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 Butte[5,6], Egle Cekanaviciute[2], Sergio E. Baranzini[1,5,7*], Sylvain V. Costes[2*]

[1] Integrated Program in Quantitative Biology. University of California San Francisco.
[2] Space Biosciences Division, NASA Ames Research Center, Moffett Field, CA 94035, USA
[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.
[6] Department of Pediatrics. University of California San Francisco.
[7] Weill Institute for Neuroscience. Department of Neurology. University of California San Francisco.

*Corresponding Authors:
Sergio E Baranzini: Sergio.Baranzini@ucsf.edu
Sylvain Costes: sylvain.v.costes@nasa.gov

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, 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 1; knowledge graph 2; transcriptomics 3

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 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. 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.

2. Materials and Methods

GeneLab data processing and analysis.

Gene expression data was 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 verses ground control data was used for analysis.

Raw data was 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) and filtered using Trim Galore! (version 0.6.4). Raw and trimmed fastq file quality was evaluated with FastQC (version 0.11.9). MultiQC (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 and RSEM 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 and imported to R (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, i.e. a spike-in mixture used for normalization) genes were used. Data was normalized with DESeq2 (version 1.26.0). Principal component analysis was performed using prcomp (stats version 3.6.3) and plotted using plotly (version 4.9.2.1). For datasets containing DNA microarray assays (GLDS-4) raw .CEL files were read in and normalized using the R script 'affyNormQC.R' which utilizes the RMA algorithm through the oligo R package [rma() with default parameters]. Quality control reports were generated via the R script 'affyNormQC.R', with parameter 'do.logtransform' set to TRUE for the generating the raw report. This microarray experiment was annotated with the R script 'annotateProbes.R' which utilized Annotation-Db class probe annotations specific to each chip from the Bioconductor repository. In cases where multiple probes mapped to the same gene ID, representative probes were selected with the highest mean normalized intensity across all samples. The results of the principal component analysis were imported to R using the GeneFab API and plotted using plotly (version 4.9.2.1).

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 (GO) gene sets. (GO biological process, GO cellular component, G molecular function). Top 50 statistically significant pathways were compared to identify overlaps. The same approach was applied to quantify the overlapping pathways between GLDS-288 and -289.

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 publically 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).

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[13 14 15]. These restart nodes, called SPOKE Entry Points (SEPs), are any concept in the input data that overlaps with a node(s) in SPOKE. 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).

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[16]. If multiple mouse genes mapped to a single human, 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 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 weren't in the same direction (i.e. if space verses ground day-29 had a positive FC and days-53-56 had a negative FC). This filter focuses 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 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 FC comparisons were converted to PSEVs by taking the dot product of the filtered FC matrix and the filtered normalized PSEV matrix. Finally, the PSEV comparisons were ranked as before.

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). Wilcoxon rank-sum 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 (Supplementary Table 1). Top nodes were selected using the most significant 2.5% per node type for Space vs. Ground and/or Space vs. Baseline (n=15,875; 4.1%).

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 = 7,567) and a set of top *Anatomy, BiologicalProcess, CellularComponent, MolecularFunction, Pathway*, and *Symptom* nodes (n = 30). 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 (Supplementary Figure 1). The p-values of the FCs used as input for PSEV creation were averaged for each group (Ground vs. Baseline, Space vs. Baseline, and Space vs. Ground). This gives us an estimate of how significant the gene FC was for a group as a whole. Each gene FC was scored based on whether the average space travel groups had a p-value that on average was more significant than Ground v Baseline (equation 1; Supplementary Figure 1 y-axis). Here, a positive value indicates that the average p-value of the FC for a given gene was more significant within space travel groups than the Ground v Baseline group. This score only judges the significance of one comparison (within a single group) to the other. Then the Wilcoxon rank-sum 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 the then averaged (Supplementary Figure 1, x-axis).

$$FC\ score = \log_2(\text{Ground v Baseline Avg P Value}) - \log_2\big((\text{Space v Baseline Avg P Value} + \text{Space v Ground Avg P Value})/2\big) \quad (1)$$

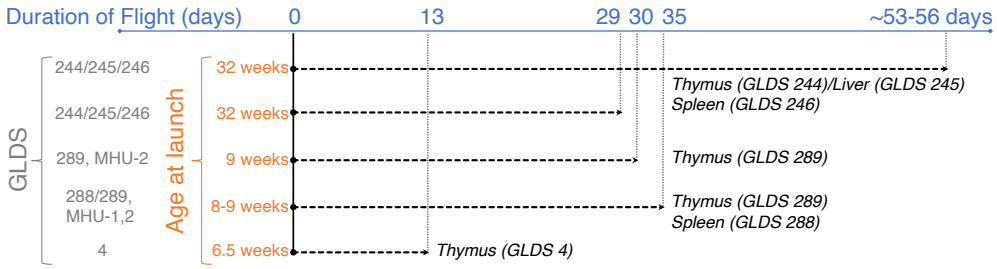


Figure 1. Summary of experimental conditions across GeneLab datasets used for the analysis. Datasets GLDS-4, -244, -245 and -246 used C57BL/6NTac mice. Datasets GLDS-288 and -289 used C57BL/6J mice for spaceflight and both C57BL/6J and Charles River mice for ground controls.

Table 1. Descriptive metadata for each NASA GLDS dataset analyzed by SPOKE.

GeneLab	Tissue	Sequencing Type	Strain	Mission/Flight	Duration in flight	Age of initiation	Age of Euthanasia	Sex	Sample size (n/cohort)
GLDS-4	Thymus	Microarray	C57BL/6NTac	STS-118	13-days (12.7h day)	~1.5-weeks	8-weeks	n/a	FLT(n=4); GC(n=4)
GLDS-244	Thymus	RNA-sequencing	C57BL/6NTac	RR-6 (SpaceX-13)	29-days (n=4; LAR); 35-36-days (n=11; ISS terminal)	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)
GLDS-245	Liver	RNA-sequencing	C57BL/6NTac	RR-6 (SpaceX-13)	29-days (n=4; LAR); 35-36-days (n=11; ISS terminal)	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)
GLDS-246	Spleen	RNA-sequencing	C57BL/6NTac	RR-6 (SpaceX-13)	29-days (n=4; LAR); 35-36-days (n=11; ISS terminal)	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)
GLDS-248	Spleen	RNA-sequencing	C57BL/6J (Eighty) Charles River Laboratories Japan (GC)	TCU (SpaceX-9)	35-days	8-weeks	12-weeks	Male	SpaceflightNAC, n=3; Spaceflight w/ centrifugation (AG, n=3; Synchronous (GC, n=3
GLDS-249	Thymus	RNA-sequencing	C57BL/6J (Eighty) Charles River Laboratories Japan (GC)	TCU (SpaceX-9 & MHU-1; SpaceX-12, MHU-2)	35-days MHU-1; 34-days MHU-2	8-weeks MHU-1; 8-weeks MHU-2	12-weeks	Male	SpaceflightNAC, MHU-1, n=3; Spaceflight w/ centrifugation (AC, MHU-1, n=3; Synchronous (GC, MHU-1, n=3; SpaceflightNAC, MHU-2, n=3; Spaceflight w/ centrifugation (AC, MHU-2, n=3; Synchronous (GC, MHU-2, n=3)

198

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

200

201

202

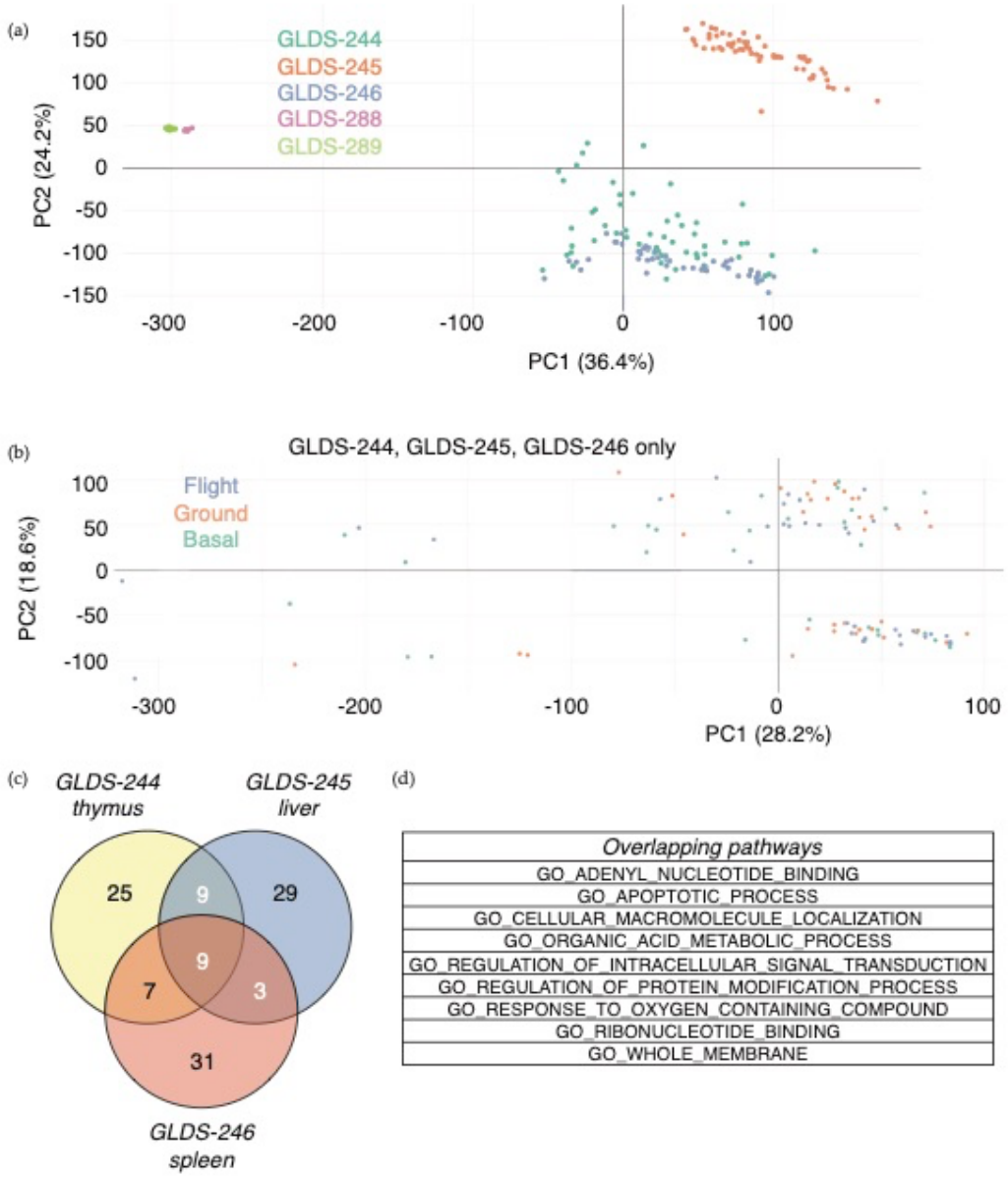


Figure 2. Transcriptomic analysis of spaceflight-associated changes in gene expression. (a). Principal component analysis of all samples, colored by dataset. **(b).** Principal component analysis of datasets GLDS-244, -245 and -246, colored by flight condition. **(c,d).** Overlapping pathways between datasets GLDS-244, 245 and 246 out of top 50 Gene Ontology pathways 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 pathways between datasets (**c**) and the list of pathways overlapping between all three datasets (**d**). Three out of top 50 gene ontology (GO) pathways overlapped between datasets GLDS-288 and -289, none of which overlapped with GLDS-244, -245 and -246.

3. Results

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 1 and Table 1). While experiments varied in their design (i.e. duration of flight, age at launch, 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 (lymphatic-associated/digestive organ, PMID:27965673)).

After data normalization, principal component analysis revealed 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 (spaceflown, ground, and baseline from the same RR-6 mission), no obvious separation between samples obtained during flight, baseline and ground was observed (Figure 2B).

Differentially expressed genes in spaceflown mice vs. ground controls after live animal return were identified in thymus, liver, and spleen of RR-6 (SpaceX-13) mission, including a set of overlapping genes across all three tissues (Figure 2C). Using these genes as an input to pathway analysis (by hypergeometric test) further showed a number of statistically significant biological functions dysregulated by space flight in thymus, liver and spleen. While some pathways were tissue-specific, nine of them were shared among the three tissues, including apoptosis, cell metabolic process, and cell membrane integrity (Figure 2D).

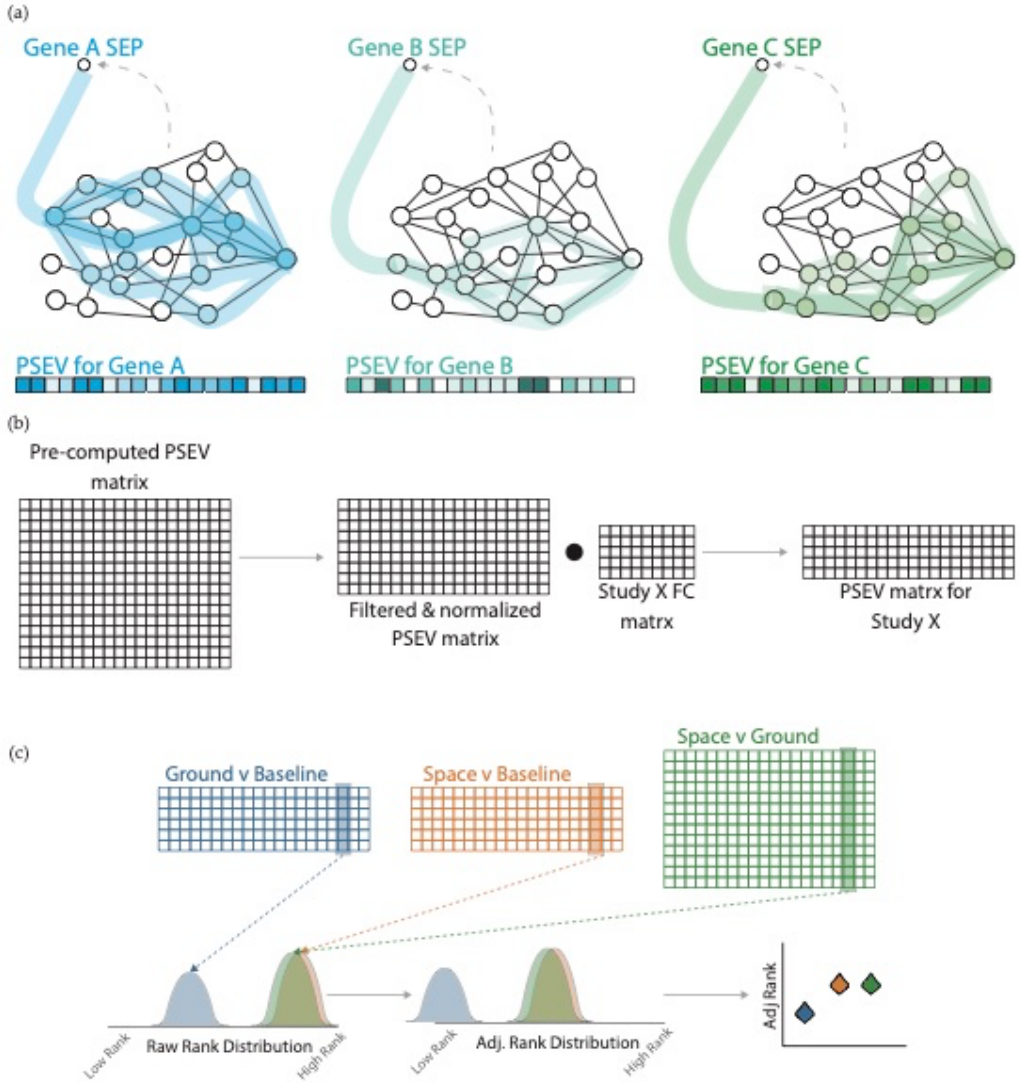


Figure 3. Generating PSEVs using gene expression fold-change. (a) PSEVs were pre-computed for all SPOKE genes. For each gene the random walker was forced to restart at that gene (probability of random jump = 0.1). After PSEVs were finished they were stored in the pre-computed PSEV matrix. **(b)** For 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 generated 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.

Fold-change enhanced Propagated SPOKE Entry Vectors

While established methods of transcriptional profiling can inform about dysregulated molecular pathways, they provide little insight on higher-order phenotypes, such as associated

255 signs and symptoms of disease. Using SPOKE, a KG that integrates information of both biological
256 and clinical database, it is possible to score every node of the graph as a function of the
257 "information flow" elicited by a defined set of quantitative inputs. SPOKE leverages the complexity
258 of hierarchical organiztion of complex organisms to identify nodes with shared information flow
259 (regardless of whether the input itself was significant or not).
260 Gene-specific Propagated SPOKE Entry Vectors (PSEVs) were generated from the selected
261 GeneLab studies prior to integrating gene expression results with SPOKE[11 12]. Each gene-specific
262 PSEV was created using a modified version of topic specific page rank[13 14 15] in which the random
263 walker was forced to restart at the corresponding *Gene* node in SPOKE (See Methods, Figure 3A).
264 This focused the random walker on nodes that are the most important for a given node (in this
265 case, *Gene* node since the input is gene expression). The amount of time a random walker spends
266 on a node was then stored in a defined element (position within) of the PSEV vector. All PSEVs
267 were then stored in the pre-computed PSEV matrix. For each gene expression study the pre-
268 computed PSEV matrix was filtered and normalized to match the genes within the study (Figure
269 3B; Methods). The dot product was then used with the normalized PSEV matrix and the $-\log_2$ fold-
270 change (FC) to produce the PSEVs for that study. After PSEVs were computed for each study, they
271 were pooled and separated into specific experimental groups to enable meaningful comparisons
272 to test the hypothesis that spaceflight alters gene expression (Ground vs. Baseline, Space vs.
273 Baseline, and Space vs. Ground) (Figure 3C).
274 Each element in a PSEV corresponds to a single node in SPOKE. Therefore, it is possible to
275 determine the overall significance of a node for spaceflight by evaluating the differential
276 distribution of node ranks in the PSEV. Wilcoxon rank-sum test [17] was utilized to compare a node's
277 rank distribution in the Ground vs. Baseline to that in either Space vs. Baseline or Space vs. Ground
278 (Supplementary Table 1).
279 Strikingly, nodes that are known to be relevant for space travel such as space motion
280 sickness (*Symptom*), regulation of blood vessel diameter (*BiologicalProcess*), taste receptor complex
281 (*CellularComponent*), Vitamin D (calciferol) metabolism (*Pathway*), and sympathetic nervous system
282 (*Anatomy*) scored among the top 5% of nodes (top 2.5% per type for Space vs. Baseline and/or Space
283 vs. Ground). Figure 4 shows violin plots from a select set of nodes (n = 22) in SPOKE that had
284 significantly different ranks in spaceflight (Space vs. Baseline and/or Space vs. Ground) compared
285 to Ground vs. Baseline. From these, 11 correspond to symptoms (pink boxed violin charts, Figure
286 4A), five to gene ontology/pathway concepts (teal boxed violin charts, Figure 4B-D), and six to
287 anatomical regions (green boxed violin charts, Figure 4E). Violin plots for each category, sub-
288 networks demonstrate how the gene expression results drive information from these 22 nodes.
289 Taken together, these results show that human physiological changes observed during
290 spaceflight can be inferred by embedding mouse gene expression data with a KG that integrates
291 observed concepts (i.e. genes) with unobserved, higher order phenotypes associated with each
292 other in a biologically meaningful manner.
293

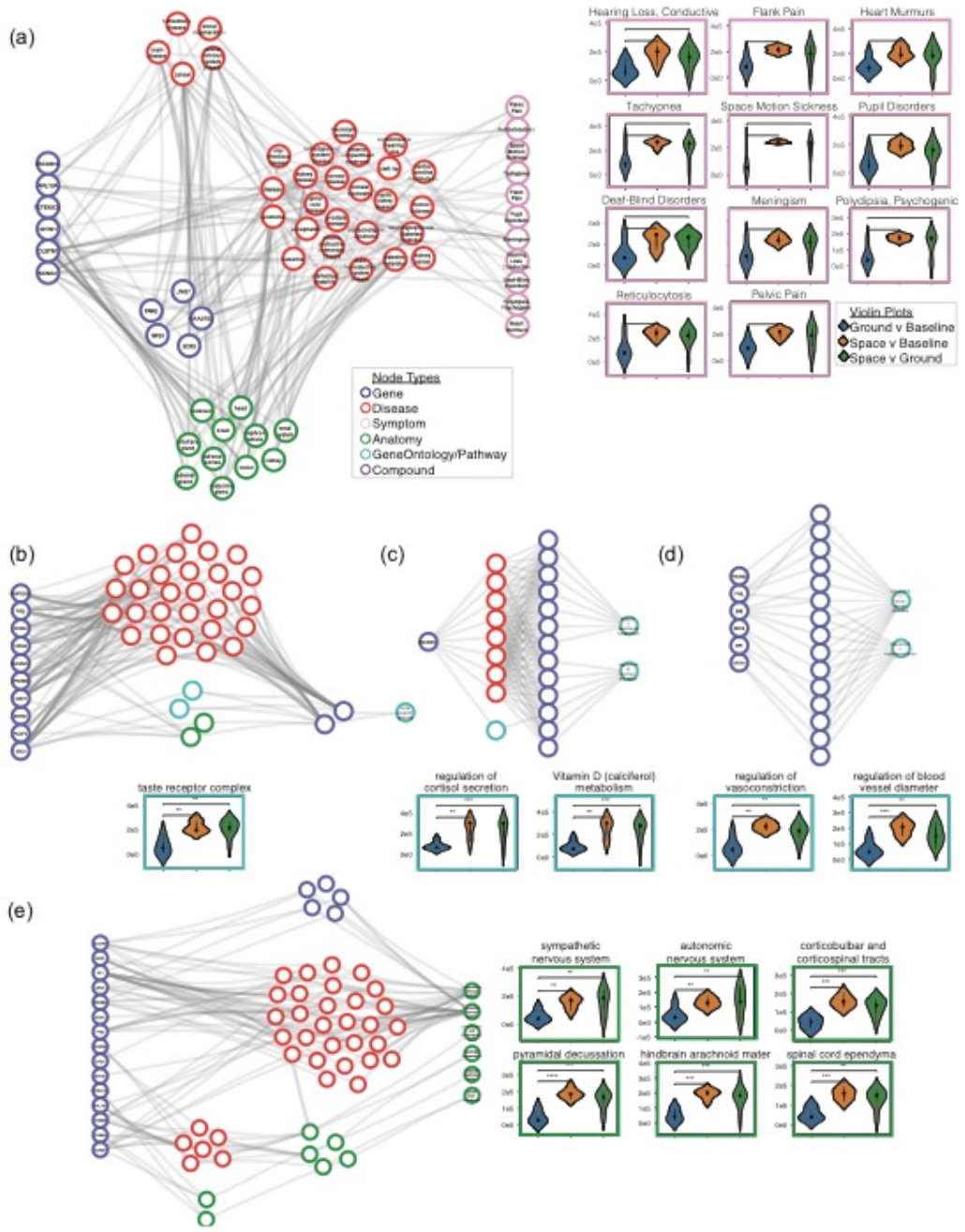


Figure 4. Retracing paths between genes and top nodes. Gene expression FC values drive information flow to nodes in SPOKE. **(a-e)** Paths were traced between genes that were partially responsible for pushing information to a set of significant nodes (n = 22). These paths were shown for **(a)** 10 *Symptom* 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).

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 multiomics data from model organisms. Additionally, NASA GeneLab data brought into the SPOKE system could be complemented by including murine phenotypical patho-physiological and biochemical non-omics data (more nodes) from the Ames Life Sciences Data Archive,[18] 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[19]. 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 such as SPOKE.

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 durations 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.

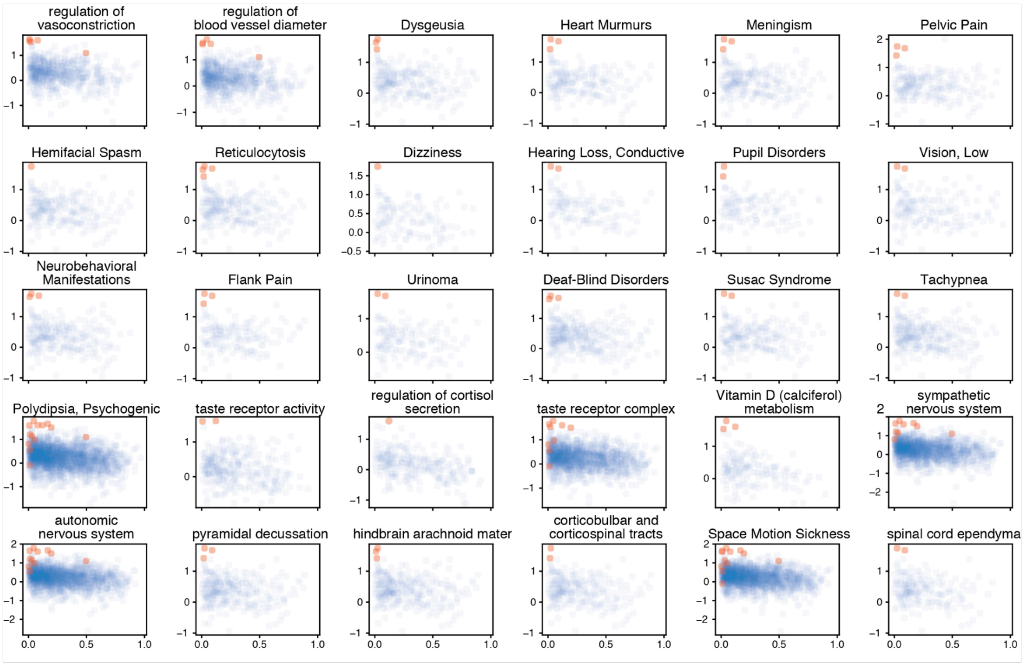
Next, we hypothesized that, though this data came from a diverse set of experiments, SPOKE embeddings (i.e. "signatures") could be used to recover space travel changes that are conserved across the studies. To accomplish this, $-\log_2$ fold-change gene expression (FC) data from each study was applied 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 input gene set and 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[20]. Thus, expanding the computational approaches from *general* comparison of spaceflight and terrestrial diseases to using an 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 [21]. 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 www.mdpi.com/xxx/s1, Figure S1: Gene selection for network paths, Table S1: Wilcoxon rank-sum test results



Supplementary Figure 1. 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 y-axis is the FC score.

Supplementary Table 1. Wilcoxon rank-sum test results for Space vs. Baseline - Ground vs. Baseline and Space vs. Ground - Ground vs. Baseline tests. Results shown for each node in SPOKE.

References

1. Afshinnekoo, E. *et al.* Fundamental Biological Features of Spaceflight: Advancing the Field to Enable Deep-Space Exploration. *Cell* **183**, 1162-1184 (2020).
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* (2020).
3. Rutter, L. *et al.* A New Era for Space Life Science: International Standards for Space Omics Processing. *Patterns*, 100148 (2020).
4. da Silveira, W. *et al.* Multi-Omics Analysis Reveals Mitochondrial Stress as a Central Hub for Spaceflight Biological Impact.
5. Malkani, S. *et al.* Circulating miRNA Spaceflight Signature Reveals Targets for Countermeasure Development. *Cell Rep*, 108448 (2020).
6. Barabasi, A.L., Gulbahce, N. & Loscalzo, J. Network medicine: a network-based approach to human disease. *Nat Rev Genet* **12**, 56-68 (2011).
7. Goh, K.I. *et al.* The human disease network. *Proc Natl Acad Sci U S A* **104**, 8685-90 (2007).
8. Nicholson, D.N. & Greene, C.S. Constructing knowledge graphs and their biomedical applications. *Comput Struct Biotechnol J* **18**, 1414-1428 (2020).
9. Wang, L., Matsushita, T., Madireddy, L., Mousavi, P. & Baranzini, S.E. PINBPA: cytoscape app for network analysis of GWAS data. *Bioinformatics* **31**, 262-4 (2015).
10. Zhou, X., Menche, J., Barabasi, A.L. & Sharma, A. Human symptoms-disease network. *Nat Commun* **5**, 4212 (2014).
11. Himmelstein, D.S. & Baranzini, S.E. Heterogeneous Network Edge Prediction: A Data Integration Approach to Prioritize Disease-Associated Genes. *PLoS Comput Biol* **11**, e1004259 (2015).
12. Himmelstein, D.S. *et al.* Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *Elife* **6**(2017).
13. Page, L., Brin, S., Motwani, R. & Winograd, T. The PageRank citation ranking: Bringing order to the web. (Stanford InfoLab, 1999).
14. Haveliwala, T.H. Topic-sensitive pagerank. in *Proceedings of the 11th international conference on World Wide Web* 517-526 (ACM, 2002).
15. 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* **10**, 3045 (2019).
16. Sayers, E.W. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **47**, D23-D28 (2019).
17. Mann, H.B. & Whitney, D.R. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, 50-60 (1947).
18. Scott, R.T. *et al.* Advancing the Integration of Biosciences Data Sharing to Further Enable Space Exploration. *Cell Reports*, 108441 (2020).
19. 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, 2020).
20. Antonsen, E.L. & Reed, R.D. Policy Considerations for Precision Medicine in Human Spaceflight. *Hous. J. Health L. & Pol'y* **19**, 1 (2019).
21. Antonsen, E. & Reynolds, R. Risk Mapping and Interaction Approach: A Special Session for HSRB Risk Custodians. (2020).

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

© 2020 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

415