edu (S.E.B.); sylvain.v.costes@nasa.gov (S.V.C.)

Abstract: There has long been an interest in understanding how the hazards from spaceflight may trigger or exacerbate human diseases. With the goal of advancing our knowledge on physiological changes during space travel, NASA GeneLab provides an open-source repository of multi-omics data from real and simulated spaceflight studies. Alone, this data enables identification of biological changes during spaceflight, but cannot infer how that may impact an astronaut at the phenotypic level. To bridge this gap, Scalable Precision Medicine Oriented Knowledge Engine (SPOKE), a heterogeneous knowledge graph connecting biological and clinical data from over 30 databases, was used in combination with GeneLab transcriptomic data from six studies. This integration identified critical symptoms and physiological changes incurred during spaceflight.

Keywords: spaceflight; knowledge graph; transcriptomics

check for updates

Citation: Nelson, C.A.; Acuna, A.U.; Paul, A.M.; Scott, R.T.; Butte, A.J.; Cekanaviciute, E.; Baranzini, S.E.; Costes, S.V. Knowledge Network Embedding of Transcriptomic Data from Spaceflown Mice Uncovers Signs and Symptoms Associated with Terrestrial Diseases. *Life* **2021**, *11*, 42. https://doi.org/10.3390/life11010042

Received: 10 December 2020
Accepted: 4 January 2021
Published: 12 January 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

NASA recognizes five main hazards of spaceflight to human health, including altered gravity (microgravity and hypergravity), ionizing radiation, isolation/confinement, hostile/closed environment, and distance from Earth. These health risks caused by the space environment resemble multiple disorders found on Earth, including muscle atrophy and bone loss, cardiovascular deconditioning, immune dysfunction, and central nervous system deficits [1]. Therefore, repurposing current FDA-approved treatments for issues that arise during spaceflight could significantly reduce the time needed to develop new therapeutics and limit their side effects.

Since its establishment in 2015, NASA GeneLab [2] has become a prominent open-source repository of data from real and simulated spaceflight studies. This platform has enabled computational analysis of multi-omics data, visualization of results, and integration with descriptive metadata, such as environmental data (e.g., space radiation dosimetry). GeneLab has already supported dozens of published studies, created a global collaboration to develop uniform standards for spaceflight-omics [3], and resulted in new space biology discoveries [4,5]. However, it has not yet been possible to use NASA GeneLab to combine

and compare space and terrestrial data. Such capability would be a major advancement in fundamental spaceflight biology and its applications, including identifying new targets or repurposing terrestrial therapeutics for spaceflight countermeasures.

NASA GeneLab is planning to set up a portal dedicated to computational modeling that enables comparisons between datasets in addition to already existing data input, query, analysis, and visualization capabilities. Knowledge graphs (KGs) would be a suitable approach to facilitate this goal by unifying disparate datasets into a human queryable framework. KGs have already been widely adopted in biomedical research to unravel the complex relationship between biological changes and disease phenotypes [6–10].

Specifically, a new massive UCSF-based KG database, the Scalable Precision Medicine Oriented Knowledge Engine (SPOKE) has transformed structured data from over 30 human biomedical databases (-omics, chemical structures, molecular and cellular responses, physiological data including e.g., patient symptoms and drug side effects, etc.) into a KG with almost 400,000 nodes of 12 types and over 10 million edges of 32 types [11,12]. Therefore, SPOKE has the potential to be combined with the NASA GeneLab modeling portal, expanding it to link terrestrial biomedical sciences to space biosciences research and space medicine.

In this study, we integrated data from six different NASA GeneLab datasets in SPOKE to enable normalization that highlighted new nodes defining systems and effects that are known to be relevant for space travel but would have been impossible to uncover without using SPOKE (workflow Figure 1a). These results suggest that SPOKE can be utilized to gain a deeper biological understanding of the health hazards associated with spaceflight and provide the proof of concept for its broader utilization to integrate space and terrestrial biological data.

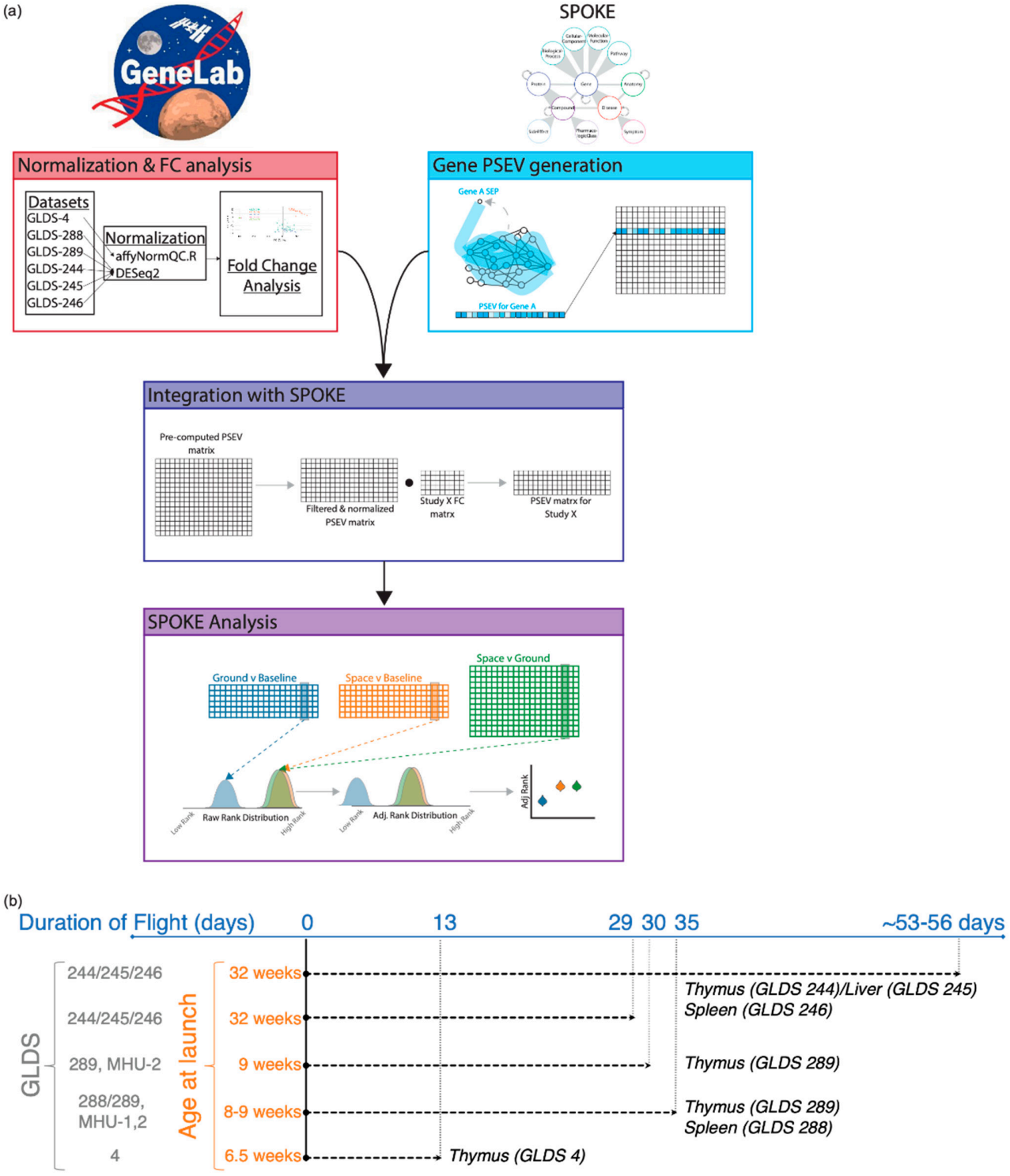

Figure 1. Analysis workflow and summary of experimental conditions across GeneLab datasets used for the analysis. (**a**) Workflow depicting the different stages of the analysis. (**b**) Datasets GLDS-4, -244, -245, and -246 used C57BL/6NTac mice. Datasets GLDS-288 and -289 used C57BL/6J mice for spaceflight and both C57BL/6J and Charles River mice for ground controls.

2. Materials and Methods

2.1. GeneLab Data Processing and Analysis

Gene expression data were downloaded from the NASA GeneLab repository (https://genelab-data.ndc.nasa.gov/), datasets GLDS-4, GLDS-244, GLDS-245, GLDS-246, GLDS-288, and GLDS-289. All data had been processed and analyzed using standard NASA GeneLab techniques detailed below. Matched flight/live animal return versus ground control data was used for analysis.

Previously, raw data were processed separately for each dataset by the NASA GeneLab data processing team. For datasets containing RNA Sequencing (RNA-Seq) assays (GLDS-244, GLDS-245, GLDS-246, GLDS-288, GLDS-289), raw FASTQ files were assessed for the percentage of rRNA using HTStream SeqScreener (version 1.1.0 for GLDS-244, GLDS-245, GLDS-246 and version 1.3.1 for GLDS-288, GLDS-289; https://s4hts.github.io/HTStream/) and filtered using Trim Galore! (version 0.6.4; https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) [13]. Raw and trimmed fastq file quality was evaluated with FastQC [14] (version 0.11.9). MultiQC [15] (version 1.8 for GLDS-244, GLDS-245, GLDS-246 and version 1.9 for GLDS-288, GLDS-289) was used to generate MultiQC reports. Mus musculus STAR [16] and RSEM [17] references were built using STAR (version 2.7.1a for GLDS-244, GLDS-245, GLDS-246 and version 2.7.4a for GLDS-288, GLDS-289) and RSEM (version 1.3.1), respectively, genome version mm10-GRCm38 (Mus_musculus.GRCm38.dna.toplevel.fa), and the following gtf annotation file: Mus_musculus.GRCm38.96.gtf. Trimmed reads were aligned to the Mus musculus STAR reference with STAR (version 2.7.3a for GLDS-244, GLDS-245, GLDS-246 and version 2.7.4a for GLDS-288, GLDS-289) and aligned reads were quantified using RSEM (version 1.3.1 from the NASA GeneLab repository).

Data representing the quantitative analysis of gene expression for each dataset was downloaded from the NASA GeneLab repository, where it had been previously analyzed and imported into R [18] (version 3.6.3). It was then combined to create a gene counts table containing the data for all samples of every dataset. For GLDS-244, GLDS-245, and GLDS-246, only non-ERCC (External RNA Controls Consortium [19], i.e., a spike-in mixture used for normalization) genes were used. Data was normalized with DESeq2 [20] (version 1.26.0). A principal component analysis was performed using prcomp (stats version 3.6.3) and plotted using plotly [21] (version 4.9.2.1). For datasets containing DNA microarray assays (GLDS-4) raw CEL files were read in and normalized using an in-house R script as described [22].

To quantify overlapping pathways between GLDS-244, -245, and -246, Entrez Gene IDs of genes that showed a significant difference ($p < 0.05$) between 29-day flight/live animal return and ground controls were used as the input to Molecular Signatures Database v7.2, GeneOntology [23–25] (GO) gene sets. (GO biological process, GO cellular component, GO molecular function). The top 50 statistically significant gene sets were compared to identify overlaps. The same approach was applied to quantify the overlapping gene sets between GLDS-288 and -289.

2.2. Scalable Precision Medicine Oriented Knowledge Engine

Scalable Precision Medicine Oriented Knowledge Engine (SPOKE) [11,12] is a population level heterogeneous knowledge graph. SPOKE was generated by unifying over 30 publicly available databases. Currently, SPOKE contains almost 400,000 nodes of 12 types (*Anatomy*, *BiologicalProcess*, *CellularComponent*, *Compound*, *Disease*, *Gene*, *MolecularFunction*, *Pathway*, *PharmacologicalClass*, *Protein*, *SideEffect*, and *Symptom*). These nodes are connected by 32 types of biologically meaningful edges ($n > 10$ million).

2.3. Gene-Specific Propagated SPOKE Entry Vectors

Propagated SPOKE Entry Vectors (PSEVs) are generated using a modified version of topic-specific page rank to learn and embed the importance of each node in SPOKE for a given restart node or set of nodes [26,27]. These restart nodes, called SPOKE Entry Points (SEPs), are any concept in the input data that overlaps with a node(s) in SPOKE [28]. In

this analysis, the SEPs were the mouse genes that have homologs to the human *Gene* nodes in SPOKE. A Gene PSEV was produced by allowing a random walker to traverse the edges in SPOKE and then forcing them to restart at a specific *Gene* SEP. The forced restart ensures that the walker will spend the majority of time on nodes that are important for that *Gene*. The significance of each node is then stored in an element of the PSEV such that the length of the PSEV is equal to the number of nodes in SPOKE (n = 389,297). Code used to generate the data in this manuscript is available at (https://doi.org/10.5281/zenodo.4408540).

2.4. Integrating Gene Expression Data and PSEVs

For each study, the $-\log_2$ fold-change (FC) mouse gene expression data was mapped to the human gene nodes in SPOKE. The homologous mapping between species was achieved using HomoloGene IDs [29]. If multiple mouse genes mapped to a single human gene, then the average FC was used. Additionally, some studies contained multiple comparisons between space and ground or baseline control mice. An example of this is the study GLDS-244 that compared mice at two space-time points (day-29 and days 53–56). In these instances, genes were removed if the FC comparisons were not in the same direction (i.e., if space versus ground day-29 had a positive FC and days-53–56 had a negative FC). This filter focuses on the data set of genes that remain consistent during space travel.

After genes were mapped and filtered for a given study, the pre-computed PSEVs for the remaining genes were extracted. This PSEV matrix was z-score normalized and then ranked such that the most important node in a given PSEV was equal to the number of nodes in SPOKE (n = 389,297) and the least important was ranked one. Then for each comparison, the filtered PSEV matrix was adjusted using the FCs. This was accomplished by taking the product of a single column in the FC matrix and the filtered normalized PSEV matrix. It is necessary for the rows (genes) of the filtered normalized PSEV matrix to be in the same order as the rows in the FC matrix. Next, each column (node) in the adjusted PSEV-matrix was summed resulting in a vector in which each element or position corresponded to a node in SPOKE (length = 389,297). Each node was then ranked as before (with the highest value in the vector ranked 389,297). In practice, this was achieved by taking the dot product of the filtered FC matrix (transposed) and the filtered normalized PSEV matrix and then ranking the resulting matrix.

2.5. Finding Significant Spoke Nodes

The PSEV comparisons from the six studies were pooled together and separated into three groups (Ground vs. Baseline, Space vs. Baseline, and Space vs. Ground). Welch's t-test was used to evaluate whether the distribution of ranks of a given node in the Ground vs. Baseline group was significantly different from that in either Space vs. Baseline or Space vs. Ground (Table S1, in the Supplementary Materials). Top nodes, those that were ranked significantly different in either space travel comparisons (Space vs. Baseline and Space vs. Ground) than in Ground vs. Baseline, were identified using the p-values from the Welch's t-test. Since 159,374 nodes had a p-value < 0.025 in either or both space travel comparisons, top nodes were further filtered by selecting the most significant 2.5% of each node type for Space vs. Ground and/or Space vs. Baseline (n = 15,801; 4.1%).

2.6. Retracing Paths from Input Gene to SPOKE Node

A high correlation between a gene's FC and the rank of a specific node suggests that the gene FC is at least partially responsible for the prioritization of the node within a PSEVs. The correlation was calculated between genes (present in >20% of FC comparisons; n = 7567) and a set of top *Anatomy*, *BiologicalProcess*, *CellularComponent*, *MolecularFunction*, *Pathway*, and *Symptom* nodes (n = 22). Next, paths were found between genes that had a high correlation (correlation > 0.6) and the set of top nodes. Gene-node pairs were then filtered to only include pairs that had the same sign (positive gene expression and positive Welch t-statistic). Then, in order to visualize paths between gene-node pairs, paths were

filtered to have a maximum of three edges and less than 100 possible combinations of nodes within the path. This left over 17,000 gene-node pairs and 234,000 possible paths.

The paths shown were selected based on their simplicity and the FC of the original genes (Figure S1, in the Supplementary Materials). The *p*-values, derived when calculating the FCs used as input for PSEV creation, were combined for Ground vs. Baseline and the space travel groups (Space vs. Baseline and Space vs. Ground together) using Stouffer's method [30]. Each gene FC was judged on whether the average space travel group had a combined *p*-value that was more significant than Ground v Baseline (Figure S1, y-axis). Then the Welch's *t*-test was used to determine whether the FC distributions were significantly different between groups. Space vs. Baseline and Space vs. Ground distributions were compared to the Ground vs. Baseline separately and then averaged (Figure S1, x-axis).

3. Results

3.1. Transcriptional Profiling of Mice after Space Flight

Here we conducted a meta-analysis of six independent transcriptomic datasets (GLDS-4, -244, -245, -246, -288, and -289) from experimental mice obtained during four different spaceflight missions (STS-118, TCU (SpaceX-9), MHU-2 (Space X-12), and RR-6 (SpaceX-13)), at five time- points of collection (13-, 29-, 30-, and 35-days live animal return (LAR); and 53–56 days (ISS terminal)), on the International Space Station (ISS) (Figure 1b and Table 1). While experiments varied in their design (i.e., duration of flight, age at launch, the genotype of mice, transcriptomic platform, time of collection), the objective of these experiments was to identify changes in gene expression induced by spaceflight in three different immune-related organs—thymus (primary lymphoid organ), spleen (secondary lymphoid organ), and liver (immune-associated/digestive organ, with lymphatic cells playing a role in its responses to injury [31]).

These sample sets were selected to include multiple immune-associated organs (thymus, spleen, liver) collected from the same space-flown mice as well as between mice flown on different missions to increase sample diversity and to include RNA sequencing and microarray as two different sequencing methods to show that both can be used as inputs to SPOKE.

After data normalization, the principal component analysis revealed a strong separation of samples by mission and tissues (Figure 2a). These findings are unsurprising, given that these variables are confounding factors of different missions/collections. However, we also observed that samples from the same time point of mission/collection from two different experiments clustered together, suggesting that some biological effects were captured. When PCA was used to plot samples from similar experimental conditions (space-flown, ground, and baseline from the same RR-6 mission), no obvious separation between samples obtained during flight, baseline, and the ground was observed (Figure 2b).

Differentially expressed genes were identified in the thymus, liver, and spleen in space-flown mice vs. ground controls after live animal return from the RR-6 (SpaceX-13) mission. Furthermore, using the differentially expressed genes as an input to pathway analysis (by a hypergeometric test) showed a number of statistically significant biological functions dysregulated by space flight in the thymus, liver, and spleen, including some that overlapped between the tissues (Figure 2c). While some gene sets were tissue-specific, nine of them were shared among the three tissues, including apoptosis, cell metabolic process, and cell membrane integrity (Figure 2d).

Table 1. Descriptive metadata for each NASA GLDS dataset analyzed by the Scalable Precision Medicine Oriented Knowledge Engine (SPOKE).

	GeneLab Study	Tissue	Sequencing Type	Strain	Mission/Flight	Flight Duration
A.	GLDS-4	Thymus	Microarray	C57BL/6NTac	STS-118	13-days (12.76 day)
	GLDS-244	Thymus	RNA-sequencing	C57BL/6NTac	RR-6 (SpaceX-13)	29-days (n = 9, LAR); 53–56-days (n = 10, ISS terminal)
	GLDS-245	Liver	RNA-sequencing	C57BL/6NTac	RR-6 (SpaceX-13)	29-days (n = 9, LAR); 53–56-days (n = 10, ISS terminal)
	GLDS-246	Spleen	RNA-sequencing	C57BL/6NTac	RR-6 (SpaceX-13)	29-days (n = 9, LAR); 53–56-days (n = 10, ISS terminal)
	GLDS-288	Spleen	RNA-sequencing	C57BL/6J (flight); Charles River Laboratories Japan (GC)	TCU (SpaceX-9)	35-days
	GLDS-289	Thymus	RNA-sequencing	C57BL/6J (flight); Charles River Laboratories Japan (GC)	TCU (SpaceX-9, MHU-1; SpaceX-12, MHU-2)	35-days MHU-1; 30-days MHU-2

	GeneLab Study	Age at Initiation	Age at Euthanasia	Sex	Sample Size (n/Cohort)	Controls	Collection Location
B.	GLDS-4	~6.5-weeks	8-weeks	n/a	FLT (n = 4); GC (n = 4)	Synchronous Ground Controls (GC)	Ground post-flight
	GLDS-244	32-weeks	36-weeks LAR; 44-weeks ISS terminal; 36-weeks LAR/ISS terminal Baseline GC; 41-weeks LAR GC; 44-weeks ISS Terminal GC	Female	LAR (n = 9); ISS terminal (n = 10); Baseline LAR (n = 10); Baseline ISS Terminal (n = 9); LAR GC (n = 9); ISS Terminal GC (n = 10)	Baseline (LAR, ISS terminal); synchronous GC (LAR, ISS terminal)	4-days post-flight (LAR); 53–56-day In-flight (ISS terminal)
	GLDS-245	32-weeks	36-weeks LAR; 44-weeks ISS terminal; 36-weeks LAR/ISS terminal Baseline GC; 41-weeks LAR GC; 44-weeks ISS Terminal GC	Female	LAR (n = 9); ISS terminal (n = 10); Baseline LAR (n = 10); Baseline ISS Terminal (n = 9); LAR GC (n = 9); ISS Terminal GC (n = 10)	Baseline (LAR, ISS terminal); synchronous GC (LAR, ISS terminal)	4-days post-flight (LAR); 53–56-day In-flight (ISS terminal)
	GLDS-246	32-weeks	36-weeks LAR; 44-weeks ISS terminal; 36-weeks LAR/ISS terminal Baseline GC; 41-weeks LAR GC; 44-weeks ISS Terminal GC	Female	LAR (n = 9); ISS terminal (n = 10); Baseline LAR (n = 10); Baseline ISS Terminal (n = 9); LAR GC (n = 9); ISS Terminal GC (n = 10)	Baseline (LAR, ISS terminal); synchronous GC (LAR, ISS terminal)	4-days post-flight (LAR); 53–56-day In-flight (ISS terminal)
	GLDS-288	8-weeks	12-weeks	Male	Spaceflight (MG, n = 3); Spaceflight w/centrifugation (AG, n = 3); Synchronous (GC, n = 3)	Spaceflight w/centrifugation; Synchronous GC	Ground post-flight
	GLDS-289	8-weeks MHU-1; 9-weeks MHU-2	12-weeks	Male	Spaceflight (MG, MHU-1, n = 3); Spaceflight w/centrifugation (AG, MHU-1, n = 3); Synchronous (GC, MHU-1, n = 3); Spaceflight (MG, MHU-2, n = 3); Spaceflight w/centrifugation (AG, MHU-2, n = 3); Synchronous (GC, MHU-2, n = 3)	Spaceflight w/centrifugation (AG, MHU-1); Synchronous (GC, MHU-1); Spaceflight w/centrifugation (AG, MHU-2); Synchronous (GC, MHU-2)	Ground post-flight

"GC" denotes ground control, "RR" denotes rodent research, "TCU" denotes transportation case units.

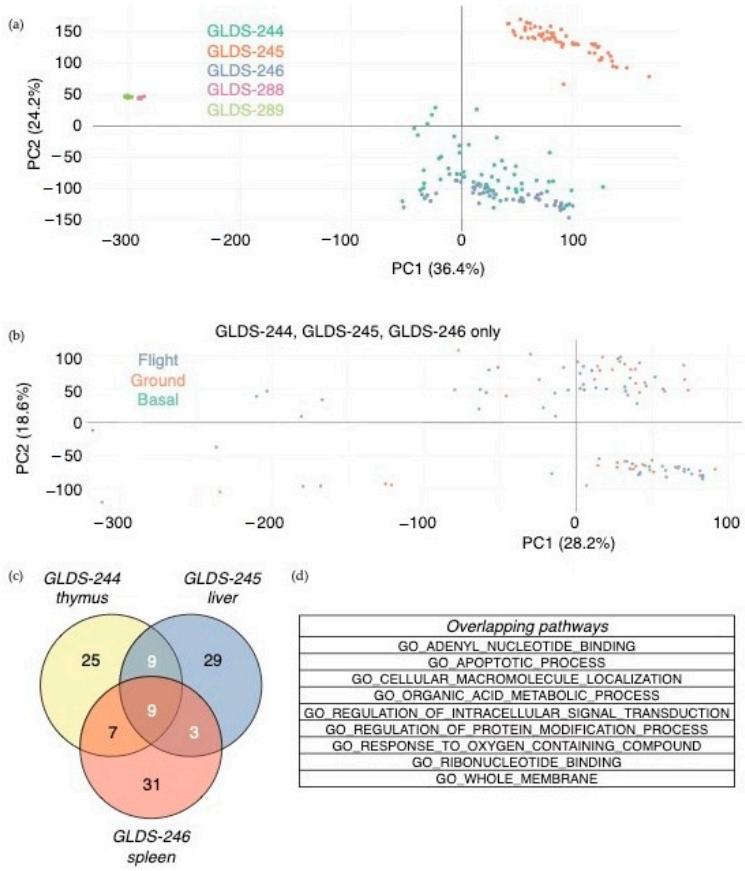

Figure 2. Transcriptomic analysis of spaceflight-associated changes in gene expression. (**a**) Principal component analysis of all samples, colored by the dataset. (**b**) Principal component analysis of datasets GLDS-244, -245, and -246, colored by flight condition. (**c**,**d**) Overlapping gene sets between datasets GLDS-244, 245, and 246 out of the top 50 Gene Ontology gene sets using significantly differently expressed genes ($p < 0.05$) between flight and ground conditions, live animal return after 29 days on the ISS. Venn diagram showing overlapping gene sets between datasets (**c**) and the list of gene sets overlapping between all three datasets (**d**). Three out of the top 50 gene ontology (GO) gene sets overlapped between datasets GLDS-288 and -289, none of which overlapped with GLDS-244, -245, and -246.

3.2. Fold-Change Enhanced Propagated SPOKE Entry Vectors

While established methods of transcriptional profiling can inform about dysregulated molecular pathways, they provide little insight into higher-order phenotypes, such as associated signs and symptoms of disease. Using SPOKE, a KG that integrates information of both biological and clinical databases, it is possible to score every node of the graph as a function of the "information flow" elicited by a defined set of quantitative inputs. SPOKE leverages the complexity of the hierarchical organization of complex organisms to identify nodes with shared information flow (regardless of whether the input itself was significant or not).

Gene-specific Propagated SPOKE Entry Vectors (PSEVs) were generated from the selected GeneLab studies prior to integrating gene expression results with SPOKE [11,12]. Each gene-specific PSEV was created using a modified version of topic-specific page rank [26,27] in which the random walker was forced to restart at the corresponding *Gene* node in SPOKE (See Methods, Figure 3a). This focused the random walker on nodes that were the most important for a given node (in this case, *Gene* node since the input is gene expression). The amount of time a random walker spent on a node was then stored in a defined element (position within) of the PSEV vector. All PSEVs were then stored in the pre-computed PSEV matrix. For each gene expression study, the pre-computed

PSEV matrix was filtered and normalized to match the genes within the study (Figure 3b; Methods). The dot product was then used with the normalized PSEV matrix and the $-\log_2$ fold-change (FC) to produce the PSEVs for that study. After PSEVs were computed for each study, they were pooled and separated into specific experimental groups to enable meaningful comparisons to test the hypothesis that spaceflight alters gene expression (Ground vs. Baseline, Space vs. Baseline, and Space vs. Ground) (Figure 3c).

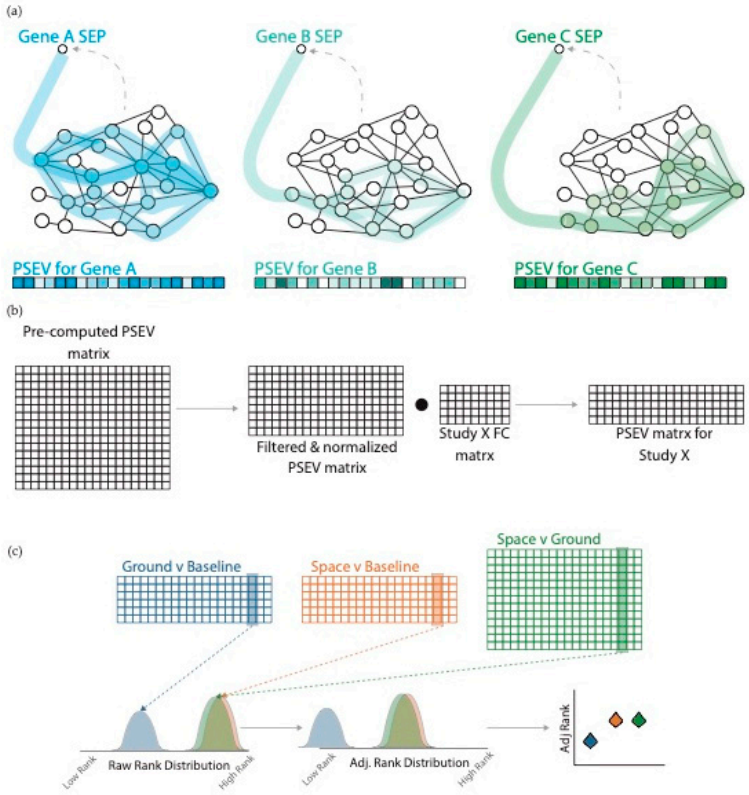

Figure 3. Propagated SPOKE Entry Vectors (PSEVs) using gene expression fold-change (FC). (**a**) PSEVs were pre-computed for all SPOKE genes. For each gene, the random walker was forced to restart at that gene (probability of random jump = 0.1). After PSEVs were finished they were stored in the pre-computed PSEV matrix. (**b**) For each study, the pre-computed PSEV matrix was filtered and normalized. Then the dot product was taken between the normalized matrix and the FC matrix to generate the PSEV matrix for that study. (**c top**) The PSEV matrices for each study were pooled together and separated into groups: Ground vs. Baseline (blue), Space vs. Baseline (yellow), and Space vs. Ground (green). (**c bottom**) The distributions of the node ranks were adjusted using the mean Ground vs. Baseline rank.

Each element in a PSEV corresponds to a single node in SPOKE. Therefore, it is possible to determine the overall significance of a node for spaceflight by evaluating the differential distribution of node ranks in the PSEV. Welch's *t*-test [32] was utilized to compare a node's rank distribution in the Ground vs. Baseline to that in either Space vs. Baseline or Space vs. Ground (Supplementary Table S1).

Strikingly, nodes that are known to be relevant for space travel such as space motion sickness (*Symptom*), regulation of blood vessel diameter (*BiologicalProcess*), taste receptor complex (*CellularComponent*), Vitamin D (calciferol) metabolism (*Pathway*), and sympathetic nervous system (*Anatomy*) scored among the top 5% of nodes (top 2.5% per type for Space vs. Baseline and/or Space vs. Ground). Figure 4 shows violin plots from a select set of nodes (n = 22) in SPOKE that had significantly different ranks in spaceflight (Space vs. Baseline and/or Space vs. Ground) compared to Ground vs. Baseline. From these, 11 correspond to symptoms (pink boxed violin charts, Figure 4a), five to gene ontology/pathway concepts

(teal boxed violin charts, Figure 4b–d), and six to anatomical regions (green boxed violin charts, Figure 4e). Violin plots for each category, sub-networks demonstrate how the gene expression results drive information from these 22 nodes. Among the other biological top nodes were nodes that reflected the results of the original studies such as those related to t-cell activity, regulation of stress, and TGFβ1 [1,33].

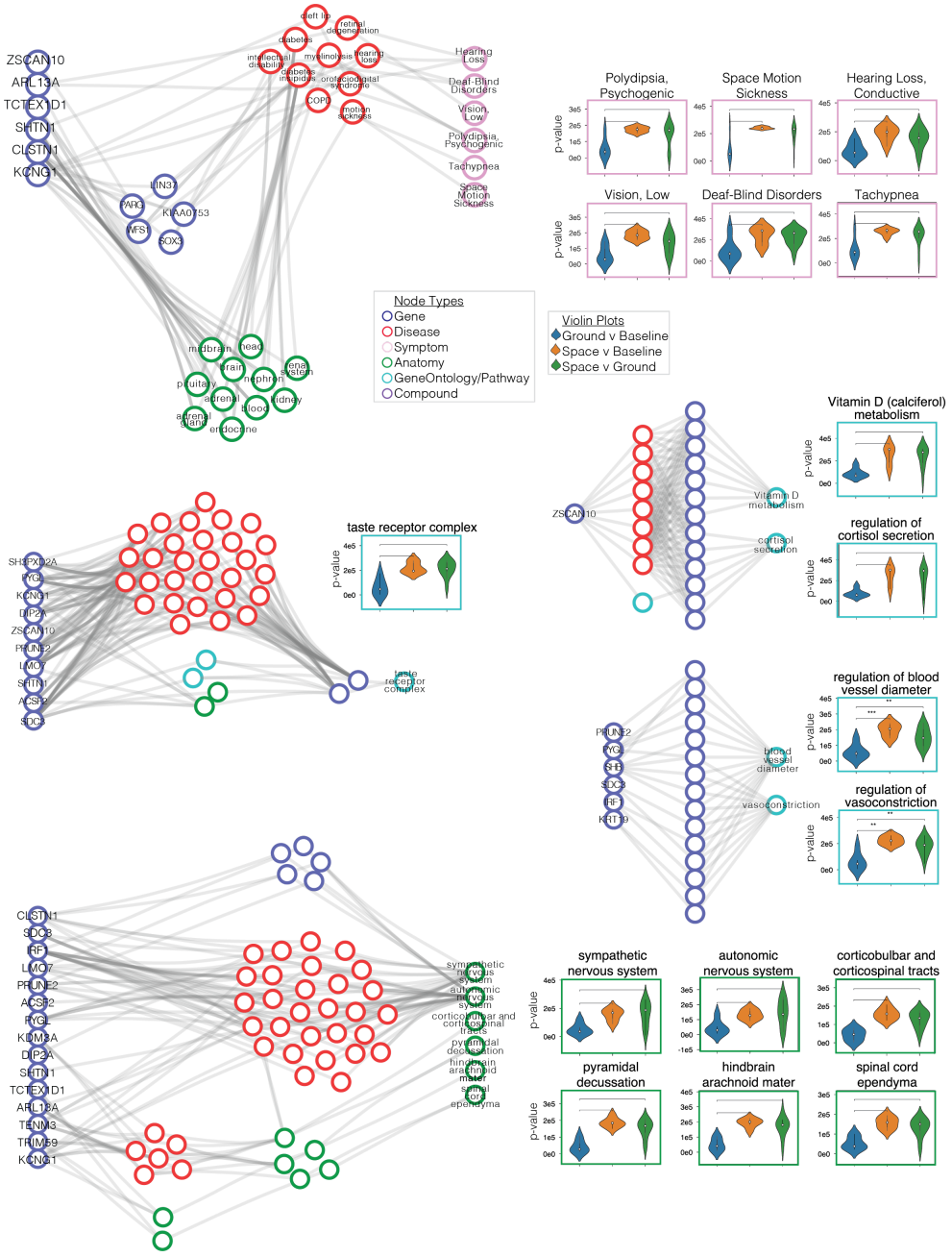

Figure 4. Retracing paths between genes and top nodes. Gene expression FC values drive information flow to nodes in SPOKE. (**a**–**e**) Paths were traced between genes that were partially responsible for pushing information to a set of significant nodes (n = 22). These paths were shown for (**a**) 10 *Symptom* nodes, (**b**) *taste receptor complex* (*CellularComponent*), (**c**) *regulation of cortisol secretion* (*BiologicalProcess*) and *Vitamin D* (*calciferol*) *metabolism* (*Pathway*), (**d**) *regulation of vasoconstriction* (*BiologicalProcess*) and *regulation of blood vessel diameter* (*BiologicalProcess*), and (**e**) six *Anatomy* nodes. Violin plots for each significant node show that the ranks within Space vs. Baseline and/or Space vs. Ground separated from the Ground vs. Baseline. In each violin plot Ground vs. Baseline (blue), Space vs. Baseline (yellow), and Space vs. Ground (green).

Taken together, these results show that potential human physiological changes during spaceflight can be inferred by embedding mouse gene expression data with a KG that integrates observed concepts (i.e., genes) with unobserved, higher-order phenotypes associated with each other in a biologically meaningful manner.

4. Discussion

One of the major objectives of biomedical research is to advance our understanding of human diseases in order to develop effective countermeasures. This aim becomes considerably more challenging when the physiological changes arise from spaceflight. Major efforts have been made by NASA GeneLab to collect and provide multi-omics data from model organisms. Additionally, NASA GeneLab data brought into the SPOKE system could be complemented by including murine phenotypical pathophysiological and biochemical non-omics data (more nodes) from the Ames Life Sciences Data Archive [34], and eventually the SPOKE system could be used for human spaceflight research data related to astronauts. However, the major challenges of analyzing any datasets generated during spaceflight are their low statistical power, considerable heterogeneity, and limited reproducibility [35]. These limitations are largely accepted by the scientific community as a reasonable trade-off for the novelty and potential for discovery these experiments entail. As a new strategy to maximize the utility of these datasets, we propose the data from model organisms can be integrated through a knowledge graph (KG) such as SPOKE. KGs including SPOKE, are bounded by present day biological knowledge. As a result, inferences made through SPOKE may change as our biological data and knowledge expands.

Here, we report the results of a KG-driven, meta-analysis of six murine transcriptomic studies (five RNAseq and one microarray) from NASA GeneLab. The samples were taken from three distinct anatomical sites (thymus, liver, and spleen) and covered multiple spaceflight duration and gravity conditions. PCAs using only gene expression data illustrated that most of the differences between the samples could be attributed to either the study or the anatomical site.

Next, we hypothesized that, though this data came from a diverse set of experiments, SPOKE embeddings (i.e., "signatures") could be used to recover space travel changes that are conserved across the studies. To accomplish this, $-\log_2$ fold-change gene expression (FC) data from each study was applied to gene-specific Propagated SPOKE Entry Vectors (PSEVs). Gene-specific PSEVs are vectors that describe how important each node in SPOKE is for a given gene. Therefore, multiplying PSEVs by FC data will highlight nodes that are both important for the input gene set and to prioritize them according to how differentially expressed the input genes are.

PSEVs from all of the studies were then pooled together and separated into three groups based on the type of FC comparison (Ground vs. Baseline, Space vs. Baseline, and Space vs. Ground). The distribution of node rank was analyzed for each node and the top 5% were selected for each node type. These top nodes were enriched for nodes for phenotypes and physiological changes known to be impacted by spaceflight. Furthermore, paths were found between the input gene set and the top node set. These paths shed light onto the underpinnings of spaceflight related health hazards and could potentially be used to identify drug targets. In the future, archived spaceflight and other experimental samples could be used to validate the predicted signatures and assess their physiological significance without the need for further experiments. Thus, we anticipate that our results are the very first steps towards a broader collaboration utilizing the SPOKE model to compare spaceflight and terrestrial phenotypes.

There is increasing interest in developing personalized risk predictions and treatments in support of long-duration deep space missions [36]. Thus, expanding the computational approaches from the *general* comparison of spaceflight and terrestrial diseases to using input from a single subject to map their *individual* risk profile would allow developing optimal medical care for individual astronauts. Notably, the power of SPOKE stems from a wide variety of its inputs that combine multi-omics, clinical, and physiological data, which

may provide a useful complement to the currently utilized risk management tools that are based upon probabilistic mathematical modeling and simulations [37].

Using a knowledge graph connecting molecular and physiological entities (among others) via biologically relevant relationships constitutes a significant advancement for complex, heterogeneous data analysis. This approach complements conventional transcriptomics analysis by extending the biological significance to higher-level phenotypes such as symptoms and side effects, which is not possible with current methods. In the long-term perspective, the SPOKE platform may also be of value to mission planners such as the NASA Human Systems Risk Board.

Supplementary Materials: The following are available online at https://www.mdpi.com/2075-1729/11/1/42/s1, Figure S1. Gene selection for network paths. There is one scatter plot for each top node used in the networks. Each one shows the genes selected for path retracing (red) and those that had paths but were not shown (blue). The x-axis is the average *p*-value for the average FC distributions and the y-axis is the difference between the $\log_2$ combined *p*-values (from FC input) in the Ground vs. Baseline and space travel groups. Table S1. Welch's *t*-test results for Space vs. Baseline—Ground vs. Baseline and Space vs. Ground—Ground vs. Baseline tests. Results are shown for each node in SPOKE.

Author Contributions: Conceptualization (C.A.N., S.E.B., A.J.B., E.C., S.V.C.), methodology (C.A.N., S.E.B., S.V.C.), software (C.A.N.), validation (C.A.N.), formal analysis (C.A.N.), investigation (C.A.N., E.C., A.U.A.), resources (C.A.N.), data curation (C.A.N., A.U.A.), writing—original draft (C.A.N., S.E.B.), writing—review and editing (E.C., S.E.B., A.J.B., A.M.P., R.T.S.), visualization (C.A.N.), supervision (S.E.B., A.J.B., S.V.C.), project administration (S.E.B., S.V.C.), and funding acquisition (S.E.B., A.J.B., S.V.C.). All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by [National Science Foundation] grant number [NSF_2033569]. S.E.B. is the Distinguished Professor in Neurology at UCSF and holds the Heidrich Friends and Family endowed chair in Neurology at UCSF. GeneLab Project at NASA Ames Research Center, through NASA's Space Biology Program in the Division of Biological and Physical Sciences (BPS) of the Science Mission Directorate. Any use of trade names is for descriptive purposes only and does not imply endorsement by the US Government.

Institutional Review Board Statement: All mouse experiments were approved by the Institutional Animal Care and Use Committee of the University of Tsukuba, JAXA, Explore Biolabs, and NASA, and were conducted according to the applicable guidelines in Japan and the United States of America.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Afshinnekoo, E.; Scott, R.T.; MacKay, M.J.; Pariset, E.; Cekanaviciute, E.; Barker, R.; Gilroy, S.; Hassane, D.; Smith, S.M.; Zwart, S.R. Fundamental Biological Features of Spaceflight: Advancing the Field to Enable Deep-Space Exploration. *Cell* **2020**, *183*, 1162–1184. [CrossRef] [PubMed]
2. Berrios, D.C.; Galazka, J.; Grigorev, K.; Gebre, S.; Costes, S.V. NASA GeneLab: Interfaces for the exploration of space omics data. *Nucleic Acids Res.* **2021**, *49*, D1515–D1522. [CrossRef] [PubMed]
3. Rutter, L.; Barker, R.; Bezdan, D.; Cope, H.; Costes, S.V.; Degoricija, L.; Fisch, K.M.; Gabitto, M.I.; Gebre, S.; Giacomello, S. A New Era for Space Life Science: International Standards for Space Omics Processing. *Patterns* **2020**, *1*, 100148. [CrossRef] [PubMed]
4. da Silveira, W.; Fazelinia, H.; Rosenthal, S.B.; Laiakis, E.; Meydan, C.; Kidane, Y.; Smith, S.M.; Rathi, K.S.; Foox, J.; Zanello, S. Multi-Omics Analysis Reveals Mitochondrial Stress as a Central Hub for Spaceflight Biological Impact. *Cell* **2020**, *183*, 1185–1201. [CrossRef] [PubMed]
5. Malkani, S.; Chin, C.R.; Cekanaviciute, E.; Mortreux, M.; Okinula, H.; Tarbier, M.; Schreurs, A.S.; Shirazi-Fard, Y.; Tahimic, C.G.T.; Rodriguez, D.N.; et al. Circulating miRNA Spaceflight Signature Reveals Targets for Countermeasure Development. *Cell Rep.* **2020**, 108448. [CrossRef] [PubMed]
6. Barabasi, A.L.; Gulbahce, N.; Loscalzo, J. Network medicine: A network-based approach to human disease. *Nat. Rev. Genet.* **2011**, *12*, 56–68. [CrossRef]
7. Goh, K.I.; Cusick, M.E.; Valle, D.; Childs, B.; Vidal, M.; Barabasi, A.L. The human disease network. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 8685–8690. [CrossRef]

8. Nicholson, D.N.; Greene, C.S. Constructing knowledge graphs and their biomedical applications. *Comput. Struct. Biotechnol. J.* **2020**, *18*, 1414–1428. [CrossRef]
9. Wang, L.; Matsushita, T.; Madireddy, L.; Mousavi, P.; Baranzini, S.E. PINBPA: Cytoscape app for network analysis of GWAS data. *Bioinformatics* **2015**, *31*, 262–264. [CrossRef]
10. Zhou, X.; Menche, J.; Barabasi, A.L.; Sharma, A. Human symptoms-disease network. *Nat. Commun.* **2014**, *5*, 4212. [CrossRef]
11. Himmelstein, D.S.; Baranzini, S.E. Heterogeneous Network Edge Prediction: A Data Integration Approach to Prioritize Disease-Associated Genes. *PLoS Comput. Biol.* **2015**, *11*, e1004259. [CrossRef] [PubMed]
12. Himmelstein, D.S.; Lizee, A.; Hessler, C.; Brueggeman, L.; Chen, S.L.; Hadley, D.; Green, A.; Khankhanian, P.; Baranzini, S.E. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *Elife* **2017**, *6*, e26726. [CrossRef] [PubMed]
13. Krueger, F. Trim galore. In *A Wrapper Tool around Cutadapt and FastQC to Consistently Apply Quality and Adapter Trimming to FastQ Files*; Babraham Bioinformatics: Cambridge, UK, 2015; Volume 516, p. 517.
14. Andrews, S. *FastQC: A Quality Control Tool for High Throughput Sequence Data*; Babraham Bioinformatics, Babraham Institute: Cambridge, UK, 2010.
15. Ewels, P.; Magnusson, M.; Lundin, S.; Käller, M. MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **2016**, *32*, 3047–3048. [CrossRef] [PubMed]
16. Dobin, A.; Davis, C.A.; Schlesinger, F.; Drenkow, J.; Zaleski, C.; Jha, S.; Batut, P.; Chaisson, M.; Gingeras, T.R. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **2013**, *29*, 15–21. [CrossRef]
17. Li, B.; Dewey, C.N. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinform.* **2011**, *12*, 323. [CrossRef]
18. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2013.
19. 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. Assessing technical performance in differential gene expression experiments with external spike-in RNA control ratio mixtures. *Nat. Commun.* **2014**, *5*, 5125. [CrossRef]
20. Love, M.I.; Huber, W.; Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **2014**, *15*, 550. [CrossRef]
21. Sievert, C. *Interactive Web-Based Data Visualization with R, Plotly, and Shiny*; CRC Press: Boca Raton, FL, USA, 2020.
22. Lebsack, T. *Microarray Analysis of Space-Flown Murine Thymus Tissue*, 4th ed.; NASA GeneLab, Ed.; NASA GeneLab: Washington, DC, USA, 2010. [CrossRef]
23. Ashburner, M.; Ball, C.A.; Blake, J.A.; Botstein, D.; Butler, H.; Cherry, J.M.; Davis, A.P.; Dolinski, K.; Dwight, S.S.; Eppig, J.T.; et al. Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **2000**, *25*, 25–29. [CrossRef]
24. Mi, H.; Muruganujan, A.; Ebert, D.; Huang, X.; Thomas, P.D. PANTHER version 14: More genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res.* **2019**, *47*, D419–D426. [CrossRef]
25. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.* **2019**, *47*, D330–D338. [CrossRef]
26. Page, L.; Brin, S.; Motwani, R.; Winograd, T. *The PageRank Citation Ranking: Bringing Order to the Web*; Stanford InfoLab: Stanford, CA, USA, 1999.
27. Haveliwala, T.H. Topic-sensitive pagerank. In Proceedings of the 11th International Conference on World Wide Web, Honolulu, HI, USA, 7–11 May 2002; pp. 517–526.
28. Nelson, C.A.; Butte, A.J.; Baranzini, S.E. Integrating biomedical research and electronic health records to create knowledge-based biologically meaningful machine-readable embeddings. *Nat. Commun.* **2019**, *10*, 3045. [CrossRef] [PubMed]
29. Sayers, E.W.; Agarwala, R.; Bolton, E.E.; Brister, J.R.; Canese, K.; Clark, K.; Connor, R.; Fiorini, N.; Funk, K.; Hefferon, T.; et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **2019**, *47*, D23–D28. [CrossRef] [PubMed]
30. Stouffer, S.A.; Suchman, E.A.; DeVinney, L.C.; Star, S.A.; Williams, R.M., Jr. *The American Soldier: Adjustment during Army Life. (Studies in Social Psychology in World War II)*; Princeton University Press: Princeton, NJ, USA, 1949; Volume 1.
31. Lukacs-Kornek, V. The Role of Lymphatic Endothelial Cells in Liver Injury and Tumor Development. *Front. Immunol.* **2016**, *7*, 548. [CrossRef] [PubMed]
32. Mann, H.B.; Whitney, D.R. On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.* **1947**, *18*, 50–60. [CrossRef]
33. Lebsack, T.W.; Fa, V.; Woods, C.C.; Gruener, R.; Manziello, A.M.; Pecaut, M.J.; Gridley, D.S.; Stodieck, L.S.; Ferguson, V.L.; Deluca, D. Microarray analysis of spaceflown murine thymus tissue reveals changes in gene expression regulating stress and glucocorticoid receptors. *J. Cell. Biochem.* **2010**, *110*, 372–381. [CrossRef] [PubMed]
34. Scott, R.T.; Grigorev, K.; Mackintosh, G.; Gebre, S.G.; Mason, C.E.; Del Alto, M.E.; Costes, S.V. Advancing the Integration of Biosciences Data Sharing to Further Enable Space Exploration. *Cell Rep.* **2020**, *33*, 108441. [CrossRef]
35. Reynolds, R.J.; Shelhamer, M. Introductory Chapter: Research Methods for the Next 60 Years of Space Exploration. In *Beyond LEO-Human Health Issues for Deep Space Exploration*; IntechOpen: London, UK, 2020.

36. Antonsen, E.L.; Reed, R.D. Policy Considerations for Precision Medicine in Human Spaceflight. *Houst. J. Health Law Policy* **2019**, *19*, 1–35.
37. Antonsen, E.; Reynolds, R. *Risk Mapping and Interaction Approach: A Special Session for HSRB Risk Custodians*; NASA: Washington, DC, USA, 2020.