Systemic Genome Correlation Loss as a Central Characteristic of Spaceflight

Anurag Sakharkar[1]

Dr. Erique Lukong[2]

Dr. Lauren M. Sanders[3]

Dr. Sylvain V. Costes[4]

Dr. Changiz Taghibiglou[5]*

Dr. Jian Yang[6]*

[1] College of Arts and Science; University of Saskatchewan; 9 Campus Drive, Saskatoon, Saskatchewan S7N5A5, Canada.

[2] Department of Biochemistry, Microbiology, and Immunology; 107 Wiggins Road, Health Sciences Building, University of Saskatchewan, Saskatoon, Saskatchewan S7N5E5, Canada.

[3] NASA Ames Research Center, Space Biosciences Division, Blue Marble Space Institute of Science, Moffett Field, California, USA.

[4] NASA Ames Research Center, Space Biosciences Division, Moffett Field, California, USA.

[5] Department of Anatomy, Physiology, and Pharmacology; 107 Wiggins Road, Health Sciences Building, University of Saskatchewan, Saskatoon, Saskatchewan S7N5E5, Canada.

[6] College of Pharmacy and Nutrition; 107 Wiggins Road, Health Sciences Building, University of Saskatchewan, Saskatoon, Saskatchewan S7N5E5, Canada.

Correspondence: jian.yang@usask.ca and changiz.taghibiglou@usask.ca

Keywords: Astronaut health, spaceflight, space omics, microgravity, radiation, International Space Station, transcriptome, space exploration, NASA, NASA OSDR

Abstract

Space exploration has captured the imagination of humanity for generations. From the first steps on the moon to the recent Mars rover and Artemis lunar exploration missions, space travel has always been an ambitious goal for humanity. However, as we venture further into space and prepare for long-term missions to other planets, the physiological and health risks associated with prolonged space travel are becoming more prominent. Most current research on astronaut health focuses on identifying individual genes or pathways for specific symptoms astronauts face. The human system is complex and delicate, and the effects of microgravity, radiation, and isolation on astronaut health during long-duration spaceflight are still not fully understood. This study used a novel ranking and analysis methodology to combine space omics data from multiple datasets in the NASA OSDR repository. The data was used to generate a multi-omic, integrative bioinformatics analysis pipeline, which identified and characterized a genome-wide spaceflight gene expression correlation loss as a central biosignature for astronaut health on the International Space Station (ISS). Our findings indicate that genome-wide correlation loss corresponds to a breakdown in gene synchronization and cooperation, showcasing the systemic symptoms spaceflight induces and their genomic roots.

Introduction

Spaceflight biosciences and medicine have been a crucial part of space exploration since its inception. The significant effects of spaceflight on the human body were recognized very early on, and efforts were made to ensure astronauts remained healthy during their missions. Human spaceflight is associated with numerous physiological and psychological changes, including bone loss, muscle atrophy, and cognitive decline, which can compromise the health and safety of astronauts during long-duration missions. Recent research has integrated multi-omic analysis methods in individual datasets from space omics databases, with great results (1, 2). The most notable successes in space health include the development of the space suit, which allowed astronauts to perform extravehicular activities, and the implementation of exercise programs to combat spaceflight-induced muscle and bone loss (3). However, the global effects of spaceflight on the human genome remain a relatively unexplored area of research. Long-term health monitoring after return from space missions is currently done through the Lifetime Surveillance of Astronaut Health (LSAH) program, and although specific pathways, such as mitochondrial dysfunction and inflammatory stress, have been identified as characteristic of spaceflight (4), very few genome-wide responses to spaceflight have been investigated so far. Concurrently, it is known that the vast physiological symptoms associated with long-term spaceflight have genomic roots (2). Identifying these underlying mechanisms and spaceflight biosignatures is critical to developing effective preventive and therapeutic interventions for astronaut health in long-term spaceflight.

In recent years, multi-omics analysis has emerged as a powerful tool for understanding complex biological systems. Multi-omics combines different types of molecular data, including genomics, transcriptomics, proteomics, and metabolomics, to obtain a more comprehensive understanding of biological processes. Pertaining to astronaut health, multi-omics analysis has the potential to identify biomarkers for different physiological systems that are impacted by space travel. However, the existing literature primarily focuses on individual omics, and there is a lack of integrative and system-wide analyses (5, 6). Moreover, most studies have small sample sizes, which limits the generalizability of the findings. The National Aeronautics and Space Administration (NASA)'s OSDR data repository (7, 8), which is an aggregation of many publicly accessible omics datasets, provides a good data resource for such analyses.

In this study, we characterize a large decrease in gene-pair correlation as a measure of whole genome regulation, as well as in individual signaling pathways, using data from multiple missions on the International Space Station (ISS). Differential expression, in combination with differential correlation, pathway analysis, network analysis, and functional analysis, was conducted for this multi-omics study. This analysis also generated a list of potential biomarkers that may help characterize, monitor, diagnose, prevent, and even treat astronaut health while in space.

Gene-pair correlation loss is relatively unexplored as a biosignature for astronaut health. Understanding this expression alteration as a biosignature could allow for a better understanding of the molecular mechanisms that underlie and control health during spaceflight. Furthermore, identifying treatment opportunities for maintaining and treating astronaut health would improve

the safety and well-being of astronauts during the long-term space missions that are undoubtedly a large part of humanity's future.

Results

Differential expression analysis

Previous studies have shown numerous differentially expressed genes in almost every tissue of astronauts on the ISS (1, 2). This mirrors the genome-wide correlation loss seen in astronauts and reinforces the understanding of spaceflight physiological issues being a multi-omics problem to solve. The DEG analysis conducted in this study comparing preflight to inflight samples, with preflight samples treated as controls, identified 6903 differentially expressed genes with a p-value cutoff of $p \leq 0.05$, which corresponds to approximately a quarter of the human genome (Supplementary Data 1). This serves as another verification of the profound and complex effect of spaceflight on genome-wide expression. When subjected to a functional analysis, it was clear that these DEGs regulate a large variance of biological, physiological, and pathophysiological functions, including many neuronal issues.

For example, the most upregulated gene in the inflight samples compared to the preflight samples, *SNORA81*, encodes for a small nucleolar RNA (snoRA81, H/ACA box 81) involved in RNA processing. This gene is particularly responsible for the conversion of uridine to pseudouridine at specific ribosomal or small nuclear RNA residues (9). *SNORA81* is known to be enriched in neurodegenerative diseases such as Parkinson's disease (10) and in stroke incidents (11). Additionally, *SNORA81* was recently identified as an important factor in RNA deregulation

in Amyotrophic Lateral Sclerosis (12), a rare neuromuscular disease characterized by the progressive degeneration of motor neurons in the spinal cord, brainstem, and motor cortex (13). Thus, the overexpression of *SNORA81* may explain observations of long-term mild stroke symptoms and motor neuron degeneration in astronauts (2). The second-most upregulated gene, *PKP1*, is predominantly associated with human melanoma metastasis. It has been identified as a hub gene and a key module that may serve as a potential tumor enhancer (13). Therefore, significant upregulation of *PKP1* in astronauts is presumed to be one of the major factors responsible for the significantly higher rates of cancer-related illnesses in astronauts over their lifetime. Similarly, the other genes identified through the DEG analysis all play important roles in symptoms astronauts experience in spaceflight. The large number of genes identified from this sole analysis already presents a significant step forward in the validation of current data and the identification of biomarkers and targets for health monitoring and treatment

Pathway analysis

The DEG analysis is a logical lead into a pathway analysis step to identify the signaling pathways and cell systems these genes are involved in. However, an important weakness in classical pathway analysis is that it only characterizes pathways from the expression of individual genes, and is not representative of the multi-omic analysis proposed in this study. Therefore, differential expression level absolute fold-changes and absolute-value correlation coefficients were multiplied to form a score termed the Correlation Index (CI) (Supplementary Data 2). Multiplication emphasizes high scores in either analysis method while retaining information from the other. These CIs were then aligned with individual genes and used in a

pathway overrepresentation analysis (ORA) using the *InnateDB* online tool (14) (Supplementary Data 3). The Gene Ontology results from this pathway analysis showcase that most of the heavily affected pathways are large-scale systems in the cell, like gene expression, the cell cycle, apoptosis, and the citric acid cycle. Not only are these processes known to be essential in general cell function and regulation, but the results from this analysis also highlight the large, multi-subsystemic impact of spaceflight on astronauts' bodies, and inform the need for multi-omics analyses of space biology data and the identification of spaceflight biosignature patterns.

Correlation analysis

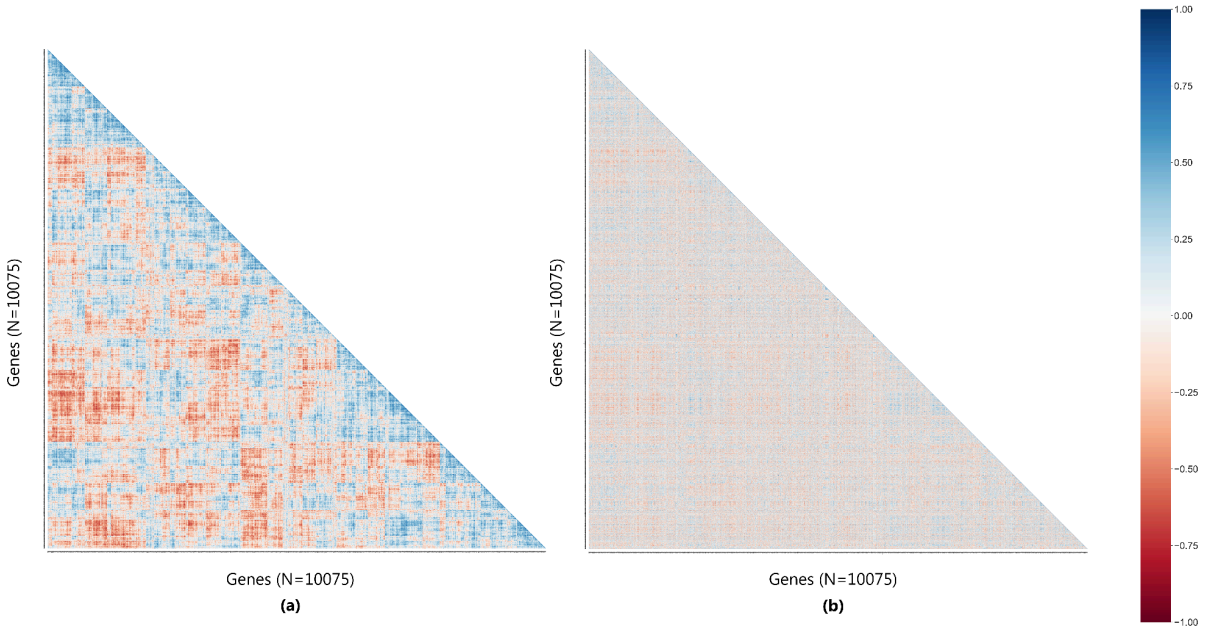


Figure 1: Spaceflight causes a significant loss in gene-pair correlation across the entire genome.

(a) Gene-pair correlations in the master preflight dataset show strong correlations across most of the genome, indicating high levels of genome coordination and synergy across pathways. (b) Gene-pair correlations in the master inflight dataset show a significant loss in gene-pair correlation across the entire genome, indicating a loss in genome regulation, both across all genes and in specific signaling pathways.

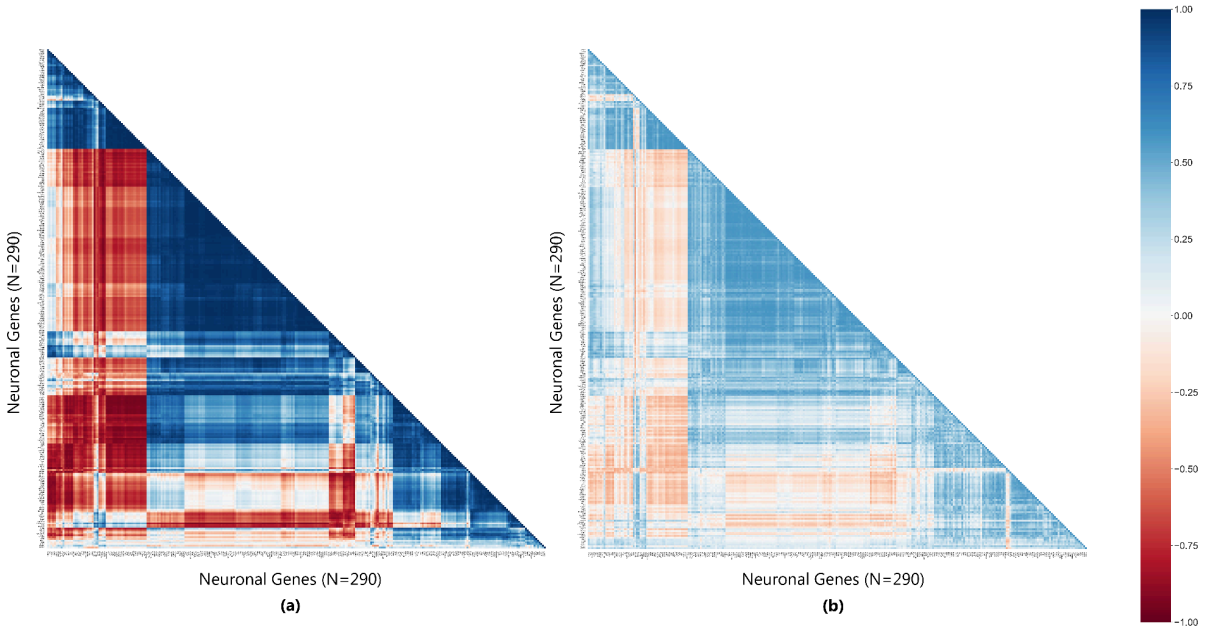


Figure 2: Spaceflight causes a significant loss in gene-pair correlation in neuronal genes. *Strong preflight correlations (a) imply high levels of gene regulation and coordination amongst neuronal genes, an example of pathway-specific gene-pair correlation being affected similarly to genome-level correlation. Correlation breakdown in (b) indicates significant neuronal dysregulation in spaceflight genomics and confirms gene-pair correlation as an important biosignature for spaceflight. This correlation loss is markedly more significant than the loss across the genome shown in Figure 2, indicating that neuronal genes might be disproportionately affected.*

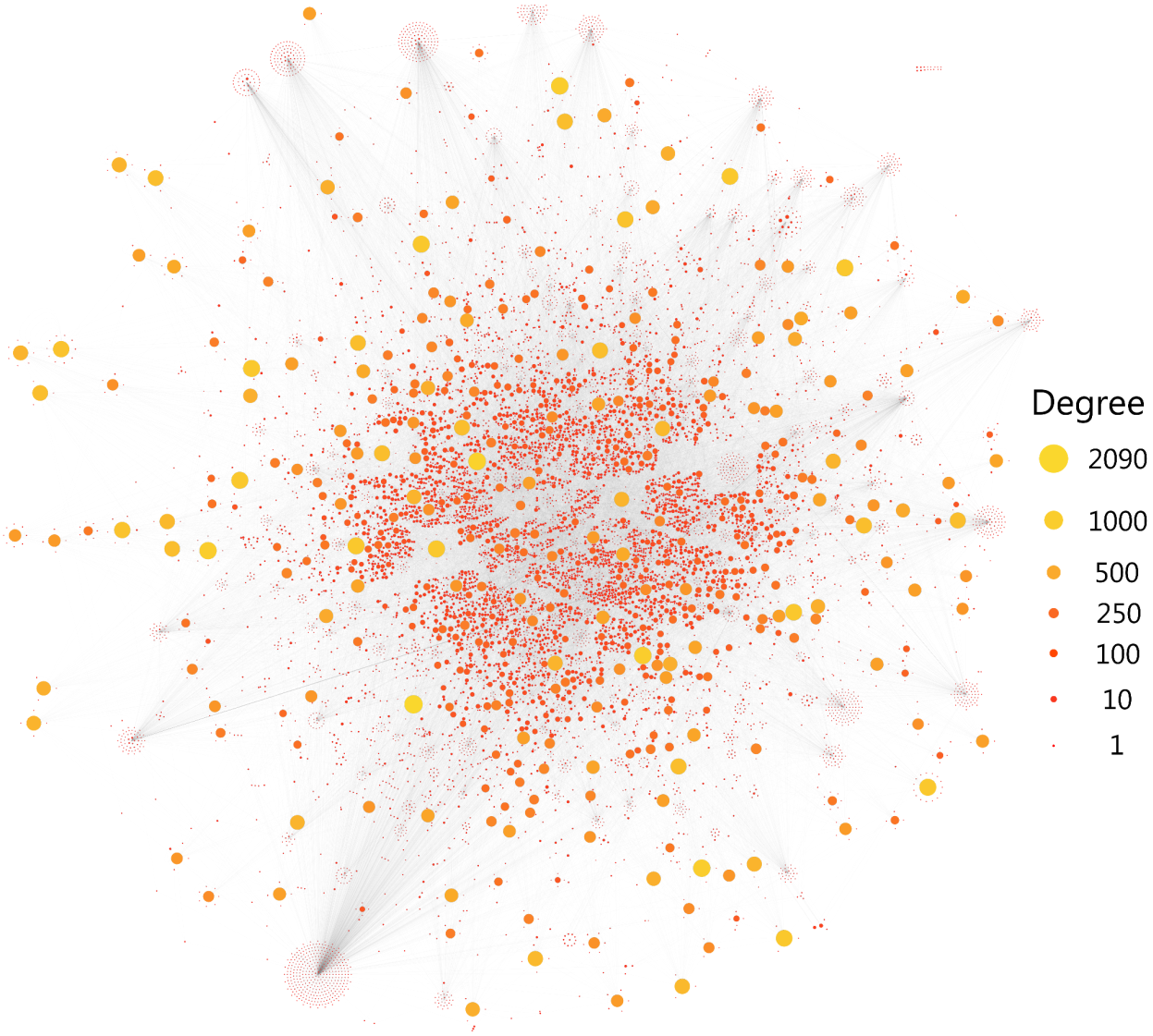
The expression of any single trait in organisms is almost never the result of a single genotypic change. Any change in the body is very likely the result of an extremely complex interplay of thousands of factors, including the expression of multiple genes, pre- and post-translational modification of expression, along with protein interactions, RNA-level interactions, and other factors affecting post-translational expression level. In fact, the fundamental and complete disorders involved in spaceflight neurological dysfunction also heavily imply that the identification of a single "spaceflight gene" is extremely unlikely to be presented as a valid conclusion (1, 2).

This need for integrative, multi-omic analysis is adequately justified by correlation analysis. Every biological, physiological or pathophysiological change in the human body requires an extremely intricate and complex coordination and interactions among multiple genes, proteins, and other modificational factors, and genome correlation has been identified as a good analysis methodology to identify gene regulation and coordination in spaceflight (1). The preflight control dataset in this analysis showed a generally strong level of correlations between genes (Figure 1a). This is expected, as the human body is an extremely ordered system, and pathways work together to accomplish tasks. For example, genes encoding promoters and inhibitors of specific pathways tend to be strongly negatively correlated, whereas two promoter genes tend to be strongly positively correlated. However, we note that these trends do not always hold true, which may be due to a variety of reasons. For example, sets of two promoter genes with antagonistic pathways have been noted to be weakly negatively correlated, and genes in autoregulatory pathways vary widely in average correlation.

We characterize a massive loss of gene-pair correlation across the entire genome in this study, which is very visible in the reduced saturation of the correlation plots shown in Figure 1b. This stark breakdown in genome correlation is an important step towards identifying gene biomarkers based on gene interaction, rather than individual gene expression levels. This loss of gene pair correlation indicates these astronauts' bodies are losing regulation, working against themselves, wasting energy and possibly even activating conflicting pathways simultaneously. The high degree of gene clustering in these correlation plots indicates that gene-pair correlation across the genome is an extremely organized characteristic of gene expression, which could also be used to classify pathways and extract functional insights from gene expression data in astronauts (Supplementary Data 4). Near-complete correlation cluster breakdown in spaceflight indicates that this organization is another key aspect of the genomic response to spaceflight stresses. However, correlation loss clusters were not found to coincide with any specific pathways or gene organization into chromosomes.

As an example of gene-pair correlation loss in a specific pathway, this breakdown in genome correlation is accentuated in correlation plots depicting only neuronal genes, as shown in Figure 2. A larger degree of correlation breakdown in this subset, showcased by higher change in average absolute value correlations across the dataset, implies that these neuronal genes are, to a limited degree, disproportionately affected in the loss in gene-pair correlation across the genome. This is another important confirmation of correlation analysis as a unique spaceflight biosignature for astronauts.

Network analysis


Degree
2090
1000
500
250
100
10
1

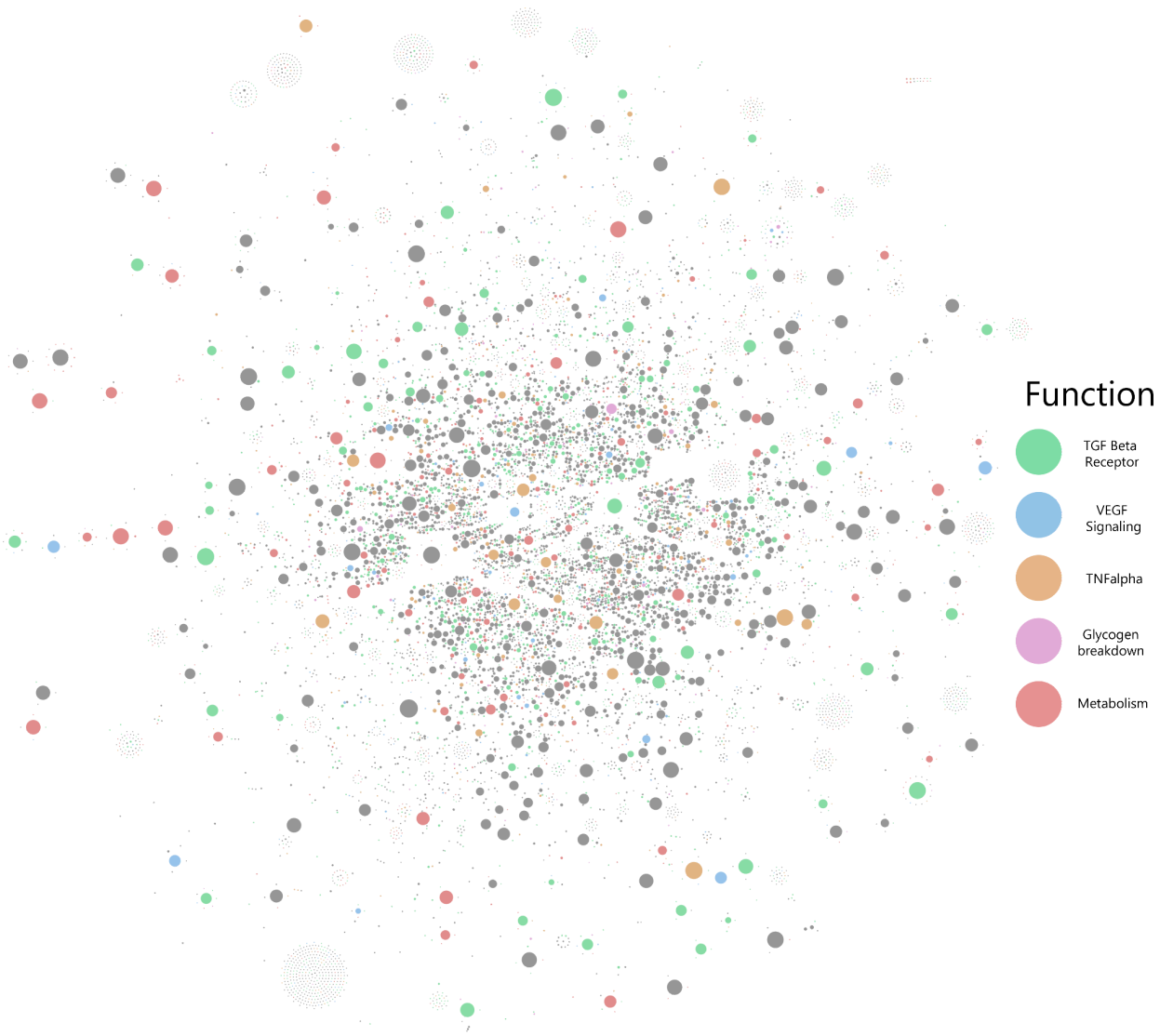
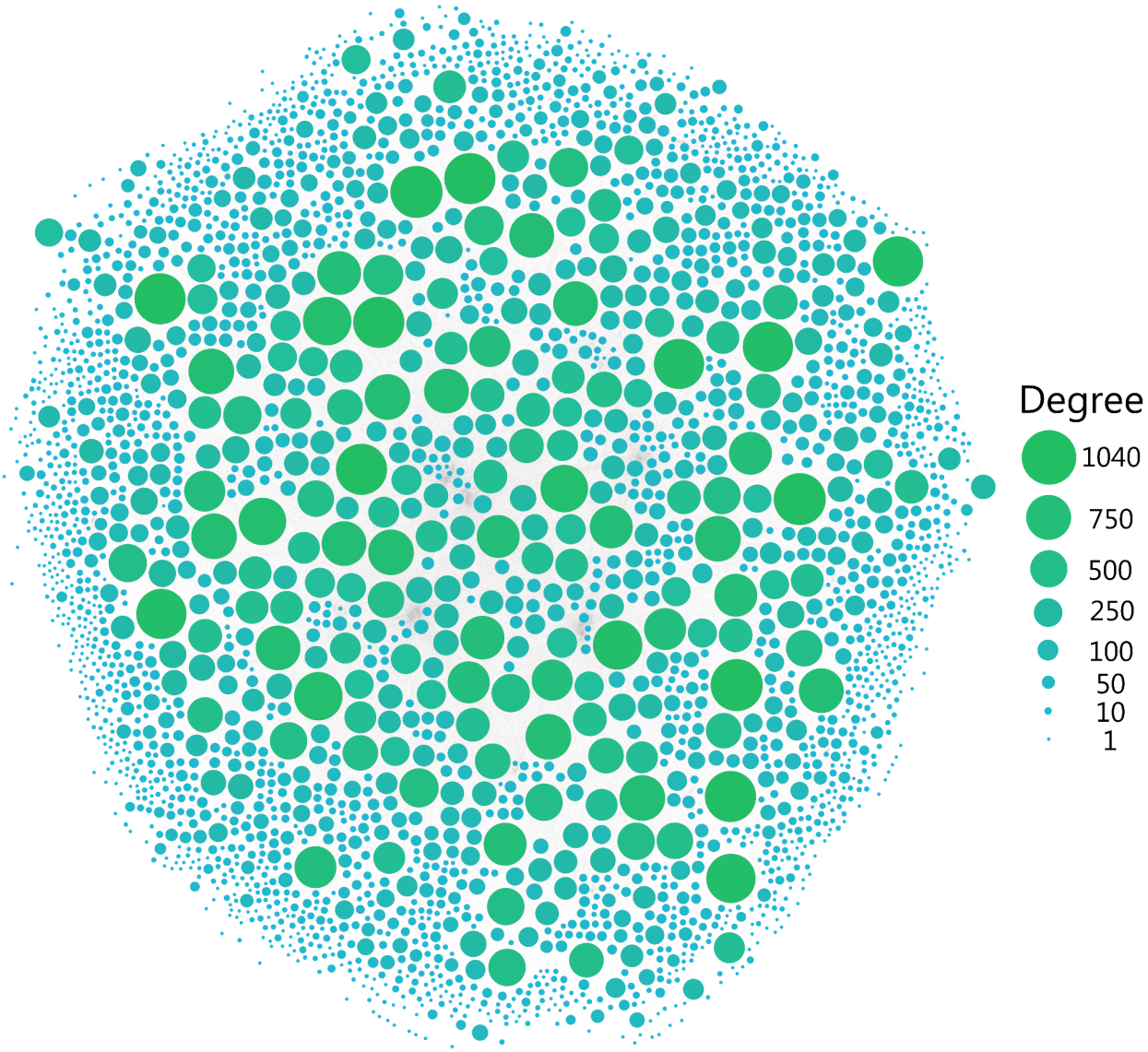


Figure 3: A high degree of protein-protein interactions in differentially correlated and expressed genes implies the existence of controlling genes. Genes with a higher degree of connection are depicted as larger and more orange.

(a) The protein-level interaction network showcases an extremely dense network of interactions, which, alongside the dense miRNA network (Figure 5), implies the existence of specific hub or controlling genes that can drive changes throughout the whole genome

through these interaction networks. Functional analysis in (b) showcases the top 5 cellular functions associated with biological hubs in (a)



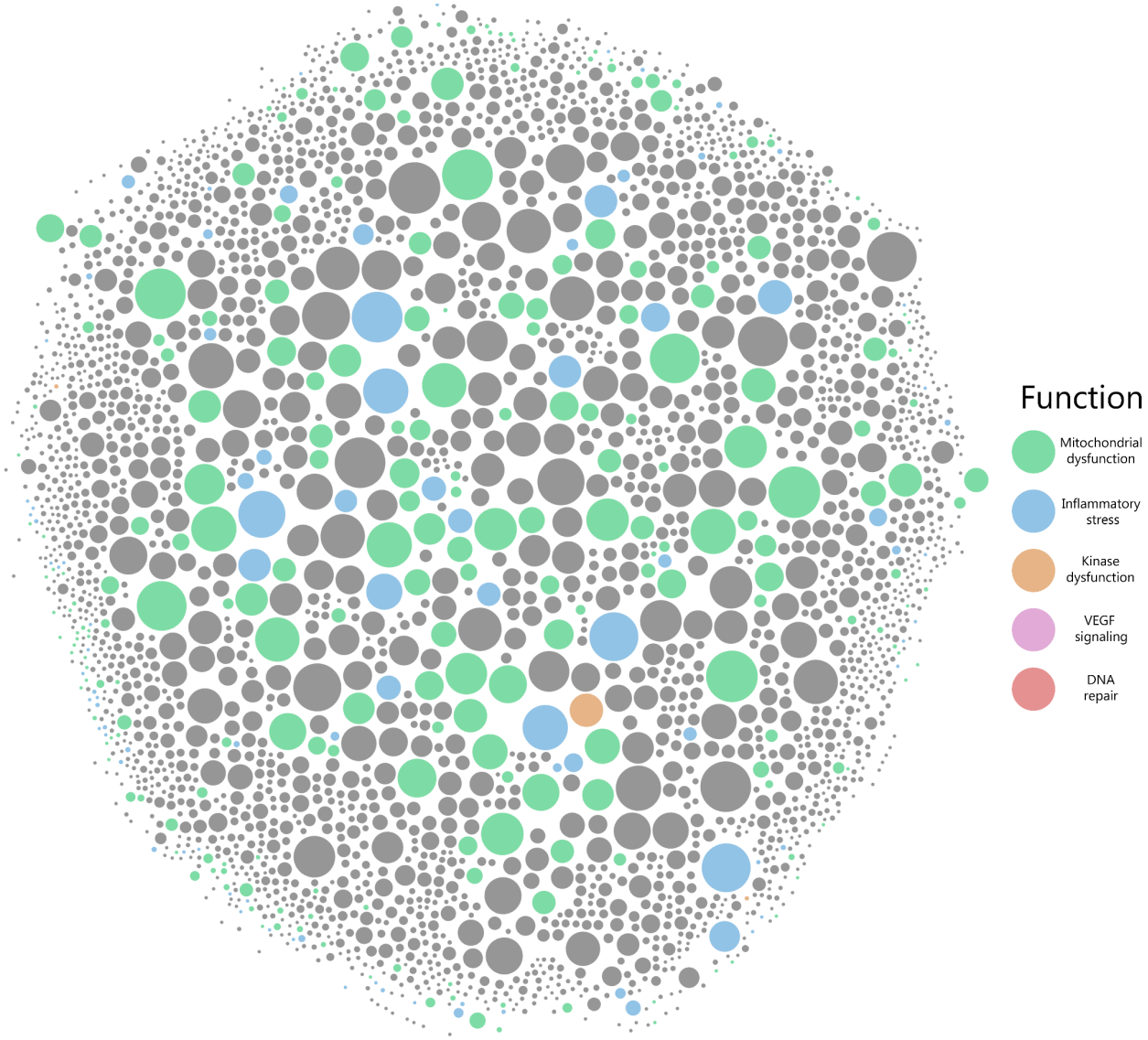


Figure 4: A high degree of miRNA-miRNA interactions in differentially correlated and expressed genes implies the existence of controlling genes.

(a)This plot for genes related by miRNA-level interactions showcases an extremely dense network of interactions, which, alongside the dense protein network shown in Figure 4, implies the existence of specific hub or controlling genes that can drive changes throughout the genome through these networks of interactions. Functional analysis in (b) showcases the top 5 cellular functions associated with biological hubs in (a).

The construction of interaction networks at both the protein-protein (Figure 4) and the miRNA-miRNA (Figure 5) interaction levels allows for the identification and classification of the connectivity and interaction profile of the entire genome using experimental and theoretical interaction databases. The top identified hub pathways have strong ties to mitochondrial dysfunction (4), inflammatory stress (2), DNA repair (2), and carcinogenesis through kinase dysfunction (4).

These network analyses allowed us to identify zero-, first-, or second-degree level interactions for genes with a large degree of connectivity in these networks. The shortlisted genes from these analyses likely act as biological hubs and regulate further networks to cause various pathophysiological symptoms.

An important consequence identified in these networks is mitochondrial stress, leading to kinase dysregulation and neurodysfunction. Mitochondrial dysfunction has been shown to be a key factor in spaceflight-induced disorders (4), and the independent identification of this pathway in this study validates our analysis pipeline in pathway and factor identification. Other notable significantly interconnected genes linked to the mitochondria include Corticotropin Releasing Hormone Receptor 1 (*CRHR1*), Collagen Type VII Alpha 1 (*COL7A1*), Amphiregulin (*AREG*), 3-Phosphoinositide Dependent Protein Kinase 1 (*PDPK1*) and CAMP Responsive Element Binding Protein 1 (*CREB1*). *CRHR1* activation is normally stress-induced, as it regulates the release of stress-related hormones that control the brain. *CRHR1* overexpression is expected in astronauts, as stress-related hormone levels are reported to be increased (2). *CRHR1* is also known to change the morphology of mitochondria in the brain, which is proposed to be part of its mechanism of action for its main functions (15). This includes an association of

CRHR1 with suicidal behavior (16) and major depressive disorder (17). *PDPK1* and *CREB1* are both involved in the tumor synthesis (18) and muscle mass control (19) pathways in *InnateDB* (14). They are also linked to mitochondrial function and control in other spaceflight studies (4).

Discussion and Significance

To the best of our knowledge, this study is the first to combine space gene expression datasets to gain insights on biomarkers to identify a central biosignature (gene-pair expression correlation loss) as avenues for diagnosis, monitoring and treatment in astronauts. Other studies in the field have also shown that individual gene expressions change significantly during spaceflight, and some genes remain altered up to 6 months from the date of return to Earth. However, our results show that the issue is universal - on a far larger scale than previously thought and has the capability to affect a significant number of biological systems.

This study employed a novel method of combining microarray and RNA-Seq datasets to increase statistical significance and reliability. Combining different types of datasets provides a more comprehensive understanding of the molecular mechanisms that underlie genomic health during spaceflight. Microarray and RNA-Seq datasets capture different aspects of gene expression, and combining them allowed us to overcome some of the limitations of each method and obtain a more complete picture of gene expression in astronauts. Data combination also allowed us to identify correlation loss as a central biosignature for spaceflight-induced symptoms in astronauts. This approach has the potential to be useful in other areas of space biology research, as well as in biomedical research. By combining datasets, researchers can obtain a

more comprehensive understanding of biological processes and improve the statistical power of their analyses, ultimately leading to more robust and reliable findings.

The identification of biosignatures for health in astronauts is also critical for understanding the potential long-term effects of spaceflight on the human body. This study incorporates a multi-omic, integrative analysis pipeline for astronaut health. This pipeline combined microarray and RNA-Seq data using differential expression analysis, genome correlation analysis, network analysis, functional analysis, interconnectivity analysis, and principal component analysis to provide a comprehensive analysis of the gene expression data, which allowed for the identification of biomarkers that were not apparent using only one method.

Firstly, principal component analysis allowed us to verify batch correction methods and ensure fair distribution across our normalized data. Differential expression analysis enabled us to identify genes that were differentially expressed in astronauts compared to control preflight datasets, while genome correlation analysis allowed us to identify co-expressed genes that were potentially involved in similar pathways. Network analysis identified groups of genes that were functionally related, while functional analysis allowed us to identify the biological functions associated with these genes. Interconnectivity analysis allowed us to identify genes that were important in connecting different functional pathways.

Space research has significantly contributed to advancements in cancer research and treatment. Previous research has found that cancer cells displayed slower growth rates and changes in gene expression patterns, while others exhibited increased sensitivity to chemotherapy drugs (1). Osteosarcoma cells cultured in microgravity on the ISS displayed morphological changes and exhibited decreased proliferation and increased apoptosis compared

to control groups on Earth (2). Microgravity has also been shown to inhibit angiogenesis and reduce the expression of proteins associated with metastasis (3). The NASA Twins Study, which aimed to reveal the molecular, physiological, and psychological effects of long-term spaceflight, discovered notable differences in several markers related to cancer, including gene expression, increases in inflammation markers, DNA methylation, and telomere length (4). Several studies have indicated that simulated microgravity affects mitochondrial activity and oxidative stress in various cancer cells. These findings highlight the broader impact of microgravity on mitochondrial function and its potential implications in cancer research, DNA damage, and mitosis.

Our analysis has identified genes such as *CRHR1*, *COL7A1*, *AREG*, *PDPK1* and *CREB1*, which are known to be linked to mitochondria dysfunction and cancer. For instance, *CRHR1* signaling was associated with the development of colitis-associated cancer (5), while *COL7A1* was identified as a novel biomarker candidate for gastric cancer (6). *PDPK1* is an upstream regulator of the c-Myc oncogene that promotes cancer cell self-renewal and survival (7). Other studies have shown that *PDPK1* is an essential regulator of cancer stem cell signal transduction (8). Loss of *PDPK1* in BRAF V600 mutant melanoma also inhibits tumorigenesis (9). These studies underscore the importance of understanding how spaceflight and microgravity conditions might influence cancer development and progression through their effects on mitochondrial function.

The identification of correlation loss as a biomarker for astronaut health has important implications for space exploration and human health. The long-term effects of spaceflight on the human body are not well understood, and identifying similar biosignatures will be critical for

understanding these effects. Furthermore, the characterization of differentially correlated genes as opportunities for maintaining and treating health during spaceflight could improve the safety and well-being of astronauts during long-term space travel. The identification of genome-wide correlation loss as a central characteristic of spaceflight samples is an important milestone in characterizing the effects of microgravity, space radiation, and other traumatic effects of spaceflight on the human body. However, further research in this area, such as proteomics and metabolomics studies, is needed to investigate the other implications of correlation loss as a biosignature and central characteristic of space genomics in astronauts.

Methods

Data acquisition

The NASA Open Science Data Repository (OSDR) (7, 8) was filtered to identify human sequencing datasets from spaceflight experiments on the ISS. 9 candidate studies were identified, 5 of which fulfilled the requirements of having both untreated control and untreated flight samples. These studies include OSD-118 (20, 21), OSD-174 (22, 23), OSD-258 (24, 25), OSD-323 (26, 27), and OSD-431 (28, 29). These studies contain data from a variety of cell types and mission lengths, allowing for a large variety in the obtained data and reducing the effect of tissue-specific expression and confounding factors in further analyses. An overview of these studies is showcased in Table 1. Raw, internally-normalized expression data files for these 5 studies were obtained through the OSDR repository.

Study ID	Data type	Cell type	Samples	Mission length
GLDS - 118	DNA microarray	Fibroblast cells	8	30 days
GLDS – 174	DNA microarray	Hair follicles	20	6 months
GLDS - 258	RNA sequencing	Cardiomyocytes	12	39 days
GLDS - 323	RNA sequencing	Intestinal cells	3	15 days
GLDS - 431	RNA sequencing	Endothelial cells	6	10 days

Table 1: Datasets obtained from the NASA OSDR database obtained for this study. *Metadata showcasing the large variation in cell type and mission length in these studies is also shown.*

Data preprocessing

Data preprocessing consisted of normalizing these datasets consistently, and processing them to allow for the combination of these datasets into master preflight and inflight tables. Our data preprocessing and consolidation workflow was based on established methods of combining RNA-Seq and microarray datasets (30).

After data acquisition, all 5 datasets were loaded into an R platform for alignment, normalization, and analysis. For microarray experiments, the *limma* package (31) in an R platform was used to complete quantitative modeling and variance assignment. Experiments and gene reads that did not pass the built-in quality assessment metric in *limma* were removed from these count tables. Probe IDs were then mapped to gene IDs with annotation files. When multiple probes were mapped to the same gene, the arithmetic mean of the values was used as the expression value of the gene.

For RNA-Seq studies, the C++ *fastp* tool (32) was used to perform quality control, filtering, and adapter trimming. This cleaned data was then mapped and quantified using the *STARsolo* tool in C++ (33). Final data pre-processing to obtain expression tables included calculating Transcripts Per Million (TPM) counts, Trimmed Mean of M (TMM) values, and excluding genes with zero counts/TPM/TMM values in all datasets before log-2 transforming count tables. Counts were used in all downstream analyses after screening for zero values. This tool allowed for the creation of count tables for the 3 RNA-Seq datasets used in this study.

Data normalization and consolidation

Once count tables for all 5 datasets were created, data consolidation was addressed. Using the *pandas* package (34) in a Python workflow, the genes in every count table were first sorted by raw expression value, then divided into 200 groups of equal size. This transformed count tables into rankings within each study. The least-square function in SciPy was then used to identify scaling factors for each gene by fitting each expression value to the quadratic equation $e = ar^2 + br + c$, where e is the expression value of the gene and r is the rank of the gene in that profile. The scaling weight w is then calculated as $w = 2ar + b$, using the scaling factors a and b identified earlier. This scaling weight was then simply multiplied by the expression value e to obtain the normalized expression value for each gene.

At this point, the count tables could simply be combined along the gene transcript ID axes. The ComBat batch effect correction function within the *sva* package (35) in R, with default parameters, was then used to analyze for batch effects. This allowed for the creation of master

preflight and inflight expression profiling datasets, combining data from all 5 experiments into two objects for bioinformatic analysis. All further analyses were conducted and verified in both all 5 initial isolated datasets and in the combined datasets to comprehensively validate the robustness of this combination process.

Principal Component Analysis (PCA) was then used to confirm normalization (Figure 5). PCA plots created to verify the normalization procedure showed high mission-based batch clustering of expression profiles before correction, indicating the presence of significant batch effects, which would affect any further analysis conducted on uncorrected data (Figure 5a). This data is secondarily clustered by assay type (Figure 5c). Post-normalization, expression profiles are notably more evenly dispersed across the plot, showcasing the effectiveness of this normalization method (Figure 5b). Post-normalization expression profiles also show significant spaceflight-factor clustering (Figure 5d), indicating good normalization for further analysis.

Finally, the *InnateDB* tool (14) accessed online was used to identify superpathways and subset the genome-level dataset into pathway-level datasets for optional further analyses. Other recent studies (36) have also used workflows to correct for batch effects, with good results for specific correction methods.

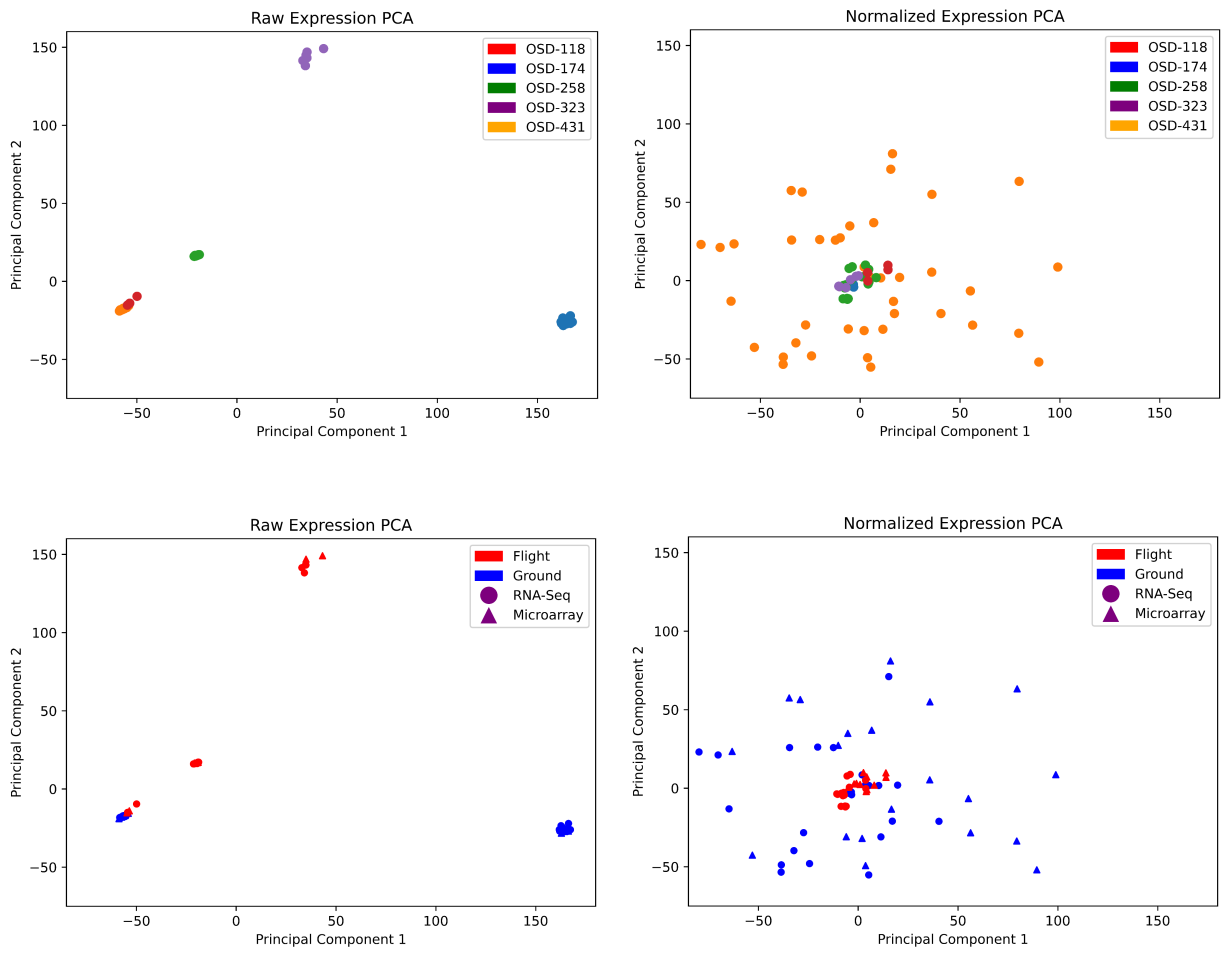


Figure 5: PCA plots demonstrate the efficacy of normalization in correcting spaceflight datasets.

(a) PCA plot of expression profiles prior to normalization reveals primary batch effects based on mission profile (b) PCA plot of expression profiles after normalization showcases a more uniform distribution of data points, indicating successful correction of these batch effects. (c) PCA plot of expression profiles prior to normalization reveals distinct separation between microarray and RNA-Seq assay groups, suggesting secondary batch effects due to assay type. (b) PCA plot of expression profiles after normalization exhibits a significantly more even and dispersed

distribution of data points, indicating a substantial reduction in batch effects and the creation of a more homogeneous dataset.

Differential expression analysis

The *DESeq2* package (37) in R was used to identify Differentially Expressed Genes (DEGs) between the preflight and inflight datasets. The Likelihood Ratio Test (LRT) function was applied, with $p \leq 0.05$ counting as significant. All genes with a p-value above the cutoff were discarded. To reduce the total number of targets and only focus on the most significant genes, DEGs with an absolute $log_2 fold - change$ ($\left|Log_2 FC\right|$) value higher than 3 (most altered genes) were then selected for pathway analysis.

Pathway analysis

The *InnateDB* tool (14) online was used with the output of the differential expression workflow to identify overrepresented pathways and other factors in the genome most altered by spaceflight. The tool's inbuilt gene ontology verified the identified pathways as playing a large variety of roles across the genome. Other genes involved in these pathways were also extracted from *InnateDB*. A correlation analysis on these sub-genesets was subsequently conducted as an exploratory analysis of the gene coordination of these pathways.

Correlation analysis

Correlation analysis was performed to classify gene coordination and overall gene regulation throughout the genome. The *df.corr* function in *pandas* (34) was used to classify the Pearson correlation coefficient between samples. Correlation coefficients were characterized for every single gene against every single other gene in the human genome.

Python and R code using the *pandas* (34) and *corrplot* (38) packages respectively were used to visualize these correlation matrices in correlation plots. These correlation coefficients and plots were then also characterized for neuron-specific genes, which were used as an example to identify and characterize correlation changes in a subset versus the whole genome. The hierarchical clustering linkage function in the *SciPy* library (39) in Python was used to cluster the genome correlation plots and identify genes with similar correlation profiles, both across the genome and in specific pathways. Genes with an absolute correlation coefficient above 0.7, which corresponds to a p-value < 0.00001, were identified as the most correlated genes. These genes were then ranked according to their correlation score and compared between preflight and inflight datasets, allowing for the identification of Differentially Correlated Genes (DCG).

Network analysis

To characterize the likelihood of a small set of genes playing a controlling role in the multisystemic disorder pathophysiology noted in astronauts, interaction networks for the most differentially correlated and expressed genes (n=200) were created. Protein-protein interactions characterized in literature and experiments were obtained via API access of the *BioGrid* database (40). micro- RNA (miRNA) -level interactions of these genes characterized in literature and experiments were obtained via download through public access of the *miRTARBase* database

(41). The interaction tables were constructed by creating a subset of these databases with the gene lists using the *pandas* package (34) in Python. These interaction networks, representing the known interactions of the proteins and miRNAs associated with these genes, were then visualized using *Cytoscape* (42). The interaction data was then screened to identify and shortlist genes with high degrees of interaction, which are most likely to play controlling roles and act as potential biomarkers and genetic factors for spaceflight-induced symptoms.

Principal component analysis

To identify likely causes of the extreme breakdown in gene-pair correlation, a Principal Component Analysis (PCA) was done on the combined master preflight and inflight datasets, as well as the neuron-specific genes using the *Scikit-learn* library (39) in Python. This was done by projecting the master expression datasets (and subsequently, the subset pathway-specific gene datasets) into high-dimensional space, with each gene mapped to one spatial dimension. Linear dimensionality reduction through PCA then allowed for the projection of this data into a lower-dimensional sub-space. Through this analysis, we identified that the top 50 altered genes correlate with more than 90% of the variation in the expression datasets provided.

Conclusions

This study uses a comprehensive, integrative multi-omics analysis pipeline to identify gene-pair expression correlation loss as a central biosignature of spaceflight-induced health issues and symptoms in astronauts. This characteristic of spaceflight is proposed to serve as a

key cause and regulator of pathological symptoms while in spaceflight. Therefore, it could be used for diagnosis, monitoring, and treatment of astronaut health while in space. Identifying gene correlation loss as a potential biomarker and treatment opportunity highlights the importance of further research and the impact of these findings on space exploration and human health. Future studies with more space datasets, such as samples from the ISS, are proposed to serve as optimization and validation platforms for this biosignature characteristic.

Data availability

Source data for this study are publicly available in the NASA Open Science Data Repository (OSDR) (https://osdr.nasa.gov/) under the accession codes OSD-118, OSD-174, OSD-258, OSD-323, and OSD-431.

Code availability

The underlying code for this study is not publicly available, but may be made available to qualified researchers on reasonable request.

Author Contributions

A.S. conceived, designed, and performed this study under the supervision of J.Y. and C.T.. L.S. and S.C. vitally contributed to all aspects of the bioinformatic analyses. A.S., E.L., and L.S. wrote the manuscript, which A.S., C.T., E.L., J.Y., L.S., and S.C. edited. E.L. provided

important support to extend the project, and insight into cancer relevance and importance. All authors approved the submitted version of the manuscript.

Competing interests

All authors declare no financial or non-financial competing interests.

Acknowledgements

This project was funded and supervised through the CPPS 432 Honors program at the University of Saskatchewan.

Citations

1. **Sakharkar, A. & Yang, J.** Designing a Novel Monitoring Approach for the Effects of Space Travel on Astronauts' Health. *Life* **13**, 576 (2023).

2. **Garrett-Bakelman, F. E. *et al.*** The NASA Twins Study: A multidimensional analysis of a year-long human spaceflight. *Science* **364**, eaau8650 (2019).

3. **Smith, S. M. *et al.*** Benefits for bone from resistance exercise and nutrition in long-duration spaceflight: Evidence from biochemistry and densitometry. *Journal of Bone and Mineral Research* **27**, 1896–1906 (2012).

4. **da Silveira, W. A. *et al.*** Comprehensive Multi-omics Analysis Reveals Mitochondrial Stress as a Central Biological Hub for Spaceflight Impact. *Cell* **183**, 1185-1201.e20 (2020).

5. **Afshinnekoo, E. *et al.*** Fundamental Biological Features of Spaceflight: Advancing the Field to Enable Deep-Space Exploration. *Cell* **184**, 6002 (2021).

6. **Gupta, U., Baig, S., Majid, A. & Bell, S. M.** The Neurology of Space Flight; How does space flight effect the human nervous system? *Life Sciences in Space Research* **36**, (2022).

7. **Ray, S. *et al.*** GeneLab: Omics database for spaceflight experiments. *Bioinformatics* **35**, 1753–1759 (2018).

8. **Berrios, D. C., Galazka, J., Grigorev, K., Gebre, S. & Costes, S.** NASA GeneLab: interfaces for the exploration of space omics data. *Nucleic Acids Research* **49**, (2020).

9. **Gu, A.-D.** . A novel experimental approach for systematic identification of box H/ACA snoRNAs from eukaryotes. *Nucleic Acids Research* **33**, e194–e194 (2005).

10. **Fiselier, A., Byeon, B., Ilnytskyy, Y., Kovalchuk, O. & Kovalchuk, I.** Scatter Irradiation of Rat Brain Triggers Sex- and Brain Region-Specific Changes in the Expression of Non-Coding RNA Fragments. *Epigenomes* **6**, 35 (2022).

11. **Zhang, X., Hamblin, M. H. & Yin, K.-J.** Noncoding RNAs and Stroke. *The Neuroscientist* **25**, 22–26 (2018).

12. **Laneve, P., Tollis, P. & Caffarelli, E.** RNA Deregulation in Amyotrophic Lateral Sclerosis: The Noncoding Perspective. *International Journal of Molecular Sciences* **22**, 10285–10285 (2021).

13. **Robberecht, W. & Philips, T.** The changing scene of amyotrophic lateral sclerosis. *Nature Reviews Neuroscience* **14**, 248–264 (2013).

14. **Breuer, K. *et al.*** InnateDB: systems biology of innate immunity and beyond—recent updates and continuing curation. *Nucleic Acids Research* **41**, D1228–D1233 (2012).

15. **Battaglia, C. R., Cursano, S., Enrico Calzia, Catanese, A. & Boeckers, T. M.** Corticotropin-releasing hormone (CRH) alters mitochondrial morphology and function by activating the NF-kB-DRP1 axis in hippocampal neurons. *Cell Death and Disease* **11**, (2020).

16. **De la Cruz-Cano, E.** Association between FKBP5 and CRHR1 genes with suicidal behavior: A systematic review. *Behavioural Brain Research* **317**, 46–61 (2017).

17. **Diana María Dionisio-García *et al.*** DNA Methylation of Genes Involved in the HPA Axis in Presence of Suicide Behavior: A Systematic Review. *Brain Sciences* **13**, 584–584 (2023).

18. **Martin-Padron, J. *et al.*** Plakophilin 1 enhances MYC translation, promoting squamous cell lung cancer. *Oncogene* **39**, 5479–5493 (2019).

19. **Kuramoto, N. *et al.*** Role of PDK1 in skeletal muscle hypertrophy induced by mechanical load. *Scientific Reports* **11**, (2021).

20. **Zhang, Y. *et al.*** Transient gene and microRNA expression profile changes of confluent human fibroblast cells in spaceflight. *The FASEB Journal* **30**, 2211–2224 (2016).

21. **Wu, H. & Zhang, Y.** Transient gene expression profile changes of confluent human fibroblast cells in spaceflight. (2017) doi:https://doi.org/10.26030/rcytqp10.

22. **Terada, M. *et al.*** Effects of a Closed Space Environment on Gene Expression in Hair Follicles of Astronauts in the International Space Station. *PLOS ONE* **11**, e0150801 (2016).

23. **Seki, M. & Terada, M.** Effects of a Closed Space Environment on Gene Expression in Hair Follicles of Astronauts in the International Space Station. (2016) doi:https://doi.org/10.26030/6sg0ng36.

24. **Wnorowski, A. *et al.*** Effects of Spaceflight on Human Induced Pluripotent Stem Cell-Derived Cardiomyocyte Structure and Function. *Stem Cell Reports* **13**, 960–969 (2019).

25. **Wu, J.** Effects of Spaceflight on Human Induced Pluripotent Stem CellDerived Cardiomyocyte Structure and Function. (2019) doi:https://doi.org/10.26030/gfew5417.

26. **Barrila, J. *et al.*** Evaluating the effect of spaceflight on the host–pathogen interaction between human intestinal epithelial cells and Salmonella Typhimurium. *npj Microgravity* **7**, 1–10 (2021).

27. **Nickerson, C. *et al.*** Evaluating the effect of spaceflight on the host-pathogen interaction between human intestinal epithelial cells and Salmonella Typhimurium. (2021) doi:https://doi.org/10.26030/hdr22940.

28. **Barravecchia, I. *et al.*** Microgravity and space radiation inhibit autophagy in human capillary endothelial cells, through either opposite or synergistic effects on specific molecular pathways. *Cellular and Molecular Life Sciences* **79**, (2021).

29. **Forcato, M.** Microgravity and space radiation exert either opposite or synergic effects on single molecular pathways. (2021) doi:https://doi.org/10.26030/hv55br83.

30. **Tang, K. *et al.*** Rank-in: enabling integrative analysis across microarray and RNA-seq for cancer. *Nucleic Acids Research* **49**, e99–e99 (2021).

31. **Ritchie, M. E. *et al.*** limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* **43**, e47–e47 (2015).

32. **Chen, S., Zhou, Y., Chen, Y. & Gu, J.** fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).

33. **Kaminow, B., Yunusov, D. & Dobin, A.** STARsolo: accurate, fast and versatile mapping/quantification of single-cell and single-nucleus RNA-seq data. *bioRxiv (preprint)* (2021) doi:https://doi.org/10.1101/2021.05.05.442755.

34. **McKinney, W.** Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science Conference* **445**, (2010).

35. **Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D.** The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**, 882–883 (2012).

36. **Sanders, L. *et al.*** Batch effect correction methods for NASA GeneLab transcriptomic datasets. *Frontiers in Astronomy and Space Sciences* **10**, (2023).

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

38. **Wei, T. & Simko, V.** Visualization of a Correlation Matrix [R package corrplot version 0.84]. (2017).

39. **Virtanen, P. *et al.*** SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods* **17**, 261–272 (2020).

40. **Oughtred, R. *et al.*** The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Science* **30**, 187–200 (2020).

41. **Huang, H.-Y. *et al.*** miRTarBase update 2022: an informative resource for experimentally validated miRNA–target interactions. *Nucleic Acids Research* **50**, D222–D230 (2021).

42. **Shannon, P.** Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research* **13**, 2498–2504 (2003).