



Transcriptomics-based Machine Learning (ML) Analysis Predicts Space-Exposed Murine Livers

Hari Ilangoan SAIC | NASA Langley Research Center
Prachi Kothiyal Scimentis LLC
Katherine Hoadley University of North Carolina - Chapel Hill
Robin Elgart University of Houston
Newton Campbell AROSE
Greg Eley Scimentis LLC
Parastou Eslami Universal Artificial Intelligence LLC



Translational Radiation Research and Countermeasures

- Formed by the Space Radiation (SR) Element under authority of the NASA Human Research Program (HRP)
- **Mission:** translate radiation research results based on animal studies to humans using bioinformatics and computational modeling
- Cross disciplinary team: SR Element Scientists, NASA Information Data & Analytics Services (IDAS) Data Scientists, Bioinformaticians & Computational biologists



Parastou Eslami
Project Scientist



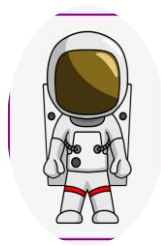
Nicholas Meyer
Project Manager



Prachi Kothiyal
Bioinformatician



Hari Ilangovan
Data Scientist



Greg Eley
Bioinformatician



Newtown Campbell
Data Scientist



Katherine
Hoadley
Computational
Biologist



Ryan Norman
Element Scientist



Brock Sishc
Cancer Discipline
Scientist



Janapriya Saha
CVD Discipline
Scientist



Janice Zawaski
Deputy Element
Scientist

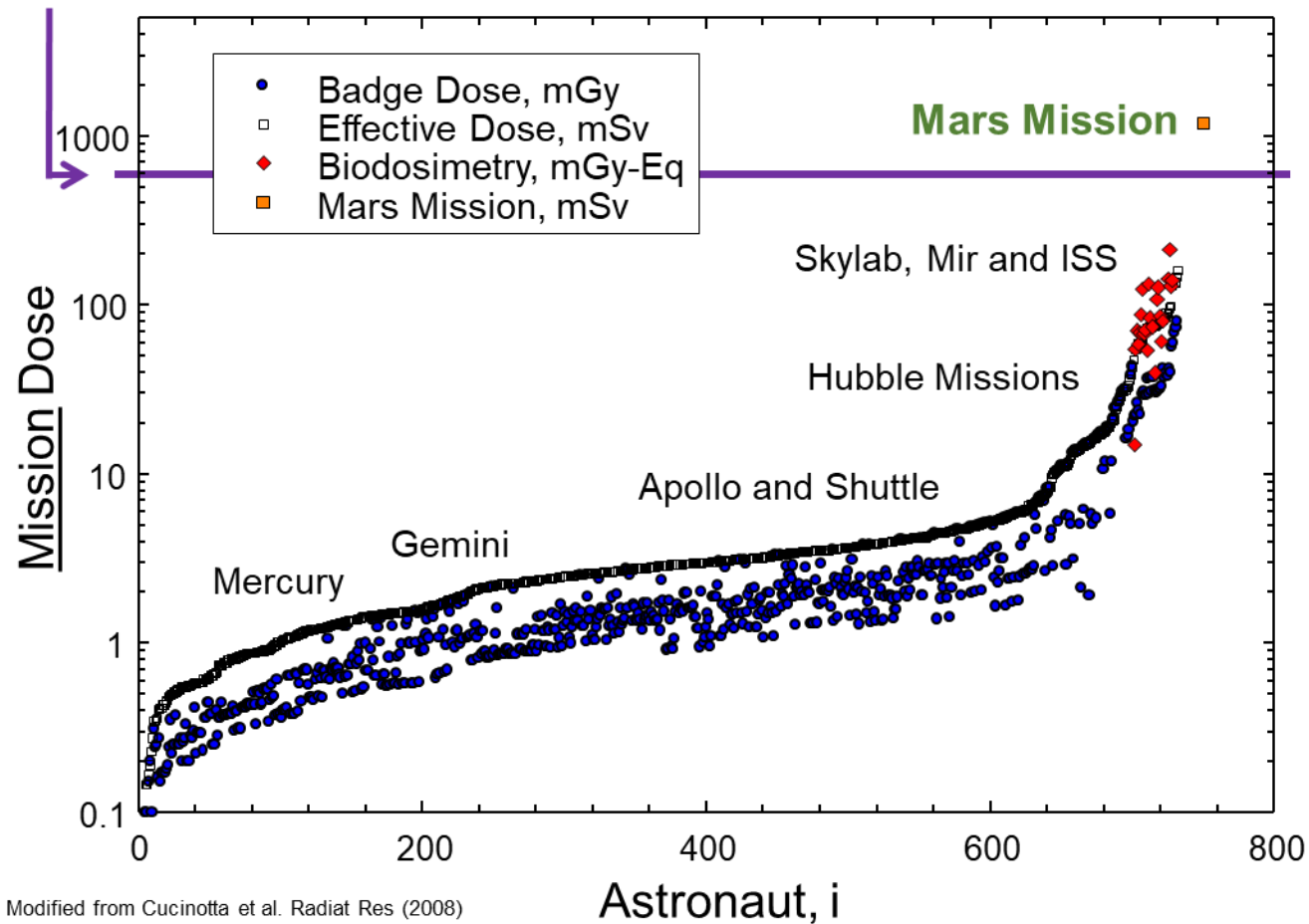


Robin Elgart
Element Scientist



Mission Radiation Environments

NASA Career Radiation Limit (600 mSv)



Long duration spaceflight (or even repeated short missions) will exceed current career radiation limits meant to minimize cancer risk

Cancer can **seriously and permanently** impact astronaut long-term health

Galactic Cosmic Radiation (GCRs) and microgravity are being studied using biological endpoints from rodent based models



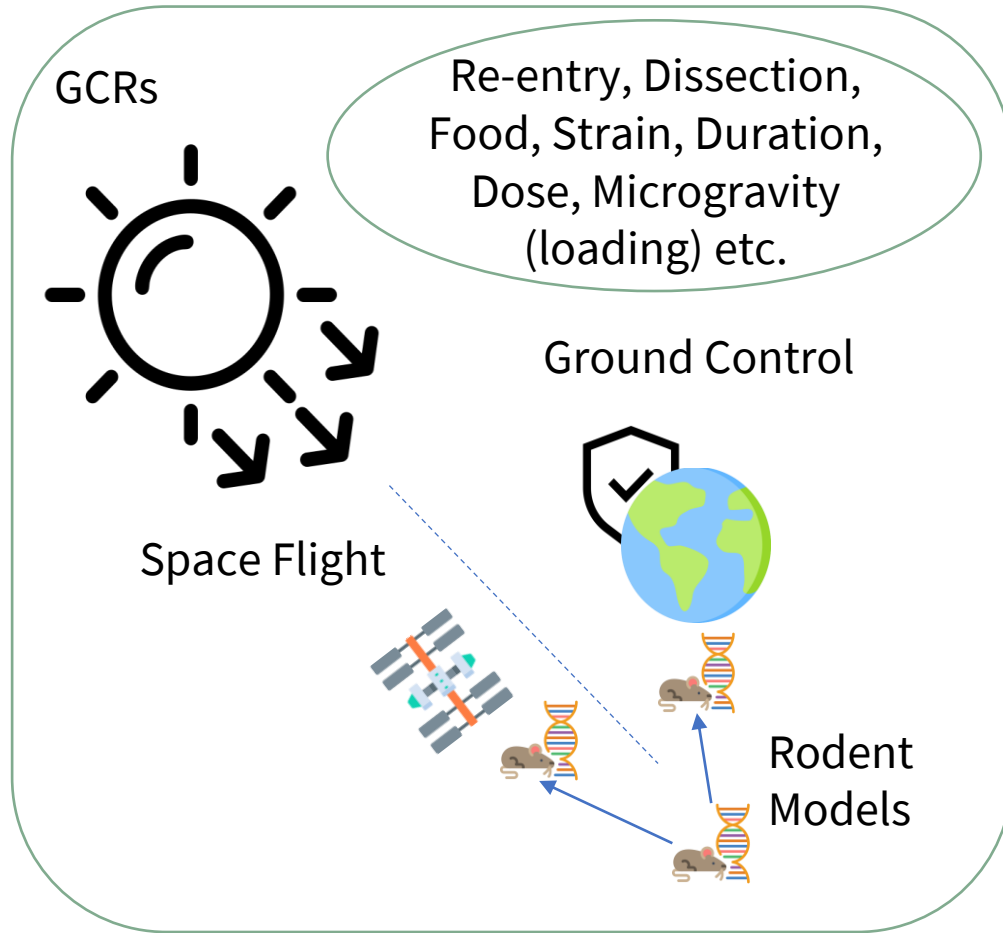
Cucinotta, Francis A, Myung-Hee Y Kim, and Lori J Chappell. "Space Radiation Cancer Risk Projections and Uncertainties – 2012," 2012, 186.

Sishc, Brock J., Janice Zawaski, Janapriya Saha, Lisa S. Carnell, Kristin M. Fabre, and S. Robin Elgart. "The Need for Biological Countermeasures to Mitigate the Risk of Space Radiation-Induced Carcinogenesis, Cardiovascular Disease, and Central Nervous System Deficiencies." *Life Sciences in Space Research*, June 14, 2022. <https://doi.org/10.1016/j.lssr.2022.06.003>.

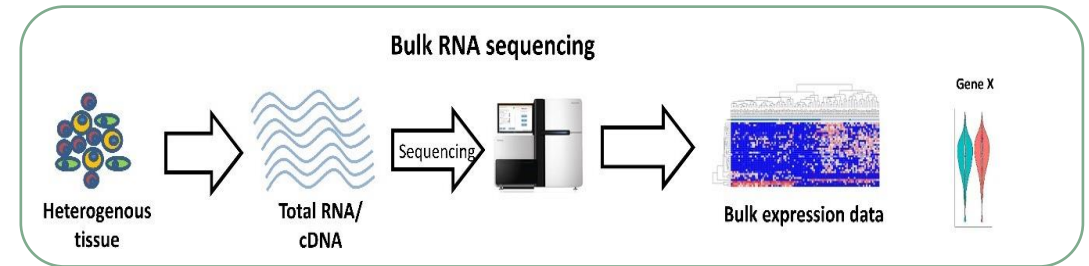


Space Studies & Bulk RNA-Seq Data

Space Flight Experiments



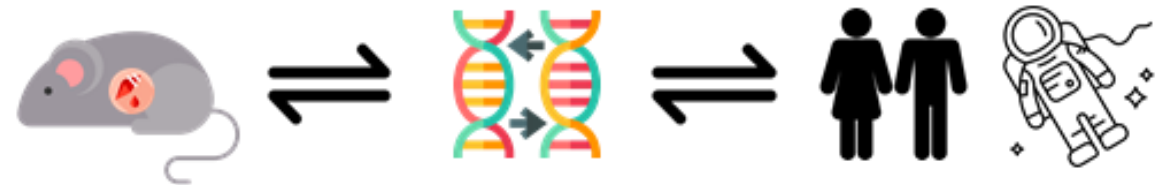
Heterogeneous Experiments



Confounding Factors

Biological and Technical Variability

Translate results to astronaut relevant data sets



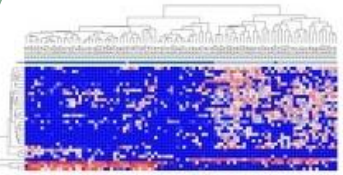


Space Studies & Bulk RNA-Seq Data



Characteristics of Studies with RNA-Seq Data

Sparsity



Bulk expression data

x_1	x_2	x_3	...	x_n	y
			...		
			...		

Sample Size, n

x_1	x_2	x_3	...	x_n	y

Between Studies:

- 1) $D \gg n$
- 2) $n_1 \neq n_2 \neq n_3$

x_1	x_2	x_3	...	x_n	y
			...		
			...		

Data Dimensionality, D

high D
small n

Differential Gene Expression Analysis & Meta Analysis

Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2

[Michael I Love](#), [Wolfgang Huber](#) & [Simon Anders](#)



Leverage best practices, lessons learned (data handling, model assumptions, etc.)

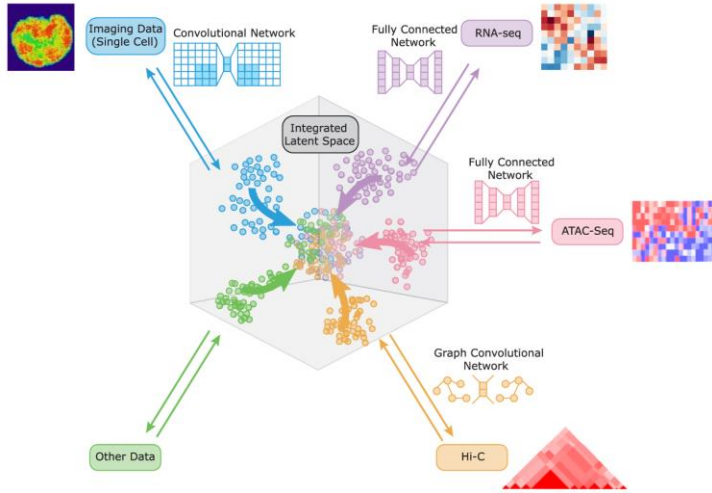


How can we apply Machine Learning techniques?



ML Applications to Omics Data & Space Biology

Scope: Predicting space-exposed murine livers using transcriptomics data



Dimensionality
Reduction

Supervised Machine
Learning Methods

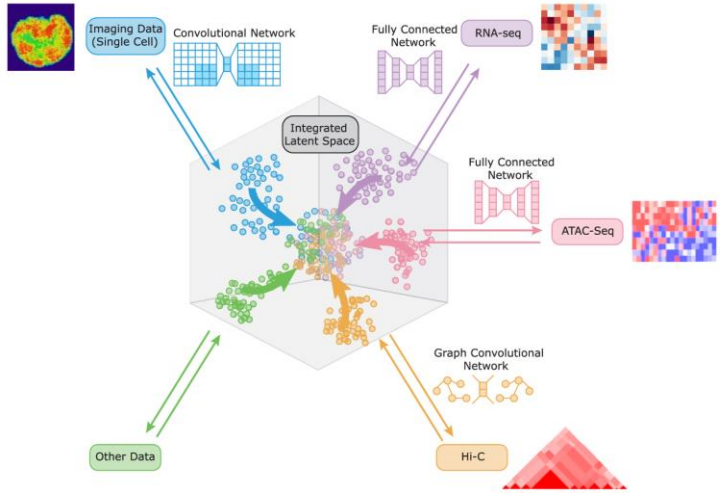
Transform/subset
features into
latent/reduced set

Classification
Space Flight vs
Ground Control



ML Applications to Omics Data & Space Biology

Scope: Predicting space-exposed murine livers using transcriptomics data



Dimensionality
Reduction

Supervised Machine
Learning Methods

Generalization

Interpretability

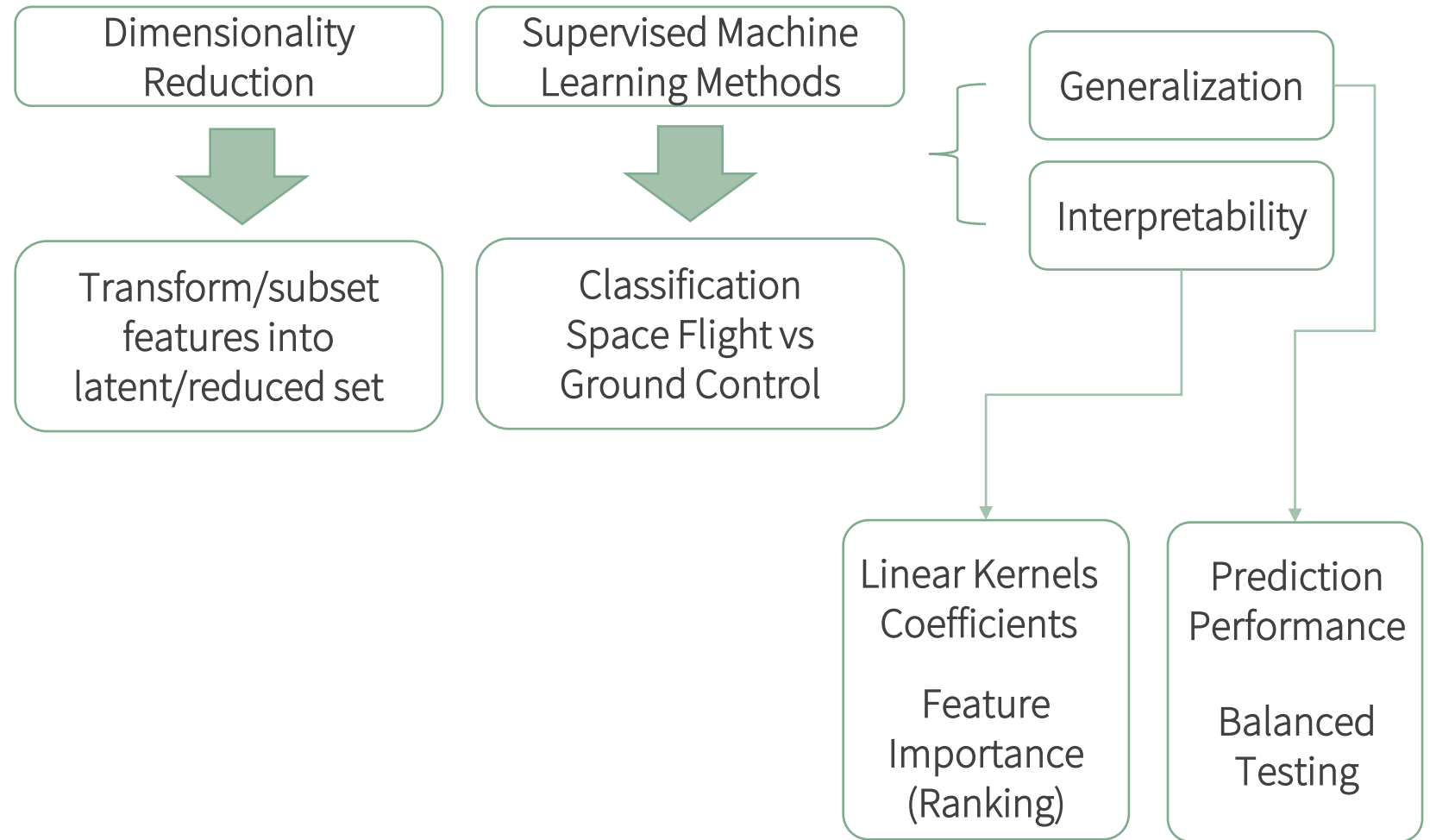
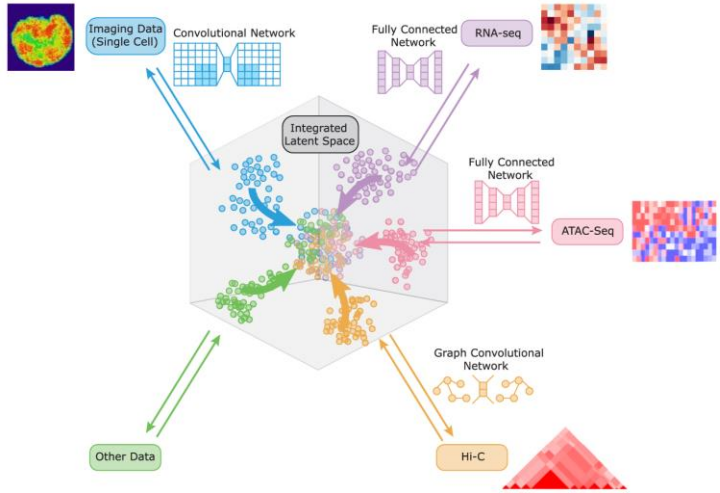
Transform/subset
features into
latent/reduced set

Classification
Space Flight vs
Ground Control



ML Applications to Omics Data & Space Biology

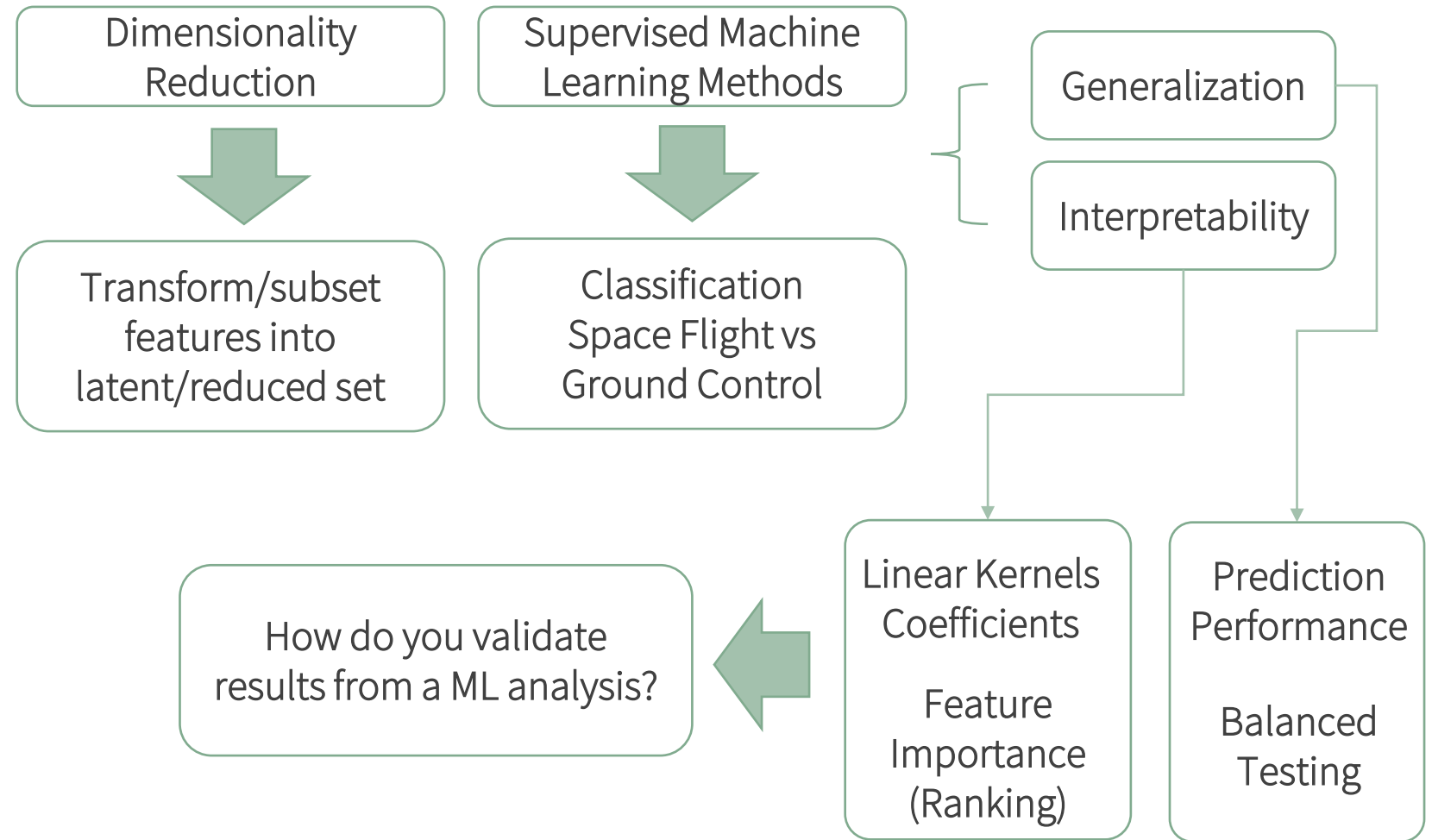
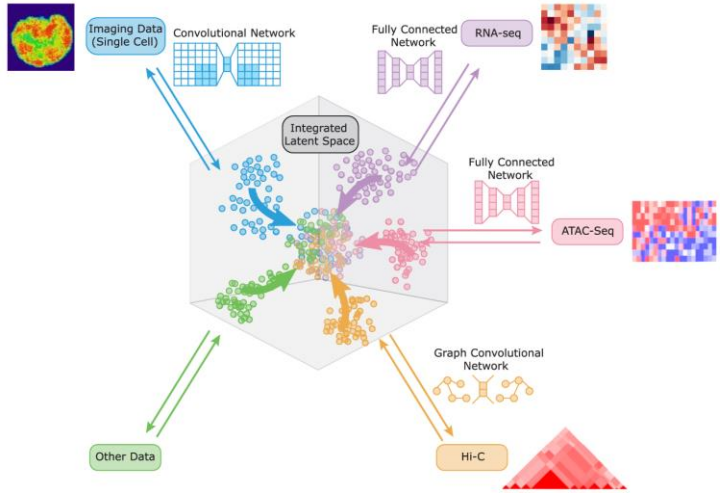
Scope: Predicting space-exposed murine livers using transcriptomics data





ML Applications to Omics Data & Space Biology

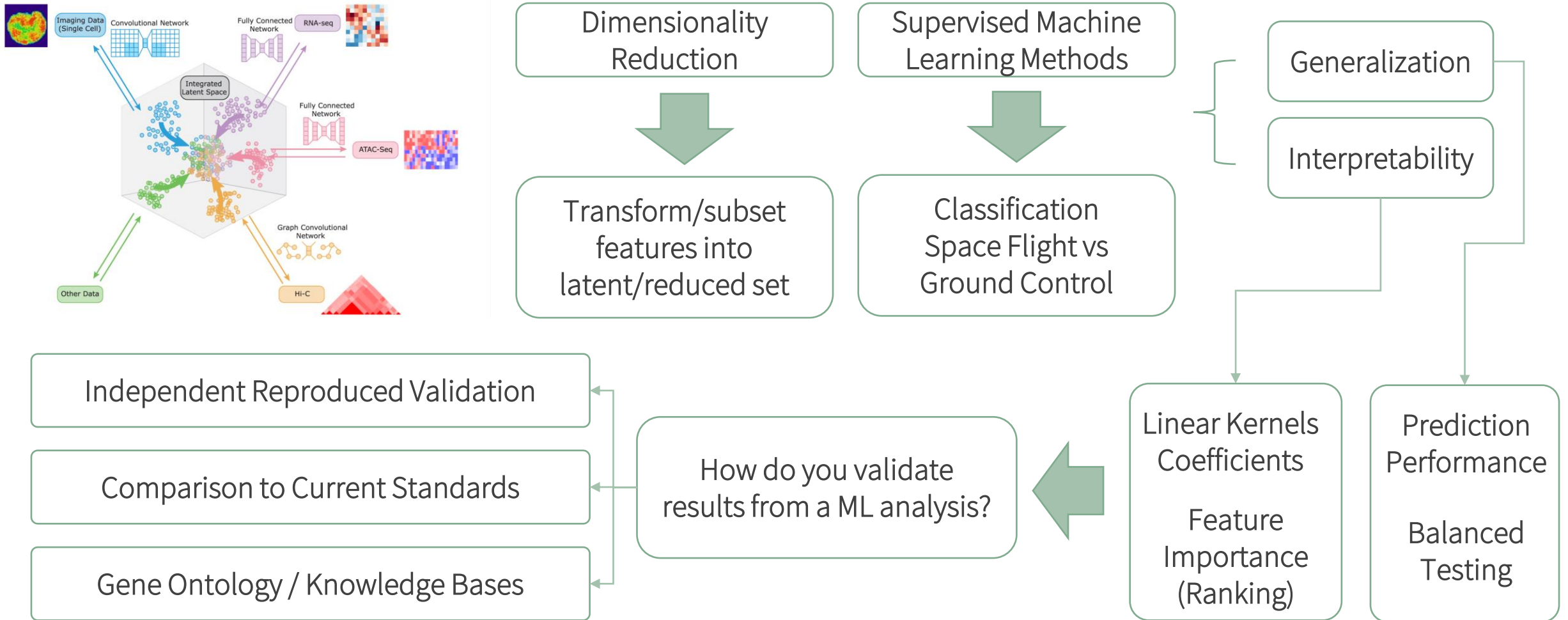
Scope: Predicting space-exposed murine livers using transcriptomics data





ML Applications to Omics Data & Space Biology

Scope: Predicting space-exposed murine livers using transcriptomics data





ML Applied to Murine Liver Datasets

Sample subset to
Ground Control & Space Flight

Tissue = *Liver*

All Studies:
 $n(\text{SF}|\text{GC}) = 81 \text{ (37|44)}$

RR-1 CASIS (9/20/14 – 10/25/14)
Duration = **21** days
Strain = C57BL/6Tac
 $n(\text{SF}|\text{GC}) = 6 \text{ (3|3)}$

RR-1 (9/21/14 – 10/25/14)
Duration = **37** days
Strain = C57BL/6J
 $n(\text{SF}|\text{GC}) = 10 \text{ (5|5)}$
Carcass Dissection only used per Nature

RR-9 (8/14/17 – 6/26/18)
Duration = **33** days
Strain = C57BL/6J
 $n(\text{SF}|\text{GC}) = 18 \text{ (5|13)}$

RR-6 (12/15/17 – 1/16/18)
Duration $\in \{1, 30, 60\}$ days
Strain = C57BL/6NTac
 $n(\text{SF}|\text{GC}) = 39 \text{ (20|19)}$

RR-3 (4/08/16 – 5/11/16)
Duration = $\in \{39, 40, 41\}$ days
Strain = BALB/c
 $n(\text{SF}|\text{GC}) = 8 \text{ (4|4)}$

Heterogeneity needs to be
addressed before models
can be trained appropriately

Combined study Analysis
preferred over Study by Study



Sample Size



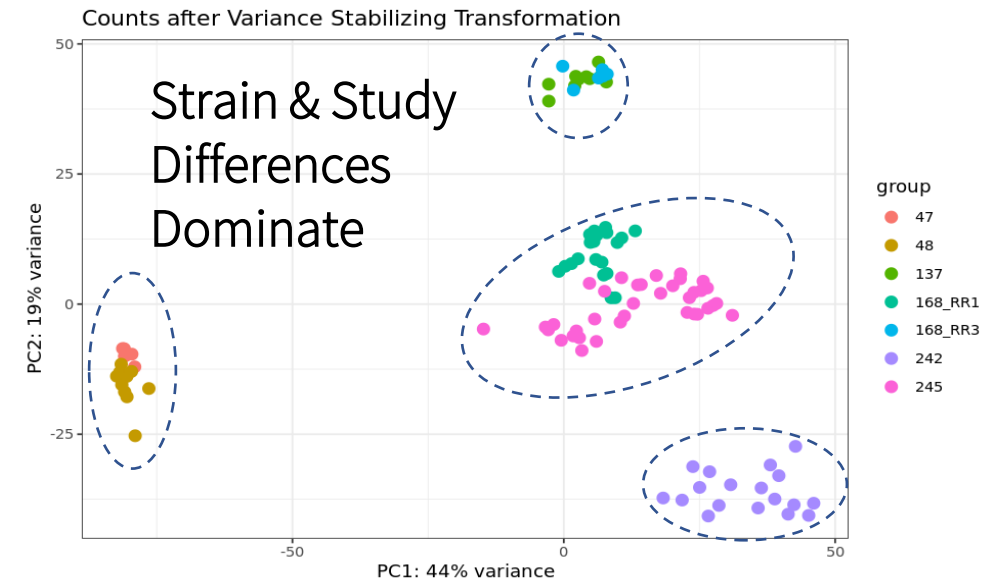
Heterogeneity



Generalization



Number of Applicable Models

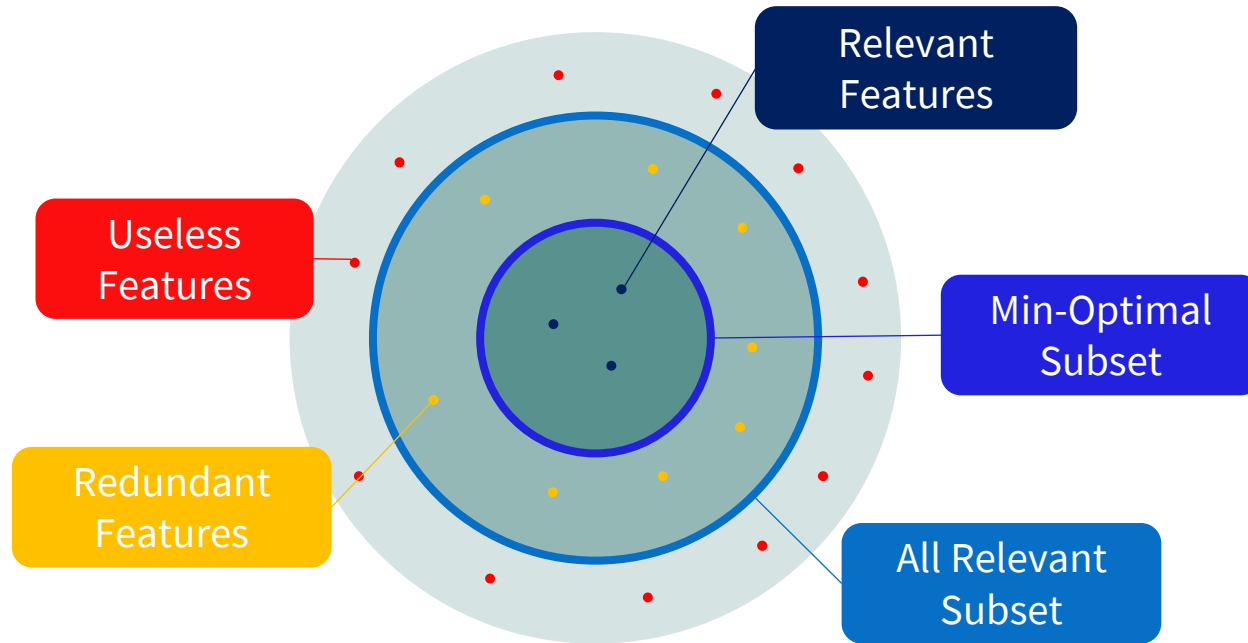




Feature Selection Methods



- Minimum Redundance Maximum Relevance (mRMR) Feature Selection



$$\text{Input data } \mathbf{X} = \begin{bmatrix} x_{00} & \cdots & x_{0m} \\ \vdots & \ddots & \vdots \\ x_{n0} & \cdots & x_{nm} \end{bmatrix} n \times m$$

$$\text{Target vector } c \text{ in the form of } \mathbf{Y} = \begin{bmatrix} y_0 \\ \vdots \\ y_n \end{bmatrix} n \times 1$$

Max Relevance

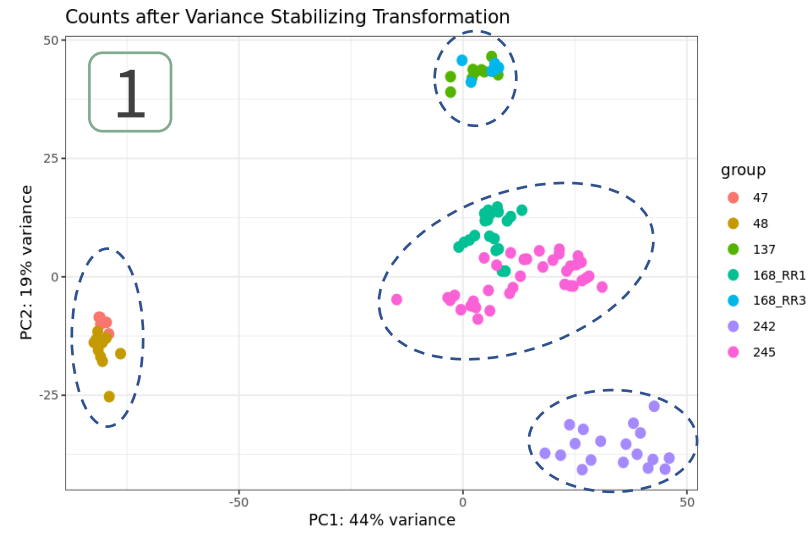
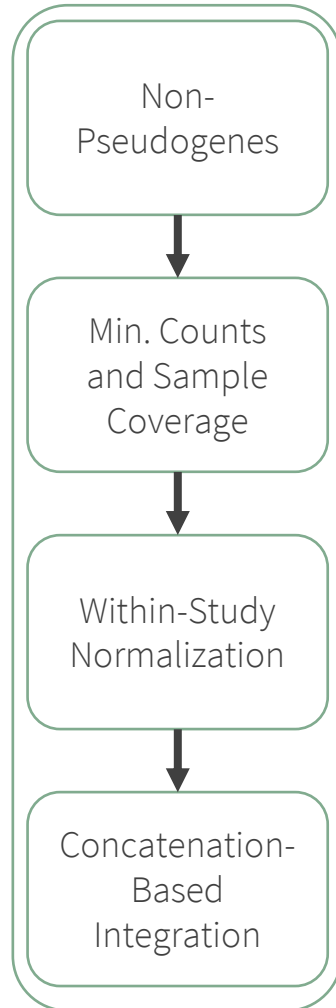
$$\max D(S, c), \quad D = \frac{1}{|S|} \sum_{x_i \in S} I(x_i; c).$$

Min Redundancy

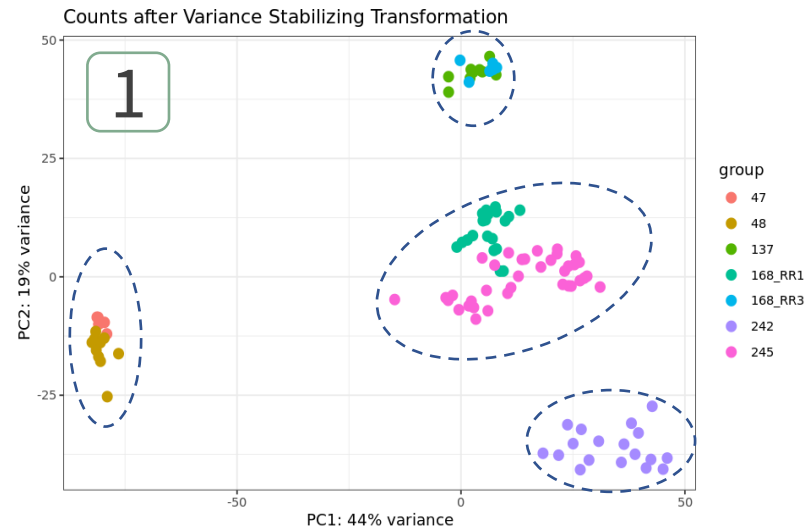
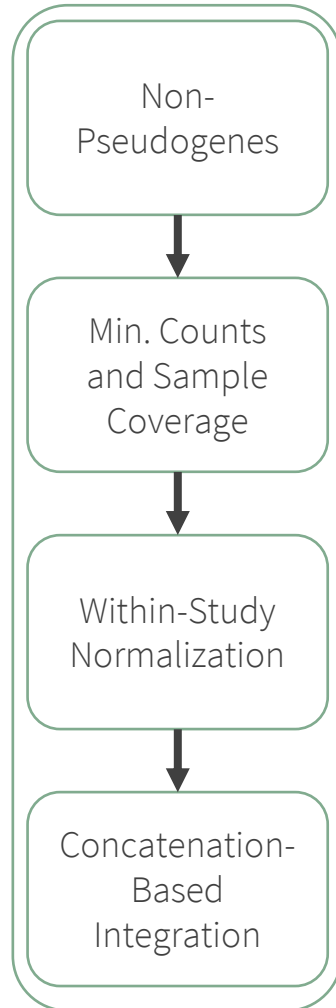
$$\min R(S), \quad R = \frac{1}{|S|^2} \sum_{x_i, x_j \in S} I(x_i, x_j).$$

$$\max \Phi(D, R), \quad \Phi = D - R.$$

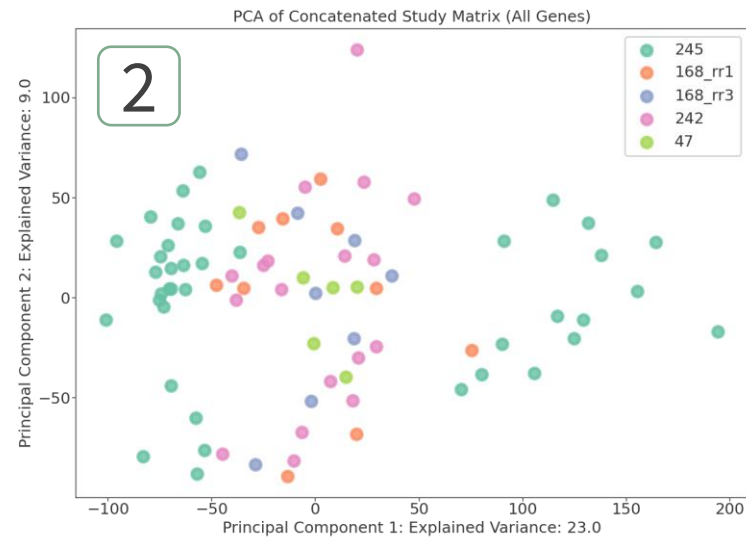
Pre-Filtering & Normalization



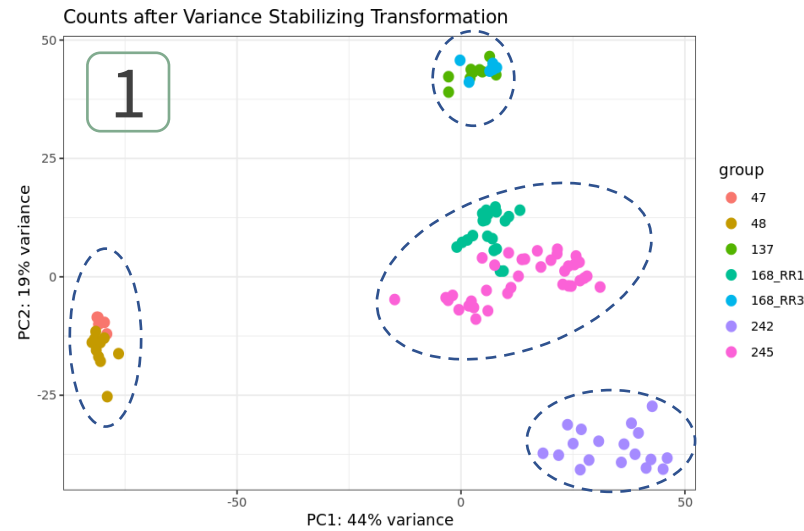
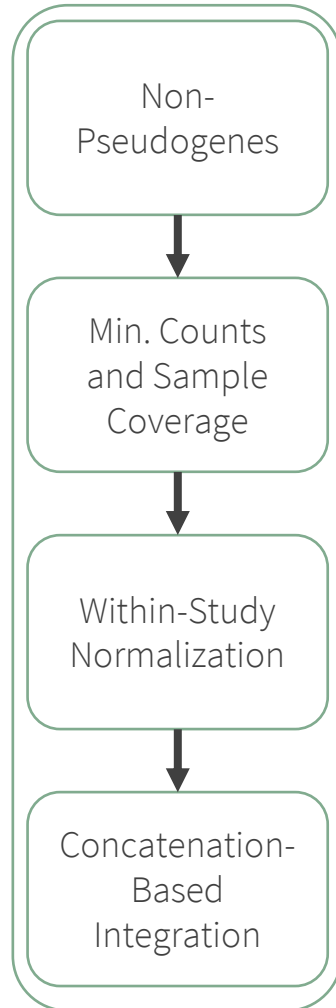
Pre-Filtering & Normalization



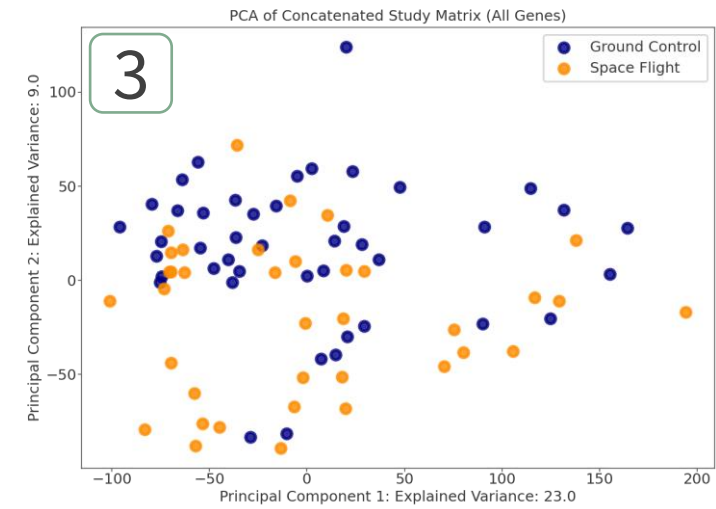
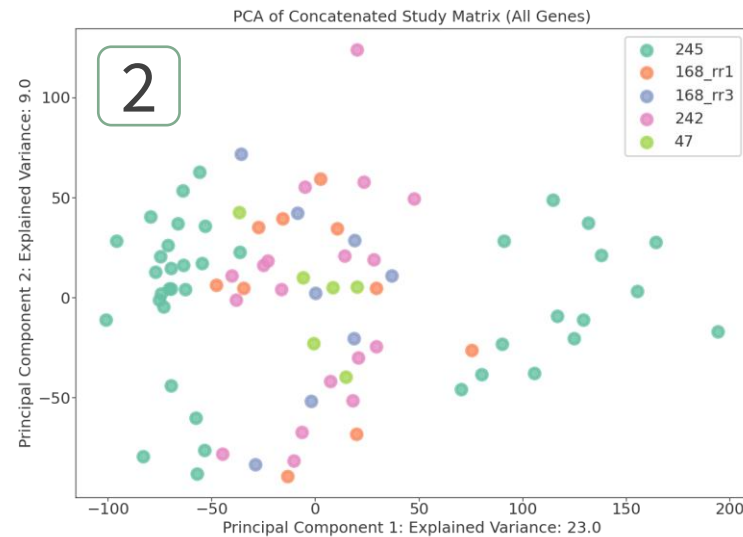
Remove technical replications
and perform transformations &
prefiltering



Pre-Filtering & Normalization



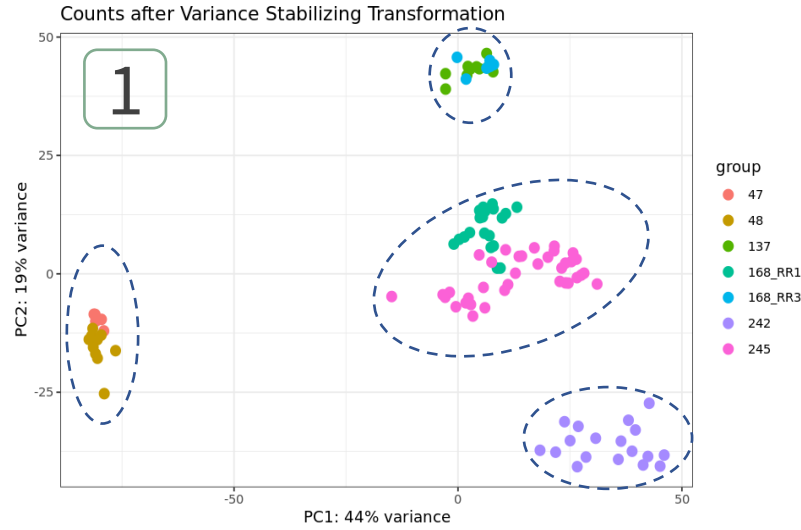
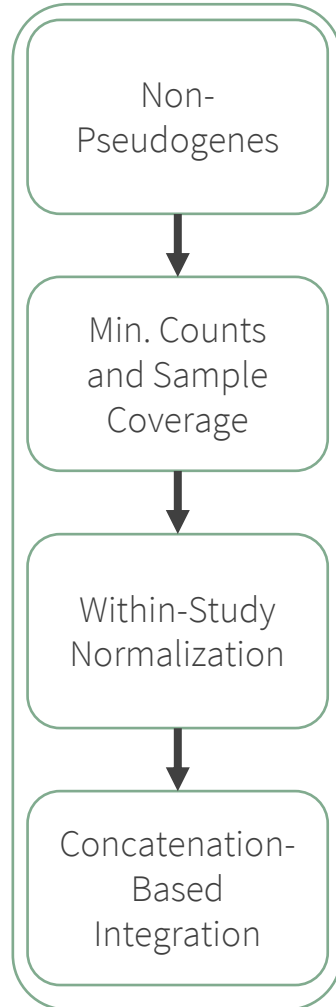
Remove technical replications and perform transformations & prefiltering



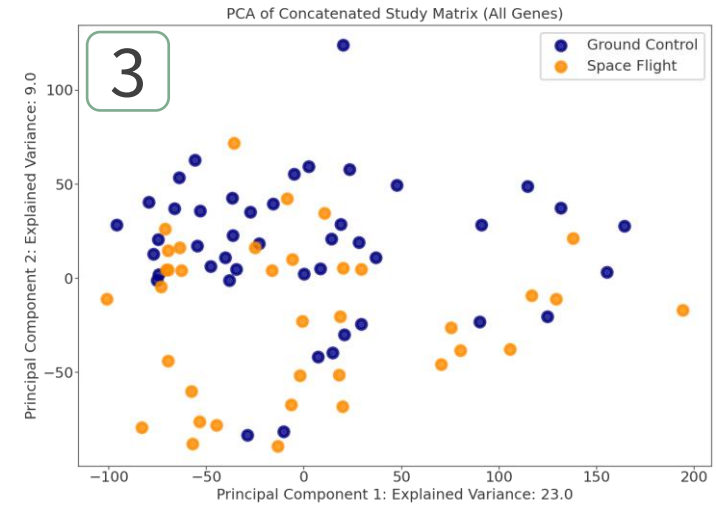
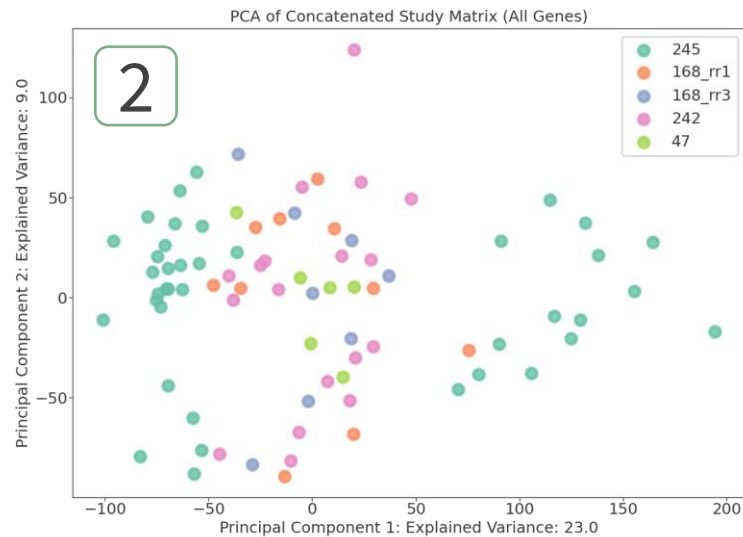
Subset to features selected via Maximum Relevance Minimum Redundance



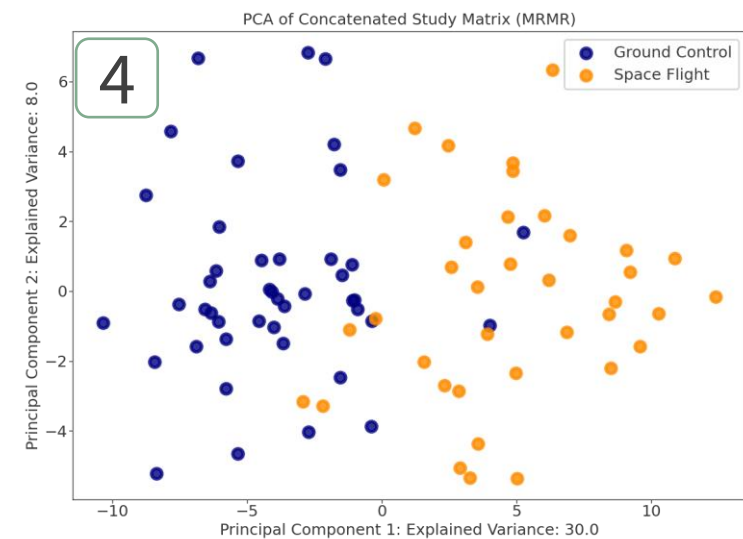
Pre-Filtering & Normalization



Remove technical replications and perform transformations & prefiltering



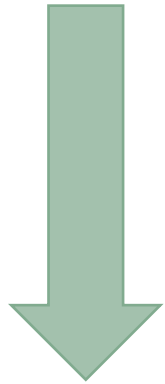
Subset to features selected via Maximum Relevance Minimum Redundance





Training of Classification ML Algorithms

Methods selected based on historical application to NGS data sets



Stratified K-Fold Cross Validation

Balanced training and test set split across treatments

Applications of Support Vector Machine (SVM) Learning in Cancer Genomics

[SHUJUN HUANG](#),^{1,2} [NIANGUANG CAI](#),² [PEDRO PENZUTI PACHECO](#),² [SHAVIRA NARANDES](#),^{2,3} [YANG WANG](#),³ and [WAYNE XU](#)^{1,2,4}

Potentialities of multivariate approaches in genome-based cancer research: identification of candidate genes for new diagnostics by PLS discriminant analysis[†]

G. Musumarra ✉, V. Barresi, D. F. Condorelli, C. G. Fortuna, S. Scirè

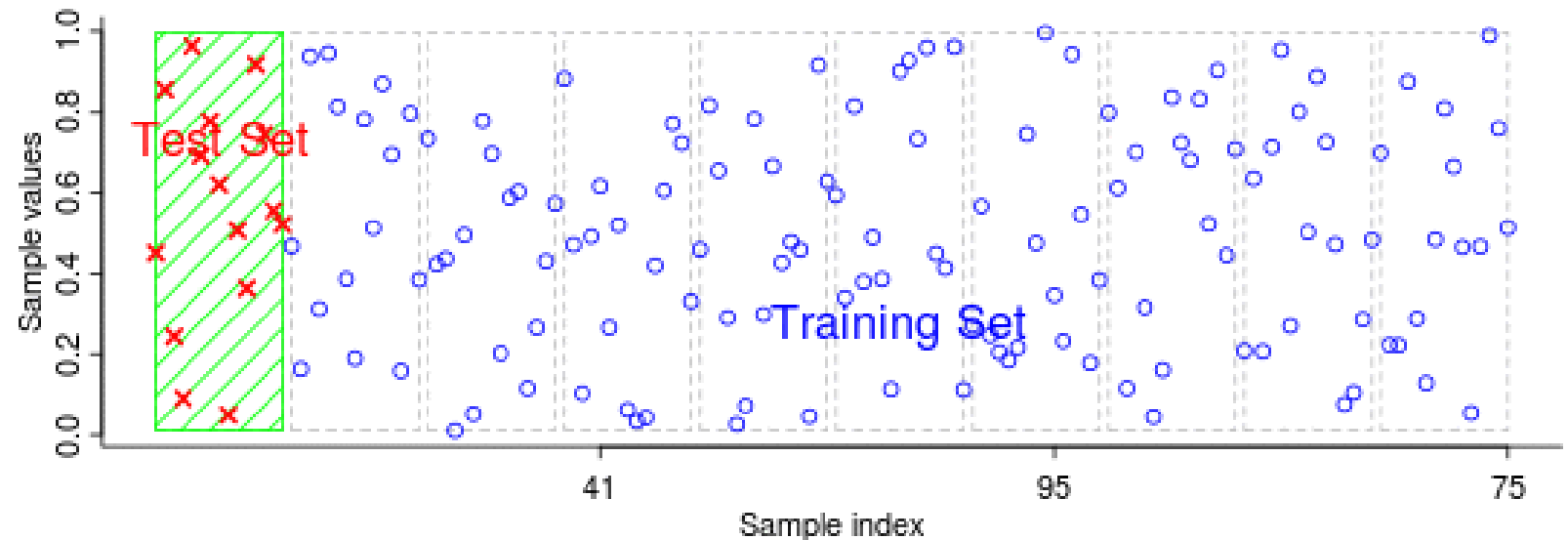
Integration of multi-omics data for prediction of phenotypic traits using random forest

[Animesh Acharjee](#), [Bjorn Kloosterman](#), [Richard G. F. Visser](#) & [Chris Maliepaard](#) ✉

[BMC Bioinformatics](#) **17**, Article number: 180 (2016) | [Cite this article](#)

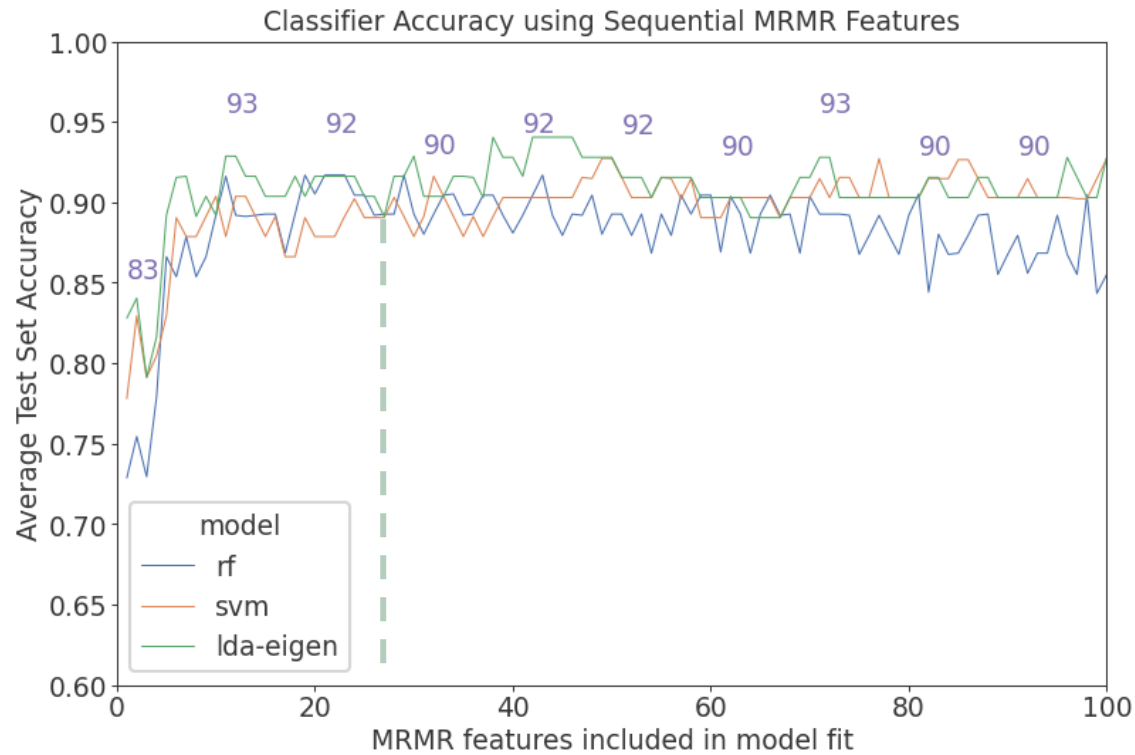


Support Vector Machines (SVM) Random Forest (RF) Linear Discriminant Analysis (LDA)





MRMR & Classification Results

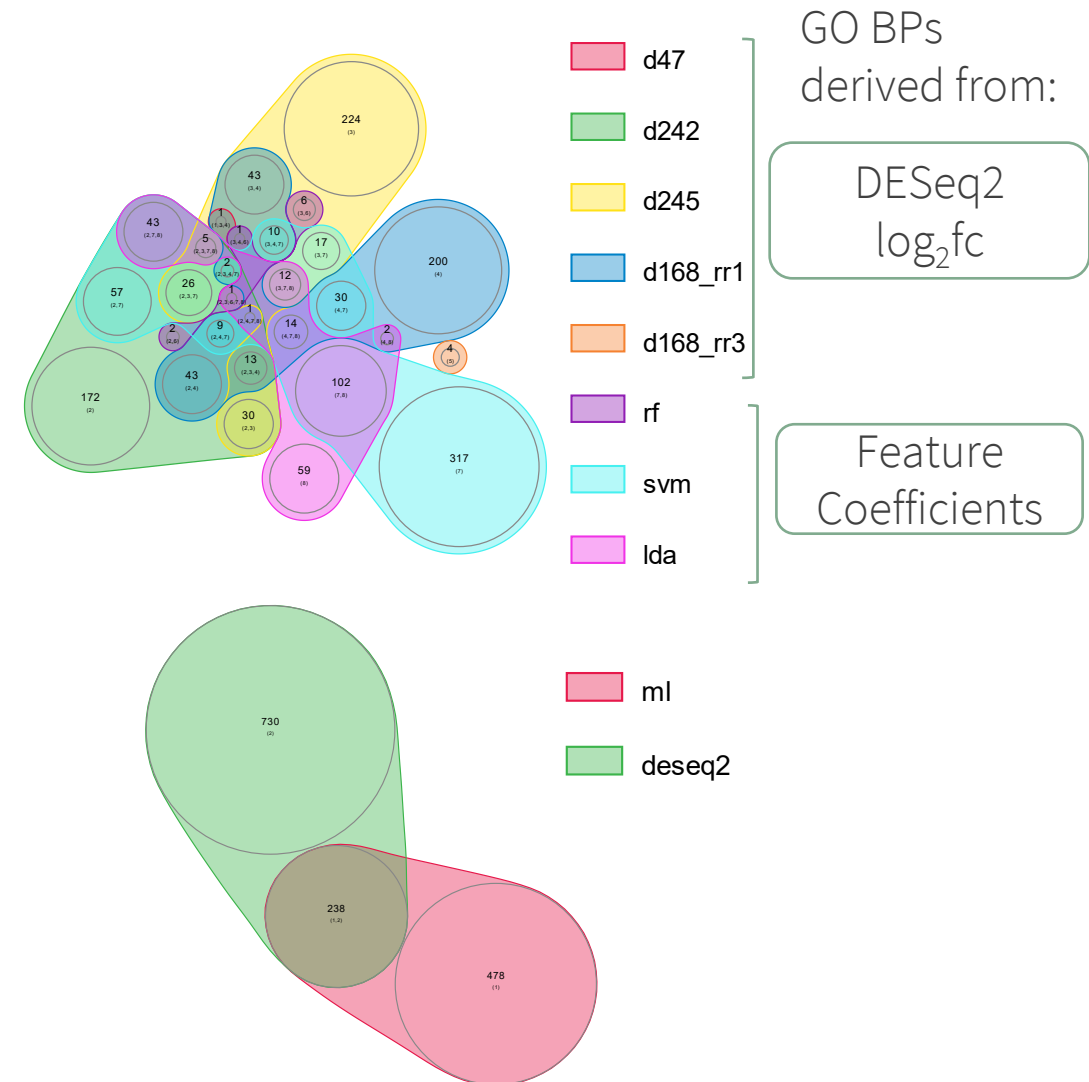


Average prediction accuracy of 90% across all models

Model parameters modified for small sample size (even with concatenated studies)

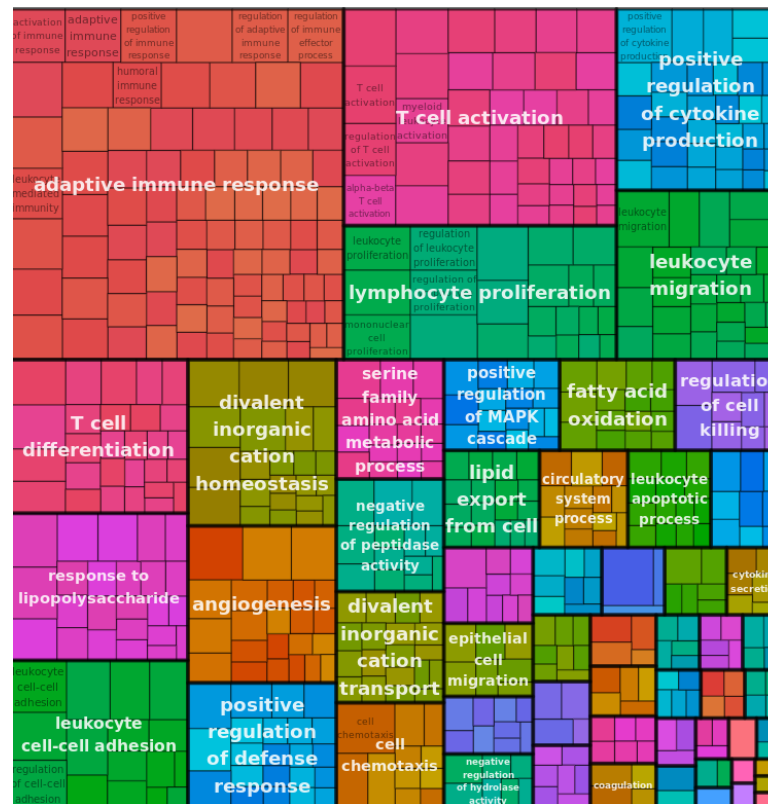
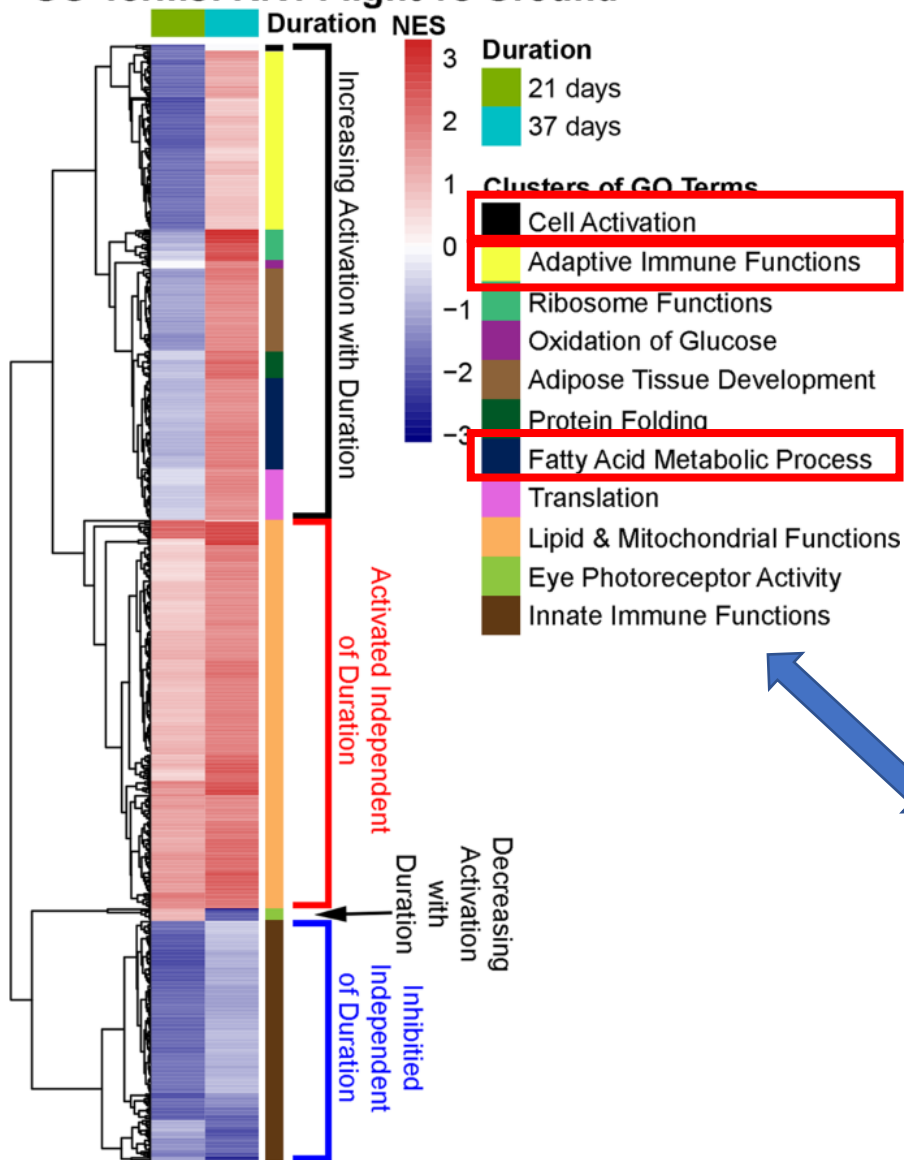
- RF Tree Depth
- SVM Regularization (c)
- LDA Shrinkage

Gene Set Enrichment Analysis (BPs)



Validation of Results

GO Terms: RR1 Flight vs Ground

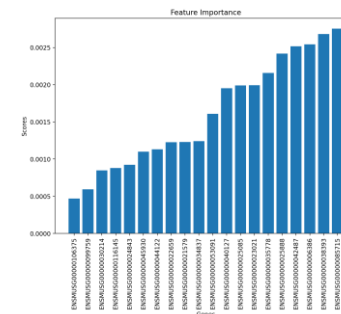
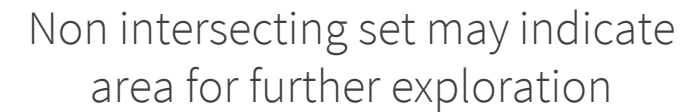
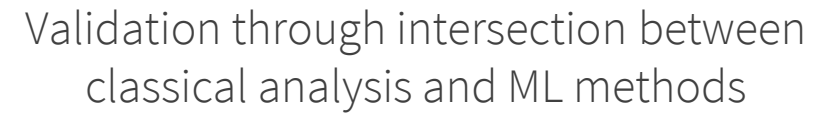


\$'0, 1, 1, 1, 0 (d242, d245, d168_rr1)'

'steroid metabolic process' 'organic hydroxy compound metabolic process' 'alcohol metabolic process' 'mitochondrial protein complex'

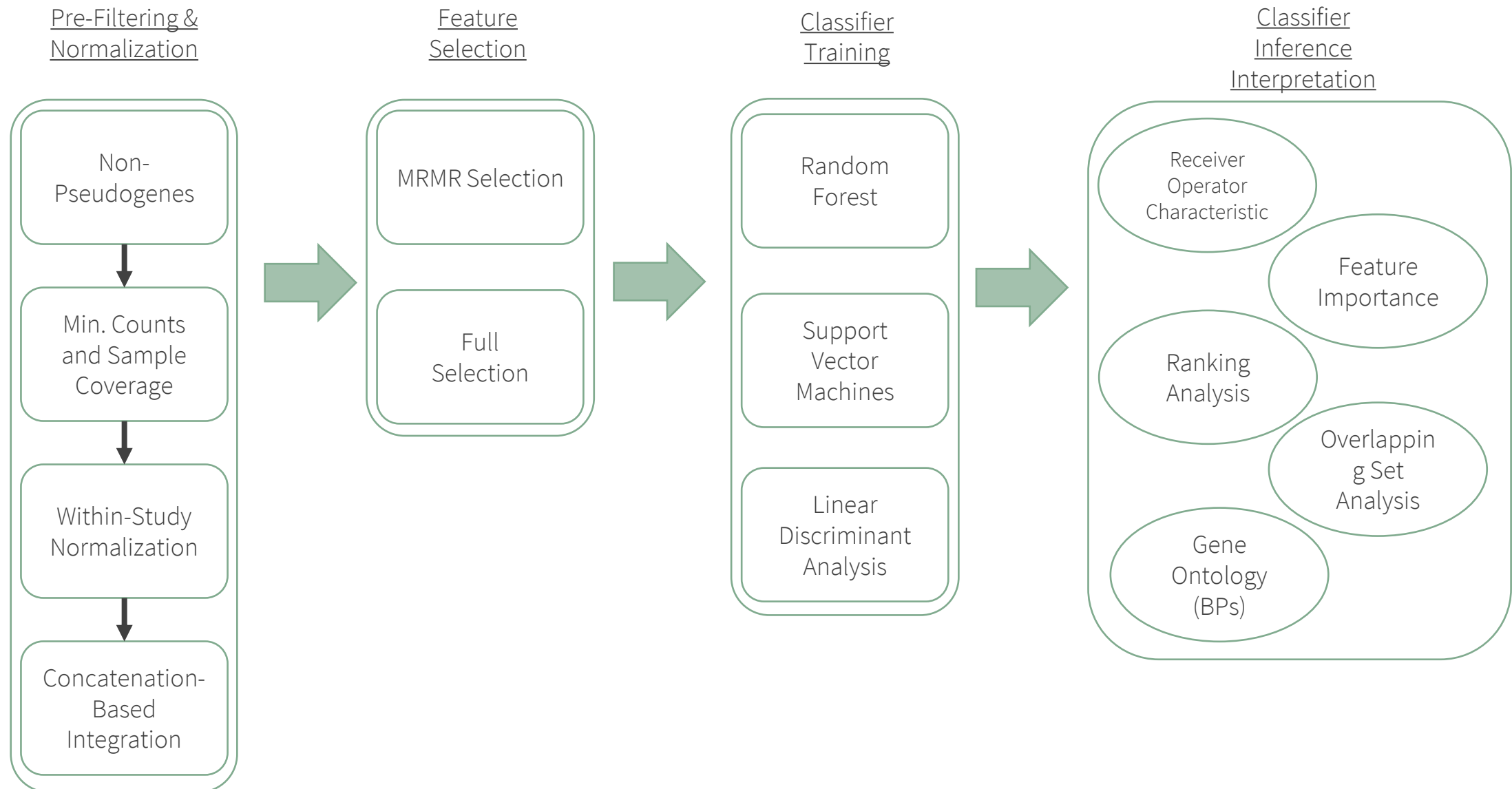
'mitochondrial matrix' 'organelle inner membrane' 'response to virus' 'defense response to virus' 'olefinic compound metabolic process'

'tetrapyrrole binding' 'organellar ribosome' 'mitochondrial ribosome' 'heme binding' 'organellar large ribosomal subunit'





Appendix: Process Flow Diagram





Appendix: SVM GO Biological Processes

