GenomicSuperSignature

- For transfer learning and efficient database search

BioC2022 Conference
Sehyun Oh, PhD

Acknowledgements

Levi Waldron

Sean Davis

CUNY SPH GRADUATE SCHOOL OF PUBLIC HEALTH & HEALTH POLICY

ISPH CUNY Institute for Implementation Science in Population Health

Funding
NCI Informatics Technology for Cancer Research (ITCR) #5U24CA180996

Contents

1. Introduction

Motivation

- Rapidly increasing number of gene expression profiles have been deposited in public archives, yet remain unused for the interpretation of most newly performed experiments.
- There have been many attempts to use the existing datasets, but
  - Hard to use (e.g. require extensive bioinformatics knowledge)
  - Requires heavy computing resources (e.g. need to train the model)
  - Works only on a specific data type (e.g. immune cell only)

➡ Our Method

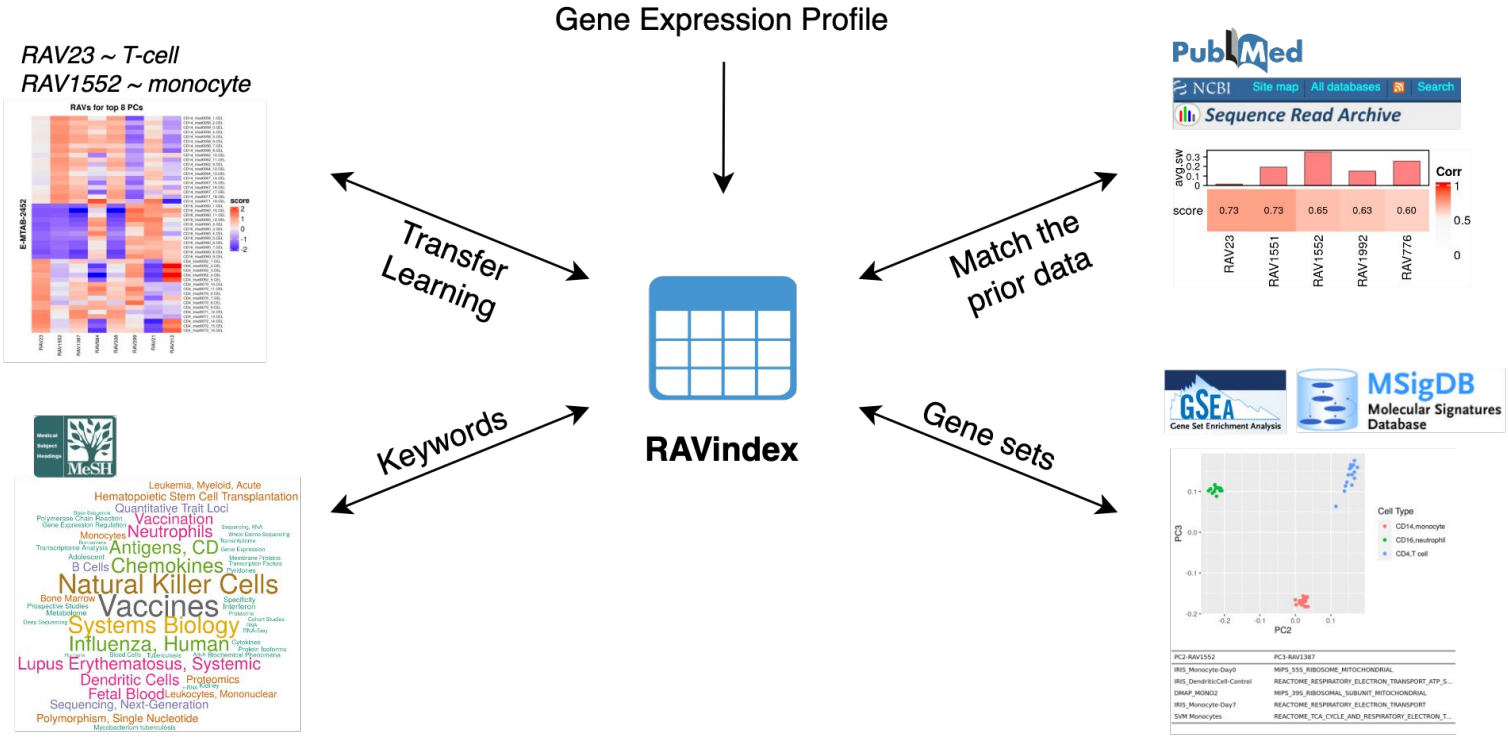
- Two components:
  - ***Pre-trained model*** (named *RAVmodel*) from large, heterogeneous public datasets
  - ***R/Bioconductor package*** (names *GenomicSuperSignature*) for easy application of the model on new data
- Robust to batch effect - applicable across platforms and different underlying biology

Applications

- Interpret gene expression profiles by comparison to published data archived in SRA and by connecting to the relevant literatures, MeSH terms, and gene sets.

- Potential Applications:
  - Find similar studies/datasets to your own gene expression data
  - Find pathways associated with your sample/dataset (*e.g.* 'annotate' PCs)
  - Comparable analysis across datasets from different platforms (*e.g.* microarray vs. RNAseq)
  - Disease subtyping using the continuous scores assigned by our model
  - Identify or inferring weak/missing signal
  - ... more

Core value: reuse and interoperability



nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-022-31411-3 OPEN

GenomicSuperSignature facilitates interpretation of RNA-seq experiments through robust, efficient comparison to public databases

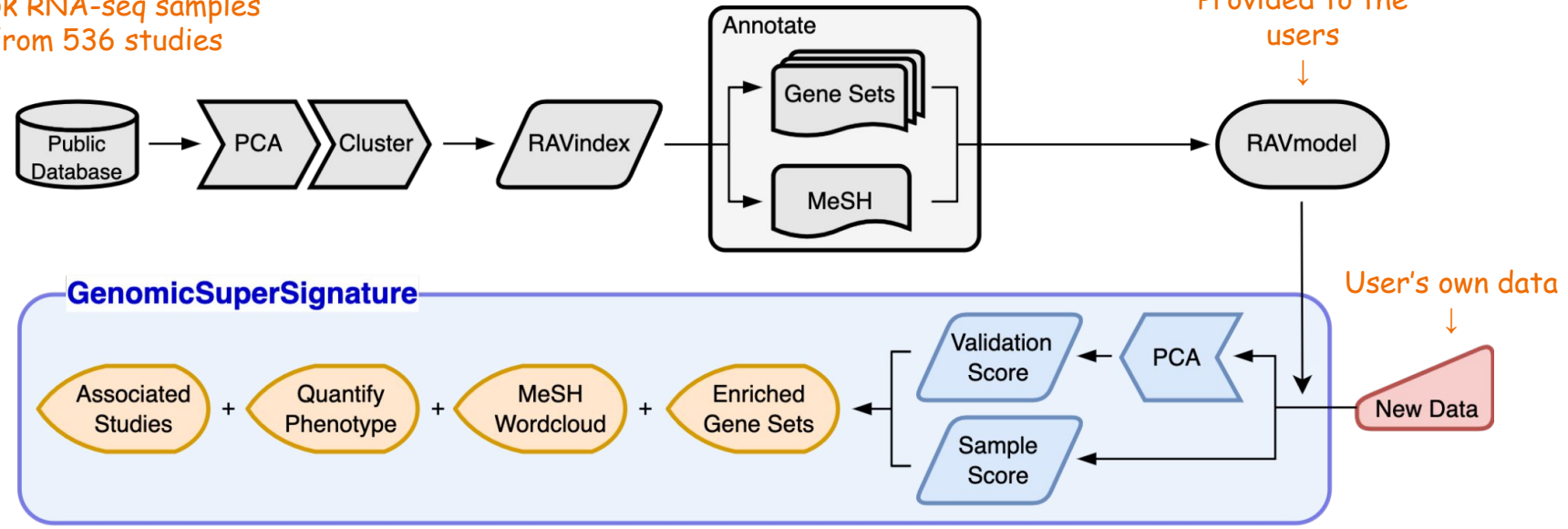
Sehyun Oh[1], Ludwig Geistlinger[2], Marcel Ramos[1], Daniel Blankenberg[3,4], Marius van den Beek[5], Jaclyn N. Taroni[6], Vincent J. Carey[7], Casey S. Greene[8], Levi Waldron[1,9] & Sean Davis[8,9]✉

2. What GenomicSuperSignature is?

RAVmodel building

(Replicable Axis of Variation)

~45k RNA-seq samples from 536 studies



Everything inside blue box done by GenomicSuperSignature

What RAV is?

- ***R***eplicable ***A***xis of ***V***ariation
- RAV construction:
  1. Collect the top principal components (PCs) of the training datasets
  2. Cluster those PCs - hierarchical clustering using Spearman's correlation
  3. Average PCs in each cluster (named *RAV*)

```
> findStudiesInCluster(RAVmodel, 221)
  studyName PC Variance explained (%)
1 ERP016798  2                  8.25
2 SRP023262  9                  1.07
3 SRP111343  3                  4.46
```

- RAV can be compared to PCs of new data (referred as '*validation*' process)

Annotation

MeSH terms

- MeSH (***Me***dical ***S***ubject ***H***eadings) terms are labels assigned to each article in Medline in order to describe what the article is about.
- Process
  1. Collect all the MeSH terms assigned to the studies used to build RAVindex
  2. Each term is adjusted by
     1) Frequency of the term
     2) Variance explained by PC
  3. Filtered with a customizable *'droplist'*

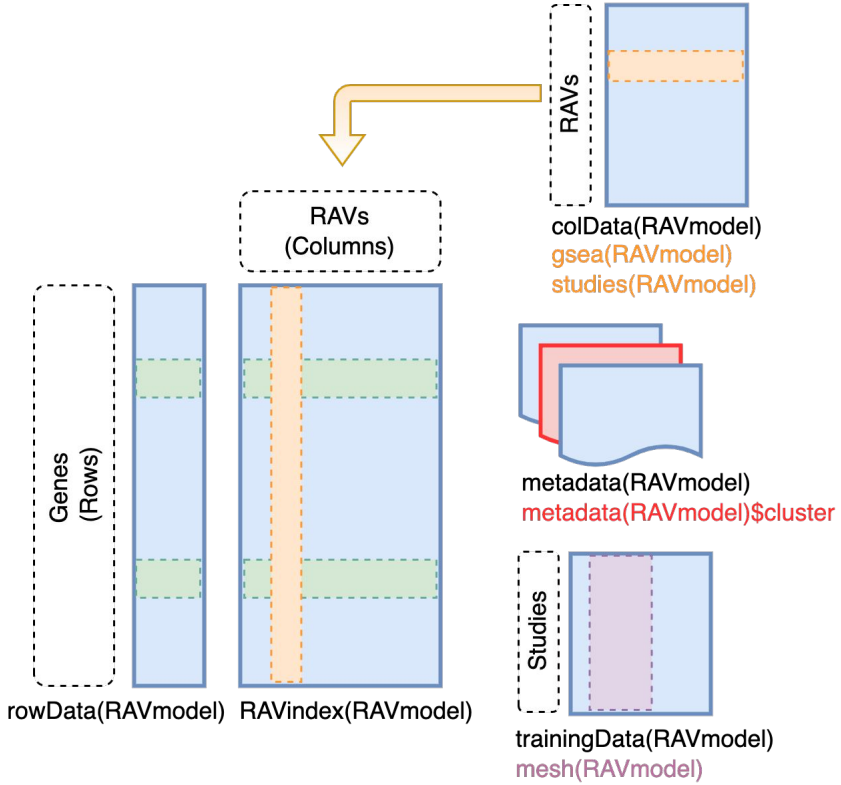
GSEA

1. Create a pre-ranked gene list from each RAV
2. GSEA on pre-ranked gene list
3. Annotate RAVs with the enriched pathways with the minimum *q-value*
4. Association strength of the enriched pathways and the RAV is ranked by normalized enrichment score (NES)

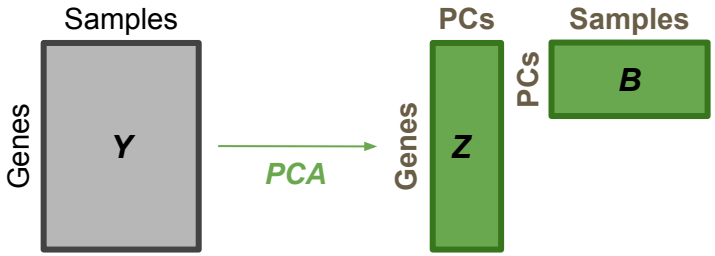
RAVmodel

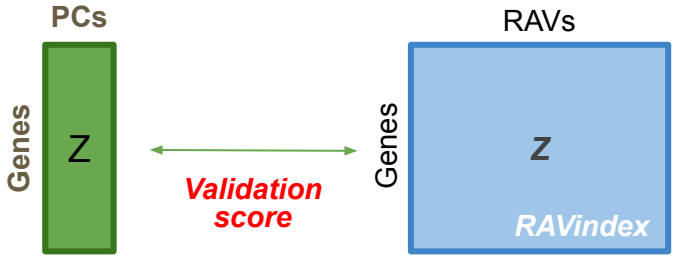
- Inherit SummarizedExperiment object
- ***RAVindex*** in the assay slot is a ‘genes x RAVs’ matrix, connecting new data to the existing database
- Information on the training datasets is stored in *colData* and *trainingData* slots.

```
> RAVmodel
class: PCAGenomicSignatures
dim: 13934 4764
metadata(8): cluster size ... version geneSets
assays(1): RAVindex
rownames(13934): CASKIN1 DDX3Y ... CTC-457E21.9 AC007966.1
rowData names(0):
colnames(4764): RAV1 RAV2 ... RAV4763 RAV4764
colData names(4): RAV studies silhouetteWidth gsea
trainingData(2): PCAsummary MeSH
trainingData names(536): DRP000987 SRP059172 ... SRP164913 SRP188526
```



‘Validation’ Process





Input from user
Procedure inside the package
Data provided by the RAVmodel
Output for user

Summary of the key terms

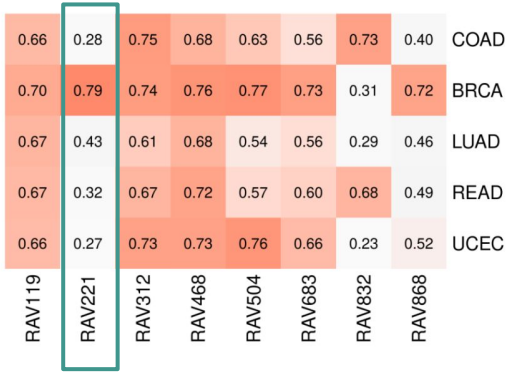
Terms	Description
RAV	A vector containing the average of loadings in each cluster.
RAVindex	A matrix containing all the RAVs. Rows are genes and columns are RAVs.
RAVmodel	Contains RAVindex, metadata on model building, and annotation. Different versions of RAVmodels are available.
Validation Score	The highest Pearson Correlation between top 8 PCs of new data and RAVs. Validation score provides a quantitative representation of the relevance between a new dataset and RAV. Process of comparing top PCs and RAVs is referred to as 'validation' and the RAV that gives the validation score is called 'validated RAV'.
Sample Score	The matrix multiplication result between the ‘genes x samples’ matrix of a new dataset and RAVindex. Similar to validation score, sample score provides a quantitative representation of the relevance between samples and the given RAV.

3. Analysis by GenomicSuperSignature

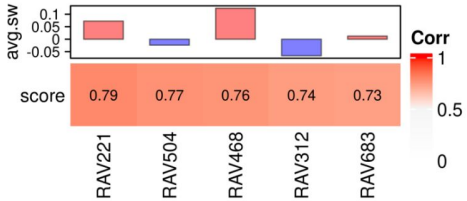
Quick connection to the existing database

5 TCGA datasets

a



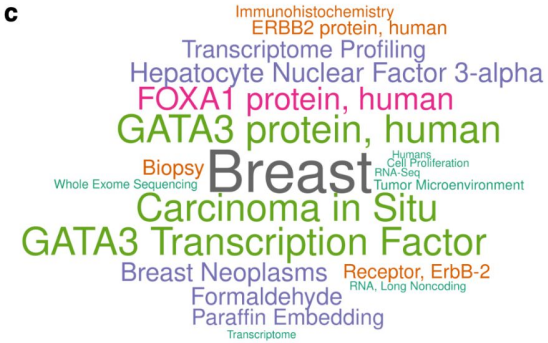
b



TCGA-BRCA

MeSH terms for RAV221

c



d

studyName	PC	Variance explained (%)	title
ERP016798	2	8.25	Whole transcriptome profiling of 63 breast cancer tumours
SRP023262	9	1.07	A shared transcriptional program in early breast neoplasias despite genetic and clinical distinctions
SRP111343	3	4.46	RNAseq analysis of chemotherapy and radiation therapy-naïve breast tumors

Relevant studies to RAV221

e

RAV221.Description	RAV221.NES
SMID_BREAST_CANCER_BASAL_DN	3.423676
SMID_BREAST_CANCER_LUMINAL_B_UP	3.119584
DOANE_BREAST_CANCER_ESR1_UP	3.081407
VANTVEER_BREAST_CANCER_ESR1_UP	3.065605
LIEN_BREAST_CARCINOMA_METAPLASTIC_VS_DUCTAL_DN	2.998661
CHARAFE_BREAST_CANCER_LUMINAL_VS_BASAL_UP	2.945720
CHARAFE_BREAST_CANCER_LUMINAL_VS_MESENCHYMAL_UP	2.926833
SMID_BREAST_CANCER_RELAPSE_IN_BONE_UP	2.890445
SMID_BREAST_CANCER_RELAPSE_IN_BRAIN_DN	2.787022
POOLA_INVASIVE_BREAST_CANCER_DN	2.729454

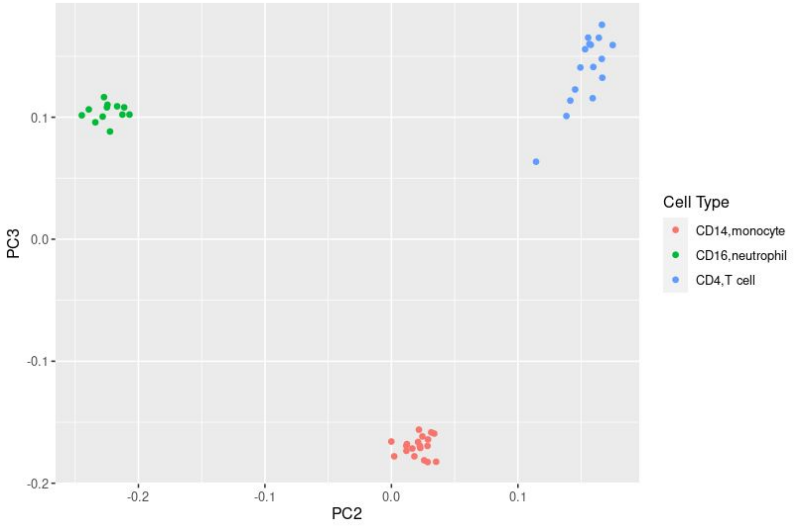
Enriched pathways for RAV221

Interpret PCs of your own data

- E- MTAB-2452 (McKinney *et al.*, 2015), a dataset comprised of isolated immune subsets from patients with autoimmune diseases.

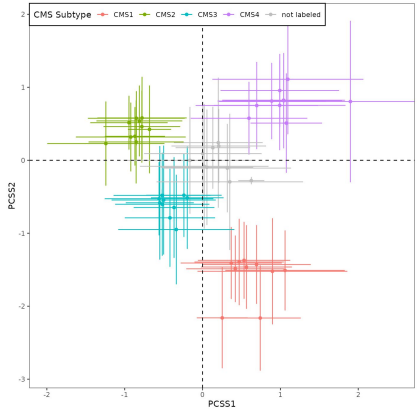
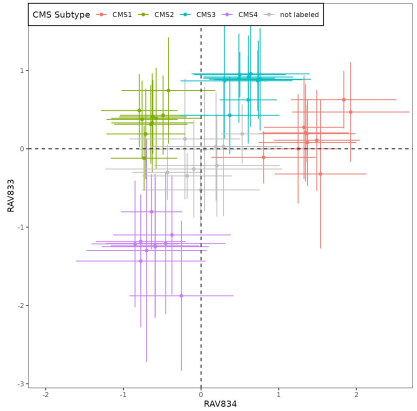
```
> annotatePC(2, val_all, RAVmodel, simplify = FALSE)
$`PC2-RAV1552`
                  Description      NES pvalue      qvalues
1         IRIS_Monocyte-Day0 2.586697  1e-10 2.680702e-09
2 IRIS_DendriticCell-Control 2.433219  1e-10 2.680702e-09
3                DMAP_MONO2 2.376574  1e-10 2.680702e-09
4         IRIS_Monocyte-Day7 2.366122  1e-10 2.680702e-09
5              SVM Monocytes 2.314221  1e-10 2.680702e-09

> annotatePC(1:3, val_all, RAVmodel, scoreCutoff = 0)
RAV1387 can be filtered based on GSEA_PLIERpriors
                       PC1.RAV23                PC2.RAV1552
1                SVM T cells CD8          IRIS_Monocyte-Day0
2          SVM T cells CD4 naive IRIS_DendriticCell-Control
3  SVM T cells follicular helper                 DMAP_MONO2
4 SVM T cells regulatory (Tregs)          IRIS_Monocyte-Day7
5        SVM T cells gamma delta                SVM Monocytes
                               PC3.RAV1387
1                MIPS_55S_RIBOSOME_MITOCHONDRIAL
2 REACTOME_RESPIRATORY_ELECTRON_TRANSPORT_ATP_S...
3        MIPS_39S_RIBOSOMAL_SUBUNIT_MITOCHONDRIAL
4 REACTOME_TCA_CYCLE_AND_RESPIRATORY_ELECTRON_T...
5         REACTOME_RESPIRATORY_ELECTRON_TRANSPORT
```



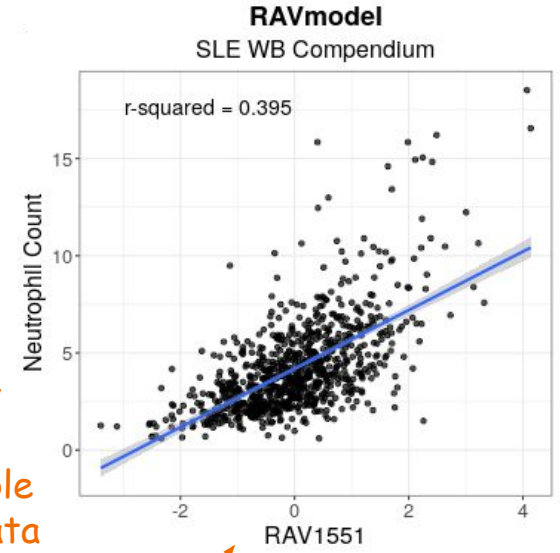
PC2.RAV1552	PC3.RAV1387
IRIS_Monocyte-Day0	MIPS_55S_RIBOSOME_MITOCHONDRIAL
IRIS_DendriticCell-Control	REACTOME_RESPIRATORY_ELECTRON_TRANSPORT_ATP_S...
DMAP_MONO2	MIPS_39S_RIBOSOMAL_SUBUNIT_MITOCHONDRIAL
IRIS_Monocyte-Day7	REACTOME_TCA_CYCLE_AND_RESPIRATORY_ELECTRON_T...
SVM Monocytes	REACTOME_RESPIRATORY_ELECTRON_TRANSPORT

Benchmark #1. Disease subtypes

	Disease-specific model	GenomicSuperSignature
Training datasets	- 8 colon cancer datasets - Microarray datasets	- 536 heterogeneous datasets - RNA sequencing datasets
Test datasets	10 colon cancer datasets (9 microarray + 1 RNA sequencing)	
Colors	4 discrete colon cancer subtypes + 1 undefined group	
		

Benchmark #2. Transfer learning

Dataset #1



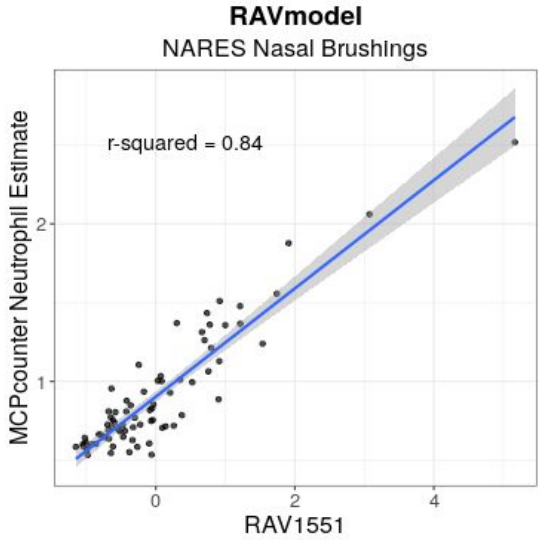
1. Available metadata

2. Identify Signature

Explain Neutrophil count by RAV1551

Verified by Estimates

Dataset #2



3. Apply Signature

Conclusions

GenomicSuperSignature demonstrates

1. Efficient and coherent database search
2. Robustness to batch effects
3. Transfer learning capacity
4. Improvements from the existing approaches :
    - Usability → Pre-computed model + R/Bioconductor package
    - Versatility → Not limited to any specific biology and robust to platforms
    - Modularity → Annotation is separated from model building
    - Scalability → Current model building takes less than a few days

A little addition...

- Future direction
  - Expand RAVmodel collections : single-cell data, mice, microbiome, etc.
  - Additional annotation : different gene sets, metadata of originating studies
- More information:
  - Paper : https://www.nature.com/articles/s41467-022-31411-3
  - Package site : https://shbrief.github.io/GenomicSuperSignature/
  - Use cases : https://shbrief.github.io/GenomicSuperSignaturePaper/

4. How to use it?

Prepare your input data

- Gene expression profile - both microarray and RNA sequencing data
- ‘Genes x Samples’ matrix - *ExpressionSet, SummarizedExperiment, Matrix*
- Follow a normal distribution (*e.g.* log2-transformed)
- Genes in gene symbol
- For dataset-level validation, you need at least 8 samples

Live Demo vignette → https://bit.ly/bioc2022_gss