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

Sehyun Oh[1], Ludwig Geistlinger[2], Marcel Ramos[1], Jaclyn N. Taroni[3], Vincent Carey[4], Casey Greene[5], Levi Waldron[1,*], Sean Davis[5,*,+]

[1]City University of New York, New York, NY, USA
[2]Center for Computational Biomedicine, Harvard Medical School, Boston, MA, USA
[3]Childhood Cancer Data Lab, Alex's Lemonade Stand Foundation, Philadelphia, PA, USA
[4]Harvard Medical School, Boston, MA, USA
[5]University of Colorado Anschutz School of Medicine, Denver, CO, USA

* equal contribution, [+]corresponding

Abstract

Millions of transcriptomic profiles have been deposited in public archives, yet remain underused for the interpretation of new experiments. We present a novel method for interpreting new transcriptomic datasets through near-instantaneous comparison to public archives without high-performance computing requirements. We performed Principal Component Analysis on 536 studies comprising 44,890 RNA sequencing profiles. Sufficiently similar loading vectors were aggregated to form *R*eplicable *A*xes of *V*ariation (RAV). RAVs were annotated with metadata of originating studies and by gene set enrichment analysis, forming a knowledge graph. Functionality to associate new datasets with RAVs, extract interpretable annotations, and provide intuitive visualization are implemented as the GenomicSuperSignature R/Bioconductor package. We demonstrated the efficient and coherent database searching, robustness to batch effects and heterogeneous training data, and transfer learning capacity of our method using TCGA and rare diseases datasets. GenomicSuperSignature will aid analyzing new gene expression data in the context of existing databases using minimal computing resources.

Abstract

PURPOSE: Millions of transcriptomic profiles have been deposited in public archives, yet remain underused for the interpretation of new experiments. Existing methods for leveraging these public resources have focused on the reanalysis of existing data or analysis of new datasets independently. We present a novel approach to interpreting new transcriptomic datasets by near-instantaneous comparison to public archives without high-performance computing requirements.

METHODS: To identify replicable and interpretable axes of variation in any given gene expression dataset, we performed Principal Component Analysis (PCA) on 536 studies comprising 44,890 RNA sequencing profiles. Sufficiently similar loading vectors, when compared across studies, were aggregated to form *R*eplicable *A*xes of *V*ariation (RAV). RAVs were annotated with metadata of originating studies and by gene set enrichment analysis, forming a knowledge graph. Functionality to associate new datasets with RAVs, extract interpretable annotations, and provide intuitive visualization are implemented as the GenomicSuperSignature R/Bioconductor package.

RESULTS: RAVs are robust to batch effects and the presence of low-quality or irrelevant studies, and identify signals that can be lost by merging samples across the training datasets. The GenomicSuperSignature package allows instantaneous matching of PCA axes in new datasets to pre-computed RAVs, cutting down the analysis time from days to the order of seconds on an ordinary laptop. We demonstrate that RAVs associated with a phenotype can provide insight into weak or indirectly measured biological attributes in a new study by leveraging accumulated data from published datasets. Benchmarking against complementary previous works demonstrates that the RAV index 1) identifies colorectal carcinoma transcriptome subtypes that are similar to but more correlated with clinicopathological characteristics than previous disease-specific efforts and 2) can estimate neutrophil counts through transfer learning on new data comparably to the previous efforts despite major differences in training datasets and model building processes with the additional benefits of flexibility and scalability of the model application.

CONCLUSION: GenomicSuperSignature establishes a knowledge graph where different prior knowledge databases are coherently linked, and enables researchers to analyze new gene expression data in the context of existing databases using minimal computing resources. The robustness of GenomicSuperSignature suggests that we can expand this approach beyond human gene expression profiles, such as single-cell RNA-seq, microbiome abundance, and different species' transcriptomics datasets.

Introduction

Vast quantities of transcriptomic data have been deposited in public archives, yet remain underused for the interpretation of new experiments. Here, we present a novel toolkit and approach for interpreting new transcriptomic datasets through near-instantaneous comparison to public transcriptomic datasets using the compute resources available on a standard laptop.

Dimensionality reduction has been broadly adopted to transform transcriptomes onto a smaller number of latent variables representing co-expressed transcripts. Gene co-expression can result from shared function or regulation[1], association with tissue composition or cell type[2], and technical batch effects[3]. In the confluence of these factors, dimensionality reduction can assist interpretability and reduce the burden of multiple hypothesis testing, but can also lead to incomplete or misleading interpretation. Valid interpretation would be improved by comparison of latent variables in new datasets to those also present in public transcriptome databases.

Many dimensionality reduction approaches, differing in the optimization and constraint criteria, are available[4] and there have been multiple attempts to detect biological and technical signals through these lower dimensional, latent variable representations. Classic methods such as Principal Component Analysis (PCA) and non-negative matrix factorization (NMF) remain widely used in their original form and as bases for newer methods. For example, scCoGAPS is an NMF method optimized for large, sparse single cell RNA sequencing datasets[5]. scCoGAPS recovers features in a source dataset, then projects a new dataset onto this learned latent space through projectR[5,6]. This approach requires users to train their own model and mostly focuses on single cell RNA sequencing datasets with similar biology. PLIER aims to improve on the interpretability of PCA by identifying latent variables that map to a single gene set or a group of highly related gene sets with positive correlations[7]. MultiPLIER applies the PLIER approach to transfer learned patterns from a large public dataset to rare diseases[8]. Other tools focus on recovering consistent signals from multiple datasets across distinct platforms[9,10], increasing interpretability[11], or transfer learning between datasets of a specific type[12]. However, none of these tools enable routine exploratory analysis of new studies through comparison to large public transcriptome databases. Also, these tools either do not provide transfer learning from large public databases, or in the case of MultiPLIER, require substantial computing resources and bioinformatics expertise.

Here, we describe GenomicSuperSignature, a toolkit for interpreting new RNA-seq datasets by dimensionality reduction in the context of a large-scale database of previously published and annotated results. As an exploratory data analysis tool, GenomicSuperSignature matches PCA axes in a new dataset to an annotated index of ***r***eplicable ***a***xes of ***v***ariation (RAV) that are represented in previously published independent datasets. GenomicSuperSignature also can be used as a tool for *transfer learning*[13], utilizing RAVs as well-defined and replicable latent variables defined by multiple previous studies in place of *de novo* latent variables. The

interpretability of RAVs is enhanced through annotations by *ME*dical *S*ubject *H*eadings (MeSH) and *G*ene *S*et *E*nrichment *A*nalysis (GSEA). Through the use of pre-built, pre-annotated, dimension-reduced RAVs, GenomicSuperSignature leverages knowledge from tens of thousands of samples and from PubMed and MSigDB[14], to the dataset at hand within seconds on an ordinary laptop. We demonstrate these functionalities in colorectal carcinoma, breast invasive carcinoma, systemic lupus erythematosus, and rare inflammatory disease. GenomicSuperSignature is implemented as an R/Bioconductor package for straightforward incorporation into popular RNA-seq analysis pipelines.

Results

The current RAVmodel is trained on 536 studies containing 44,890 human RNA sequencing profiles. This RAVmodel is associated with 18,798 (4,430 unique) MeSH terms and 70,687 (1,784 unique) MSigDB curated (C2) gene sets. This assembly of multiple existing databases represents a 'knowledge graph' (Fig.1B), accompanied by tools in the GenomicSuperSignature R package for applying this knowledge graph to the interpretation of new datasets (Fig. 1A and Supplementary Fig. 1C). We demonstrate this novel application of public data in three examples. First, using TCGA datasets, we showed that new data can be rapidly associated with related studies, gene sets, and MeSH terms. Second, we showed that the knowledge graph trained from diverse RNA-seq experiments identified colon cancer transcriptome subtypes more closely associated to clinicopathological variables than the subtypes previously identified by meta-analysis of a focused colorectal carcinoma (CRC) microarray compendium. Lastly, we showed that neutrophil counts of two independent datasets can be interpreted and inferred through a single RAV, providing a novel quantitative measure of neutrophil count from transcriptome data. These examples, along with sensitivity analyses and simulations, demonstrate that the RAVmodel and GenomicSuperSignature are a robust, general-purpose method for the interpretation of transcriptome data.

Sensitivity Analysis and Simulation

Model training methods were optimized for robustness, simplicity, computational cost, and validity (Supplementary Methods). Briefly, the RAVmodel was trained on the RNA-seq Sample Compendia of refine.bio[15]. We analyzed TPM count data using Principal Components Analysis (PCA) following log-transformation, then identified clusters of similar principal components from independent datasets using hierarchical clustering on Spearman distance and ward.D agglomeration. This approach was compared to alternatives based on 1) ability to group synthetic true-positive principal components, 2) separation of synthetic true-negative principal components added to training data, 3) magnitude of changes in the results compared to the simplest method, and 4) maintenance of RAVs identified from a focused training dataset when adding unrelated datasets. Alternative approaches considered but not selected for training the final model included NMF, Independent Component Analysis (ICA), PLIER[16] and MultiPLIER[8], Variance-Stabilizing Transformation (VST)[17] in place of log transformation, combining training datasets into a single dataset instead of analyzing them independently, increasing the number of

principal components included per dataset, and alternative clustering algorithms including graph-based clustering. These assessments are described in Supplementary Methods "Sensitivity analysis for model building".

Connecting new data with the existing databases

To demonstrate the ability to match datasets under new analysis to relevant published datasets, we applied RAVmodel on five TCGA datasets (Fig. 2A). Based on the correlation to principal components in these datasets, we identified RAVs specific to breast invasive carcinoma (RAV221 and RAV868) and to colon and rectal adenocarcinoma (RAV832). When RAVmodel was applied to the TCGA-BRCA dataset, RAV221 was assigned with the highest validation score (Fig. 2B, Supplementary Table 1) and the associated MeSH terms were mostly breast-related terms, such as 'breast' and 'breast neoplasms' (Fig. 2C, `drawWordcloud` function). We extracted three breast cancer studies contributing to RAV221 (Fig. 2D, `findStudiesInCluster` function). GSEA annotations on RAVs were queried and the top 10 enriched pathways were all breast-cancer associated (Fig. 2E, `subsetEnrichedPathways` function). We also checked RAV832 on its association with TCGA-COAD and TCGA-READ datasets. RAV832 was assigned with the second highest validation score for both COAD and READ datasets (Supplementary Fig. 4A and 4B, respectively) and contained MeSH terms such as 'colorectal cancers', 'colon', and 'adenocarcinoma' (Supplementary Fig. 4C). We also recognized that three out of five training data in RAV832 directly represented colon-associated illnesses (Supplementary Fig. 4D) and the top enriched gene set was an upregulated pathway in colorectal adenoma (Supplementary Fig. 4E). In summary, we confirmed that RAV serves as a specific and robust index, coherently connecting expression profile, gene sets, related studies, and MeSH terms to build knowledge graph (Fig. 1B), and ultimately enables the interpretation of new datasets in the context of existing databases.

RAVs to characterize colorectal cancer

To compare the utility of GenomicSuperSignature relative to the focused use of data from a single disease, we compared RAVs to two previous studies that employed CRC gene expression databases to identify CRC molecular subtypes. The CRC Subtyping Consortium used 18 CRC datasets from multiple platforms comprising 4,151 patients to define four discrete Consensus Molecular Subtypes (CMS) observed across numerous patient cohorts[18,19]. Ma *et al.* subsequently proposed a continuous scoring system (PCSS) based on an analysis of 8 CRC microarray datasets comprising 1,867 samples, and found it was more closely correlated to microsatellite instability (MSI)[20,21], grade, stage, and tumor location[18,19]. Importantly, these previous efforts both employed curated databases of only CRC transcriptomes, whereas the training set of the current RAVmodel consists of less than 2% CRC studies (Supplementary Table 3). We identified the RAVs most highly associated with CMS subtypes (RAV834/833) and PCSSs (RAV1575/834) (Supplementary Results), and confirmed that these RAV pairs showed comparable or higher performance on colon cancer subtyping than CRC subtyping efforts defined by bespoke methods in focused datasets (Fig. 3A, Supplementary Fig. 5A).

Using training and validation data of the original CRC studies, we compared associations between different subtype models and RAVs with the same clinicopathological variables. Notably, these data were *not* part of RAV training and are microarray datasets whereas the RAVs were trained exclusively from RNA-seq data. As described previously[18], we used the likelihood-ratio test (LRT) to compare the different subtype models for association with clinicopathological variables. A p-value near 1 ($-\log_{10}$p-value near 0) means that no additional information is provided by a full model composed of two subtype definitions compared to a model with only one. CMS-associated RAVs performed better than discrete CMS on all four phenotypes and also outperformed PCSSs except on tumor location (Fig. 3B-C). Interestingly, PCSS-associated RAVs were still better than CMS but slightly worse than PCSSs, while CMS-associated RAVs were better than both CMS and PCSSs, indicating that RAVs contain more comprehensive information (Supplementary Fig. 5B-C). This performance improvement became more significant using only the 10 original validation datasets, excluding 8 datasets used to train the PCSS model (Supplementary Fig. 6). In conclusion, RAVs trained from heterogeneous datasets, not specific to CRC, captured biologically-relevant signatures for CRC as well or superior to focused efforts using CRC-specific databases, suggesting that RAVs are for general-use and can be applied to describe other diseases as well.

Identify common biological attributes across different datasets

For practical and technical reasons, biological datasets often contain missing information or signals buried in noise. GenomicSuperSignature can fill out those gaps by uncovering weak or indirectly measured biological attributes of a new dataset by leveraging the existing databases. To evaluate this 'transfer learning' aspect of the GenomicSuperSignature, we compared the neutrophil count estimation by RAVs across two different datasets, systemic lupus erythematosus whole blood (SLE-WB)[22] and nasal brushing (NARES)[23] datasets as described in the previous study[8]. We searched for the SLE pathology-relevant RAV in three different ways using the SLE dataset[22]. First, we identified RAV1551 based on the highest validation score with the positive average silhouette width. Second, we searched for the keyword, *neutrophil*, in GSEA-based annotation. Eleven RAVs, including RAV1551, had two keyword-containing, enriched pathways. Lastly, we used the neutrophil count of the SLE-WB dataset to find the metadata-associated RAV. For the continuous variables like neutrophil count, we compared the R^2 between the target variable and all RAVs, where RAV1551 showed the maximum R^2, 0.395 (Fig. 4A). Neutrophil is a terminally differentiated cell type and potentially under-detected in the active gene expression profile, so we used the neutrophil estimate from MCPcounter[24] and further evaluated the correlation between RAV1551 score and neutrophil estimate as the previous study[8]. A stronger correlation between RAV1551 score and neutrophil estimate was observed (Fig. 4B). We concluded that RAV1551 is the SLE pathology-relevant RAV, specifically associated with neutrophil counts, and tested whether this information can be expanded beyond the SLE dataset. For that, we applied RAV1551 on NARES dataset, which is a gene expression profile of nasal brushings obtained from granulomatosis with polyangiitis (GPA) patients, a condition that causes inflammation of blood vessels affecting ears, noses, throats, lungs and kidneys[23]. RAV1551 was not a top validated signal, ranked 14th with the validation score 0.41

with PC1 (data not shown), implying that neutrophil phenotype is not a major feature of NARES dataset. However, R^2 between the neutrophil estimate from MCPcounter and RAV1551 score was 0.84 (Fig. 4C). This suggests that RAV can serve as a new measure to compare different datasets and provide interpretation of potentially subtle biological signals.

Pathway separation

We benchmarked one of the model validation measures, named pathway separation, from the previous study[7,8]. Briefly, pathway separation is defined as the ability of the signature model to keep non-overlapping signatures that can differentiate biologically similar pathways. Three biological subjects were tested - type I versus type II interferon, neutrophil versus monocyte, and G1 versus G2 cell cycle phases. GenomicSuperSignature can successfully separate them either with top one or top five enriched pathways (data not shown).

Discussion

A key innovation of GenomicSuperSignature is the creation of an index of ***r***eplicable ***a***xes of ***v***ariation (RAVindex) consisting of principal components repeatedly observed in independent analysis of multiple published datasets (Fig. 1A). Compared to approaches that merge training data, this strategy is highly scalable, can identify latent variables specific to small training datasets, and ignores technical artifacts that are not observed across multiple datasets. RAVindex is annotated with publication citations, *ME*dical *S*ubject *H*eadings (MeSH) terms, and gene sets, all of which is stored as the 'RAVmodel'. Assembly of this information through RAVmodel creates a knowledge graph which can be rapidly applied to new datasets on a standard laptop (Fig. 1B). GenomicSuperSignature augments standard RNA-seq exploratory data analysis by providing modes of interpretation and hypothesis testing that were previously impractical to apply.

GenomicSuperSignature contains information learned from a large body of existing studies that can be "transferred" to newly collected data. For example, the RAVindex contains cancer type-specific RAVs (Fig. 2A), including RAVs that are more closely related to colorectal carcinoma (CRC) clinicopathological variables than transcriptome subtypes previously identified through intensive analysis of CRC-specific databases bespoke subtyping efforts (Fig. 3, Supplementary Fig. 5). Such "transfer learning" is broadly applicable but particularly beneficial to the study of rare diseases and to small datasets where weak and under-represented, but biologically meaningful, signals can be identified[8]. To demonstrate this, we identified an RAV that was highly correlated to neutrophil content using a systemic lupus erythematosus (SLE) dataset not in model training data, and used this RAV to estimate neutrophil content in a nasal brushing (NARES) dataset that lacks neutrophil count information (Fig. 4). In addition to data inference, GenomicSuperSignature can be useful for analyzing disease progress, comparing phenotypes across independent datasets, and identifying weak biological signals. The current RAVindex contains 4,764 RAVs and 3,386 out of them are previously observed in two or more

independent datasets, which can be expected to have many other such applications through transfer learning.

GenomicSuperSignature is expected to be robust to batch effects because it characterizes clusters of highly similar latent variables from two or more independent studies through cluster sizes, enabling to ignore any that are unique to a single study. We demonstrated the robustness of GenomicSuperSignature through sensitivity analysis and benchmarking against prior disease-specific analyses[8,18]. While trained exclusively on RNA-seq datasets, the performance of GenomicSuperSignature was not diminished when applied to microarray datasets. Furthermore, we observed transfer learning functionality compatible with the results from recount2-MultiPLIER[8], even though that model used a different matrix decomposition method, was trained at the sample level instead of dataset level, and used a training set with only ~10% overlap of samples (Supplementary Methods). We conclude that GenomicSuperSignature is robust to different technological platforms and to heterogeneity of training datasets, and enables interpretation of divergent datasets without subject-specific models.

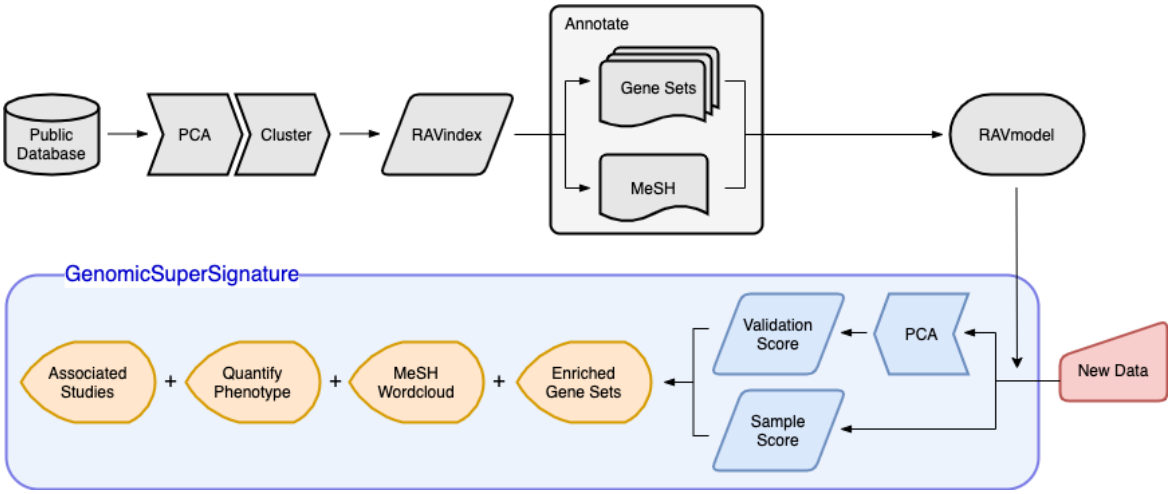
GenomicSuperSignature offers significantly improved usability over the existing tools by adopting user-friendly application schemas. First, the pre-built models greatly reduce computational requirements for users: whereas training the current model took several days on 24 cores with 128 Gb memory, its application can be performed in seconds on a conventional laptop computer. Its implementation as an R/Bioconductor package allows ready incorporation into widely-used RNA-seq analysis pipelines. GenomicSuperSignature R/Bioconductor package also enables a large research community to reuse public data for more accurate analyses of new data.

The approach taken for GenomicSuperSignature is flexible and can be extended to other large publicly available databases. We plan to develop RAVmodels trained on microarray, single-cell RNA sequencing, and spatial transcriptomic data, from model organisms such as mouse, and to metagenomic data from microbiome studies. Cross-species RAVmodels can help extend the discoveries from model organisms to humans[25]. These planned efforts will generate an expanded knowledge graph that incorporates the new ‘knowledge nodes’. For example, the GSEA annotation part of the model is independent from the RAVindex building process, so we can easily build multiple versions of RAVmodel with different gene sets or even any combination of gene sets. Also, we can connect RAVs with additional information on the training data because RAVs keep information on their specific source data. While the collection of RAVmodels grows as described, the GenomicSuperSignature package will be maintained as a stand-alone toolbox equally applicable to these different RAVmodels. GenomicSuperSignature and its associated data resources will provide biomedical researchers with a new set of data exploration tools exploiting knowledge gained from hundreds and eventually thousands of existing public datasets.

Figures

Fig 1. Overview of GenomicSuperSignature

A) Schematic illustration of RAVmodel construction and GenomicSuperSignature application. Building the RAVmodel (grey) was performed over several days on 24 cores with 128 Gb memory. Users can apply RAVmodel on their data (red) using the GenomicSuperSignature R/Bioconductor package (blue), which operates on a time scale of seconds for exploratory data analysis (orange) on a typical laptop computer.



B) Knowledge graph assembled by GenomicSuperSignature. GenomicSuperSignature connects different public databases and prior information through RAVmodel (Supplementary Fig. 1B), creating the knowledge graph illustrated here. Through GenomicSuperSignature and the accompanying RAVmodel, users can instantly access and explore the knowledge graph from multiple entry points such as gene expression profiles, publications, study metadata, keywords in MeSH terms and gene sets.

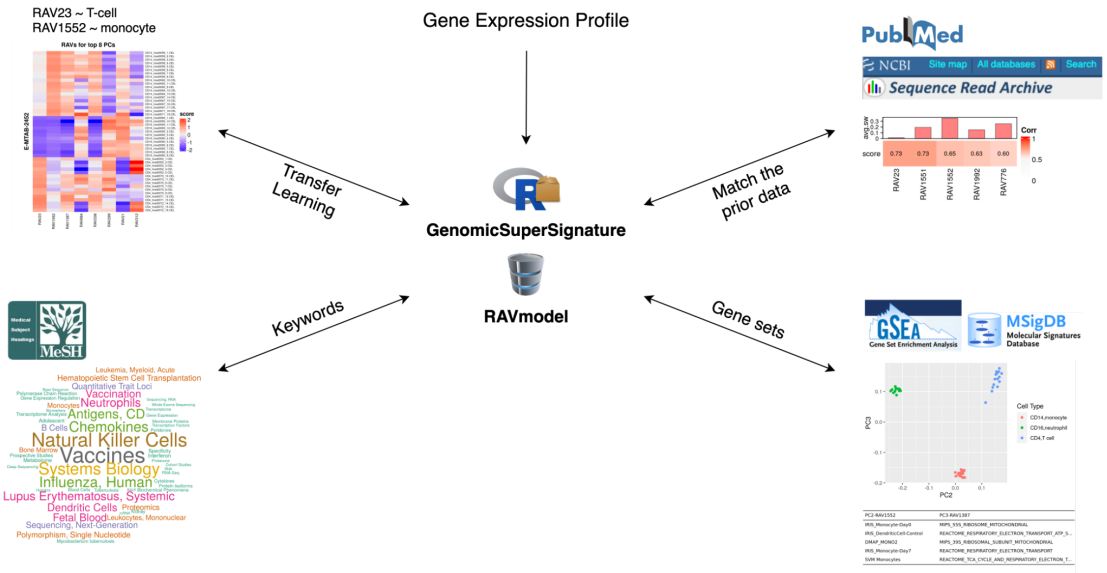


Fig 2. Connecting new datasets to existing databases

GenomicSuperSignature provides a rich resource for understanding new or user-supplied datasets in the context of existing datasets summarized in the RAVmodel. A) Validation of multiple TCGA RNA-seq datasets. Each dataset was subjected to PCA and Pearson correlation coefficients between top PCs and all possible RAVs were calculated. RAVs with Pearson coefficient above 0.7 in at least one dataset were displayed here. RAV221 and RAV868 indicate association with breast cancer while RAV832 is associated with colon and rectal cancer. (COAD: Colon Adenocarcinoma, BRCA: Breast Invasive Carcinoma, LUAD: Lung Adenocarcinoma, READ: Rectum Adenocarcinoma, UCEC: Uterine Corpus Endometrial Carcinoma) B) Validation of TCGA-BRCA. From panel A, we showed RAV221 is associated with breast cancer and confirmed RAV221 is one of the top validated RAVs for TCGA-BRCA. Top 5 validated RAVs (*score*, bottom panel) and their average silhouette width (*avg.sw*, top panel) are shown. C) A word cloud of MeSH terms associated with RAV221. We collected MeSH terms assigned to the publications belonging to RAV221 and weighted them based on their prevalence and the contribution to any given RAV (Supplementary Methods). This word cloud shows that RAV221 is heavily composed of principal components from studies of breast neoplasms. D) Three studies contributing to RAV221. E) Top 10 enriched pathways in RAV221.

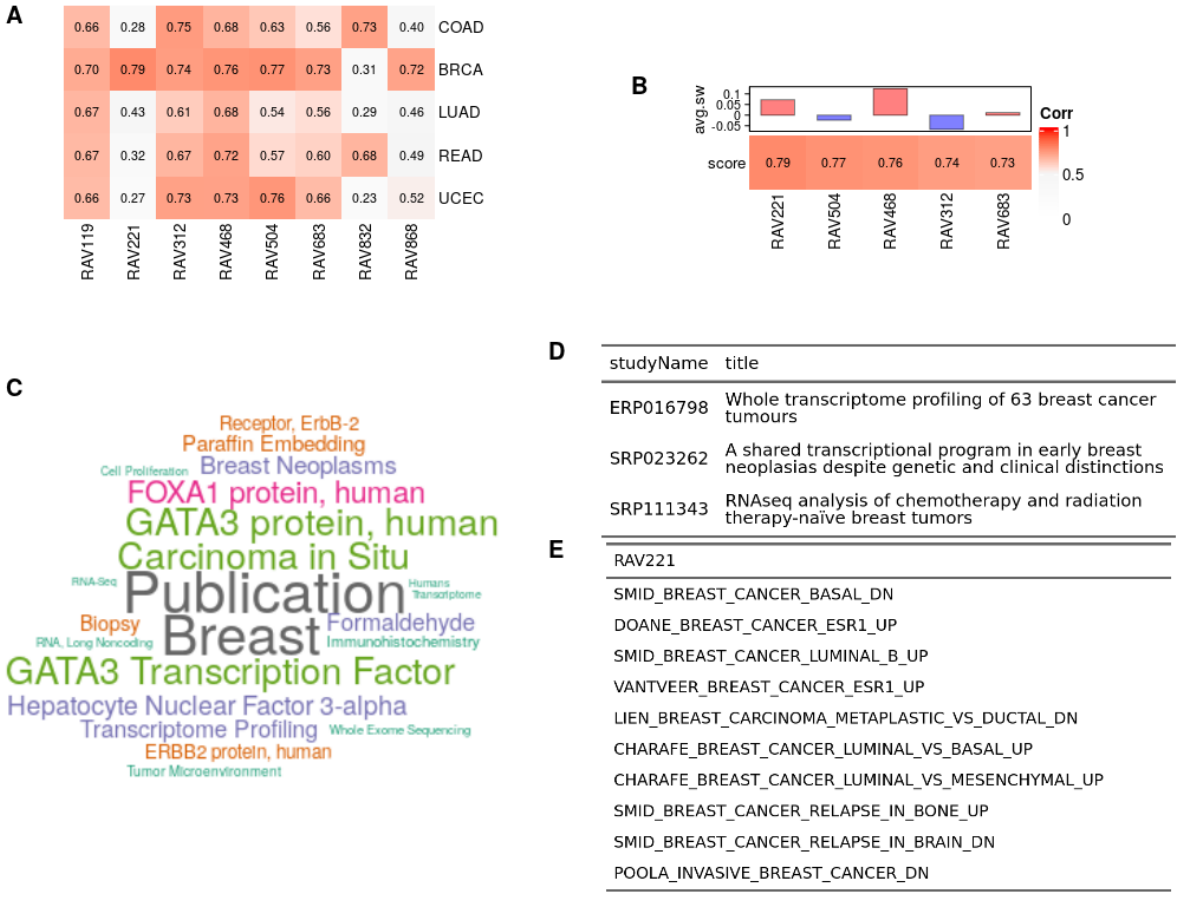


Fig 3. Sample scores for disease subtyping and metadata characterization

We assigned sample scores to 3,567 tumor samples from 18 CRC studies. A) The samples in each of 18 datasets, assigned to either (i) one of the 4 previously proposed CMS subtypes by CRC Subtyping Consortium[18,19] or (ii) not assigned to a CMS subtype (so 5 x 18 = 90 total groups), are represented by the mean (point) and standard deviation (error bar) of sample scores. CMS subtypes (colors) separate when plotted in RAV coordinates. We further evaluated the capacity of RAVs to demonstrate clinicopathological characteristics of colon cancer. B) Clinical phenotypes were regressed on discrete CMS subtypes and RAV834/833-assigned sample scores as covariates. LRTs were used to compare the full model to a simplified model containing only CMS subtype (left) or RAV834/833-assigned sample scores (right) as predictor. RAV834/833-only model shows -$\log_{10}$p-value near 0, implying that CMS is not providing additional information. C) The same regression and LRTs as in panel B were done using PCSS1/2 and RAV834/833-assigned sample scores as covariates. RAV834/833 outperforms PCSS1/2 on explaining colon cancer phenotypes except tumor location.

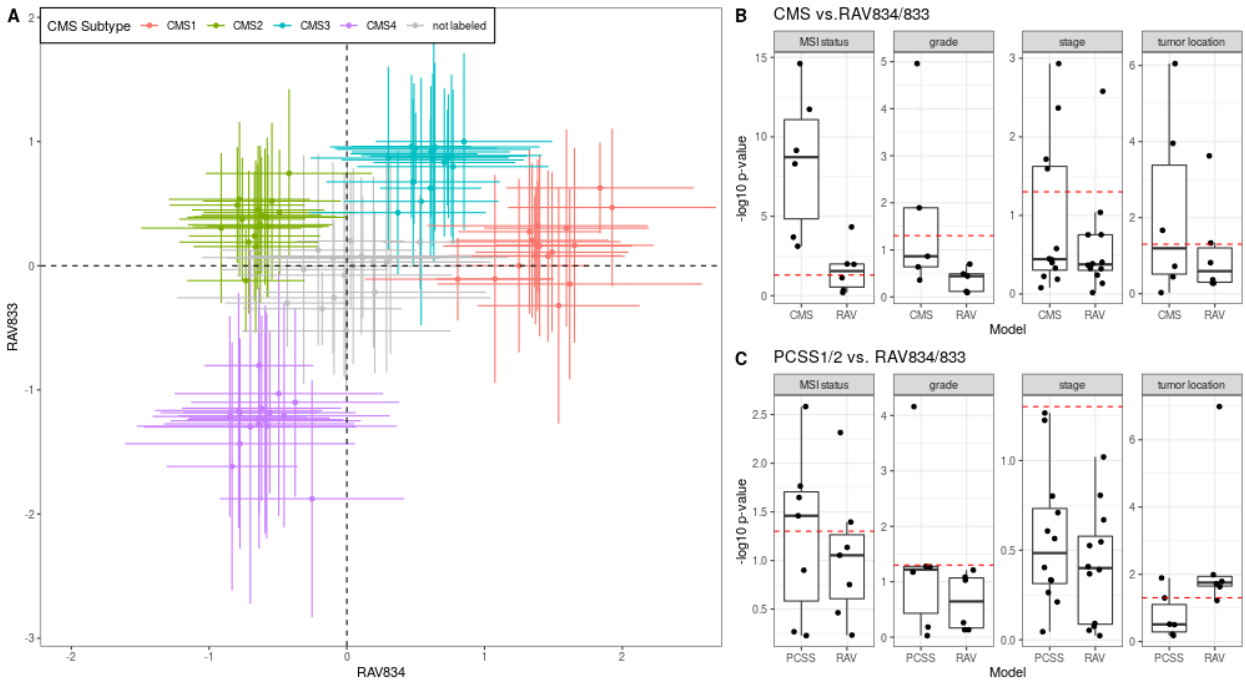
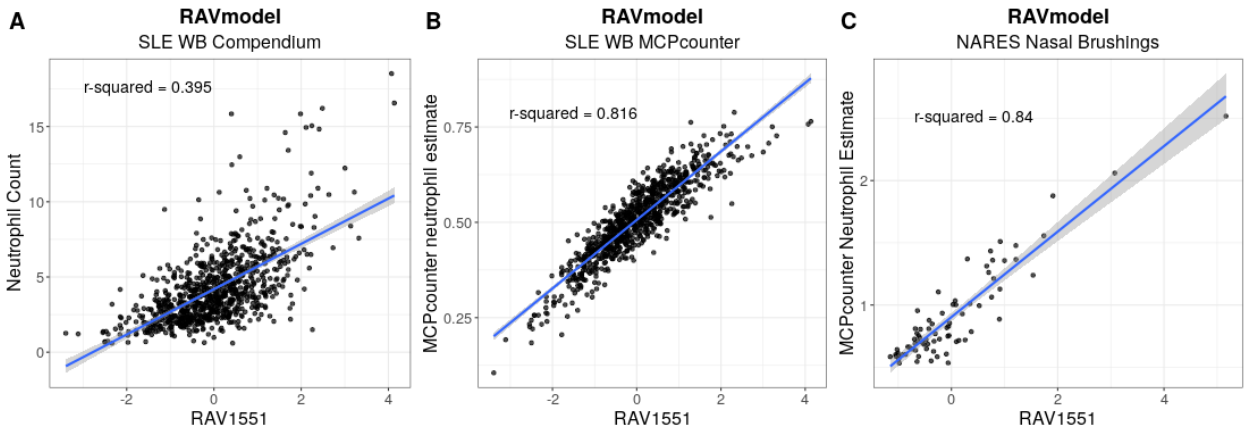


Fig 4. Estimate biological features of a new dataset using the signatures learned from public databases

RAVs encompass biological signals applicable across different platforms and independent datasets. We demonstrate this transfer learning capacity of RAVs by identifying the neutrophil-associated RAV from systemic lupus erythematosus whole blood (SLE-WB)[22] data and using the same RAV to analyze nasal brushing (NARES)[23] dataset. A) Neutrophil counts of SLE-WB data were plotted against RAV1551-assigned sample scores. B) Neutrophil count estimates by MCPcounter[8,24] were plotted against sample scores assigned by RAV1551. C) Neutrophil count of NARES samples were estimated by MCPcounter and plotted against RAV1551-assigned sample scores.



Methods

Source data

We used human RNA sequencing datasets from RNA-seq Sample Compendia in refine.bio[15], which hosts uniformly processed gene expression data from EBI's ArrayExpress, NCBI's GEO and SRA. Data were downloaded on April 10th, 2020, and we selected studies based on the following criteria: 1) Exclude studies with more than 1,000 samples because they are more likely to be single-cell RNA sequencing datasets. 2) Exclude studies assigned with a *Me*dical *S*ubject *H*eadings (MeSH) term, "Single-Cell Analysis". 3) Exclude studies with less than or equal to 50 successfully downloaded and imported samples (Supplementary Table 2 and Supplementary Fig. 2). After filtering, the complete compendium includes 536 studies (defined as a single SRA study) comprising 44,890 samples.

Processing training datasets

Training data included each sample's `quant.sf` file from Salmon outputs[26], not aggregated or normalized. We imported `quant.sf` files using `tximport`, scaling transcripts-per-million (TPM) using the average transcript length across samples and the library size (**"`lengthScaledTPM`"**)[27], followed by the log2 transformation. Ensembl transcript names were converted into gene symbols using the AnnotationDbi package[28]. Row normalization was done on all samples together, not at the individual study level, because correcting variability at the study level could remove the signals we want to capture[29]. For model building, we used 13,934 common genes among 536 studies' top 90% varying genes, where the variation cutoff was based on their study-level standard deviation (Supplementary Fig. 1A and Supplementary Methods).

Build RAVmodels

We performed PCA on pre-processed gene expression matrices independently for each study using the `stats::prcomp` R function for each gene centered but not scaled. Loading vectors of top 20 PCs from 536 studies (total 10,720 PCs) were clustered via hierarchical clustering. For hierarchical clustering, we calculated the distances between loadings using Spearman's correlation coefficient and clustered them with ward.D agglomeration method. The number of clusters was set to the minimum number that can separate up to 50 negative controls (Supplementary Methods and Supplementary Fig. 3). PCs in each cluster were averaged and the resulting 'genes x averaged loadings' matrix, RAVindex, was combined with associated metadata, GSEA and MeSH annotations into a unified data structure that we term a RAVmodel. We performed a detailed evaluation of model building criteria including varying gene sets used for annotation, the size of training datasets, the number of clusters, and so on (Supplementary Table 4, Supplementary Methods).

Annotation

We subjected each RAV to *G*ene *S*et *E*nrichment *A*nalysis (GSEA) to interpret the biological signals associated with it. Genes were ordered by loading value from each RAV and supplied as a `geneList` input for `clusterProfiler::GSEA`[30]. We excluded gene sets with fewer than 10 genes and more than 500 genes. Enriched pathways with adjusted p-value < 0.05 and the subset of GSEA results - `NES`, `Description`, `pvalue`, and `qvalues` - were included in the RAVmodel. The version of RAVmodel used in this study is annotated with Molecular Signatures Database (MSigDB) curated gene sets (C2, version 7.1)[31,32].

MeSH terms[33] were assigned to each study using the NCBI Medical Text Indexer (MTI) tool[34]. The relevance of MeSH terms in each RAV was assessed through the bag-of-words model: all the MeSH terms associated with the training datasets were considered as the 'universe' and each term in the cluster was reverse-weighted by the frequency of the given term in the universe. MeSH terms were also weighted by the variance explained by the principal component that they came from (Supplementary Methods).

Validation datasets

Five TCGA RNA sequencing datasets (COAD, BRCA, LUAD, READ, and UCEC) were acquired from GSEABenchmarkeR[35]. Any genes with count-per-million (CPM) less than 2 were excluded and the count matrix was log2-transformed and centered but not scaled before PCA. Eighteen colon cancer microarray datasets from curatedCRCData were also used for validation[36]. Missing and infinite values were removed from these microarray data and the remaining expression values were centered at each gene level. To evaluate the ability to perform transfer learning to metadata, we used preprocessed versions of nasal brushing (NARES)[23] and the systemic lupus erythematosus (SLE) datasets[22].

Reproducible Research

All analyses presented here are reproducible using code accessible from https://github.com/shbrief/GenomicSuperSignaturePaper/.

Supplementary Materials

Supplementary Methods

Sensitivity analysis for model building

Datasets for methods searching

During the optimization process for RAVmodel building, we used the well-characterized and smaller datasets: 8 colon cancer datasets from curatedCRCData, 10 ovarian cancer datasets from curatedOvarianData, and recount2 datasets used for recount2-multiPLIER model. Training datasets for recount2-multiPLIER model and the current version of RAVmodel are partially overlapping: recount2-multiPLIER model used 37,027 runs from 30,301 unique samples from 1,466 studies and GenomicSuperSignature was constructed from 44,890 runs from 34,616 unique samples from 536 studies. Among them, only 6,839 runs from 5,260 unique samples from 87 studies were used by both models. In addition to the different combinations of these datasets, we created synthetic datasets that can serve as positive and negative controls.

Dimensionality reduction methods

We assessed multiple dimensionality reduction methods for RAVindex building. Nonnegative matrix factorization (NMF) was excluded because there is no clear criterion to select representative components, such as variance explained by each principal component in PCA. Non-orthogonal relationship between components captured by NMF is potentially a more relevant representation of biological data, but by combining replicative principal components, we overcome the orthogonality constraint imposed by PCA. We also ruled out independent component analysis (ICA) because it assumes that subcomponents are independent to each other, which we considered not the case for biological data. We narrowed down to two dimensionality reduction methods, PCA and PLIER[7], and investigated the types of signatures when they were applied at the dataset-level or sample-level. This comparison was done within four different conditions: perPCA (PCA on each dataset and cluster top PCs), megaPCA (PCA on all samples), perPLIER (PLIER on each dataset and cluster latent variables (LVs, equivalent to principal components from PCA)), and megaPLIER (PLIER on all samples, identical to multiPLIER). One of the downsides in megaPLIER approach was that the direct link between LVs and the training data was not available. Also, the annotation database was inseparable from the model building, making it harder to scale. perPLIER approach blended LVs in each cluster and lost distinct signatures. Like megaPLIER, megaPCA did not maintain the links between signature and it's source data. Additionally, megaPCA picked up only a handful of strong signatures in top PCs, which we can still capture through the perPCA approach. Overall, we decided to use the perPCA approach for our model building because it is more scalable, keeps

the link between signature and source data, and captures both pan-data and per-data signatures.

Data transformation

We did log2-transformation and row-normalization across all samples, not at the dataset level, to keep the differences in the datasets. VST transformation[17] was excluded because it requires significantly more computing resources without any meaningful improvement on capturing biological signatures over log2-transformation.

Subset genes

We searched for the minimum set of genes that keeps the biological information, because more genes require more computing resources to process and some genes were measured only in certain training datasets. Also, low- or non-expressing genes can be rather a noise, making it harder to interpret the result. First, instead of using a fixed cutoff for 'low-expressing' genes, we selected genes based on their expression variance within the dataset because we suspect that genes with the stable expression level within a dataset convey less information to capture. We also tried the common genes between training datasets and annotation databases before PCA. However, it didn't improve the accuracy of GSEA compared to the common genes among training datasets and made the model less scalable because RAVindex needs to be rebuilt for RAVmodels with different annotation databases even for the same training datasets. So for the model building, we used the common genes among the top 90% varying genes from each training dataset.

The number of PCs to collect

After the 8th PC, less than 5% of the variance was explained by a single PC for all 536 training datasets (data not shown). As a conservative choice we included the top 20 PCs for model building in case any large, complex dataset still contains biologically relevant signals in the lower PCs. Using a varying number of PCs would add a complexity to the process that seemed unjustified given that the variance explained by each PC does not vary much by study size.

Clustering methods

To group the replicative PCs, we tried centroid-based clustering such as k-means, graph-based clustering, and connectivity-based clustering like hierarchical clustering. We tried them on top 20 PCs from 8 colon cancer datasets and for evaluation, the cluster membership was compared with previously identified signatures (PCSSs) using Jaccard index. We also tested them on top 5 PCs from 10 positive and 10 negative controls, which were synthetic datasets created through bootstrap and random sample selection, respectively. We evaluated each clustering method based on how well top PCs from positive controls were clustered together and top PCs from negative controls were assigned into different clusters.

For the centroid-based clustering methods (k-means and k-medoids), we searched the optimum number of clusters using multiple measures including elbow method, mean silhouette width, and

within sum of squares. However, the number of clusters required to separate unrelated PCs was too high to keep the related PCs together, which couldn't be improved with different distance metrics. We suspect that PCs from biological data do not possess spherical or ellipsoidal symmetry required for centroid-based clustering to work.

We evaluated graph-based clustering and hierarchical clustering with the different combinations of distance metrics and agglomeration methods (for hierarchical clustering) on the same datasets used for the centroid-based clustering tests. From this evaluation, the clustering schema was narrowed down into two versions: for graph-based clustering, edge-betweenness clustering using edge weight by Spearman correlation and for hierarchical clustering, using Spearman distance and ward.D agglomeration. When we applied these clustering approaches on larger datasets, however, there were scalability issues with graph-based clustering: it formed a very large cluster, containing more than 5% of PCs, that cannot connect related datasets due to the extreme distribution of the cluster sizes. So we decided to use hierarchical clustering based on Spearman distance with ward.D agglomeration.

Choose the optimum number of clusters for hierarchical clustering

We used the negative-control dataset to explore the optimum number of clusters for hierarchical clustering. First, we constructed 50 synthetic datasets by randomly selecting 50 samples from 44,890 samples. We scrambled genes in each of 50 synthetic datasets and added random value between -0.1 and 0.1. The mean and standard deviation of 44,890 samples were used for row normalization of the synthetic datasets. We confirmed that these synthetics datasets can serve as a negative control based on the distance matrix: the minimum and maximum distance of PC1s from the synthetic datasets ranged approximately between 1st and 3rd quarters of the distance distribution of PCs from the actual training datasets, which we want to separate during the clustering process (Supplementary Fig. 3A). We collected top 20 PCs from 536 training datasets and PC1s from varying numbers of synthetic datasets (10, 20, 30, 40, and 50) and performed hierarchical clustering with the different numbers of clusters. Nine different cluster numbers were applied to each datasets: those nine cluster numbers were `round({#ofPCs}/d)`, where d is 7, 6, 5, 4, 3, 2.75, 2.5, 2.25, and 2. All negative controls were separated when `d = 2.25`, regardless of the number of negative controls (Supplementary Fig. 3B and 3C). So for the current versions of RAVmodel, we selected 4,764 clusters (= `round((20x536)/2.25)`).

Model validity

To test whether heterogeneous datasets can maintain the signatures from the focused dataset, we first built RAVindex from the focused training datasets and gradually diversed the training datasets with the unrelated datasets. A rate of overlapping enriched pathways over correlated pathways was monitored, from which we confirmed that our RAVindex building process reliably maintains the dataset-specific signatures from the heterogeneous training datasets.

Weighting MeSH terms in each cluster

The significance of MeSH terms associated with each cluster was evaluated based on their exclusivity. However, the simple sum of associated MeSH terms can be inappropriate in some cases. For example, potential noises can be a predominant signal in small clusters and common MeSH terms, such as 'human' or 'RNA sequencing' for the current version of RAVmodel, can be overrepresented and cover all the other terms. To handle these extreme situations, we applied several filtering and normalization terms. If the cluster contains less than 8 PCs, we considered any MeSH terms appearing half down of 'cluster size x 0.5' as noise and removed them. If the cluster has more or equal to 8 PCs in it, any MeSH terms appearing less than or equal to 4 times were eliminated. These cutoff values for 'noise' can be modified by users to fit their needs. Remaining MeSH terms were assigned with the new score based on the variance explained by PCs and the sum of scores for a given MeSH term was divided by the frequency of that term in the universe, where the 'universe' is defined as the complete pool of MeSH terms associated with all the training datasets. This final score can be displayed as a table or a word cloud, using `meshTable` or `drawWordcloud` functions, respectively.

Prepare input dataset

The GenomicSuperSignature takes gene expression profiles generated from both microarray and RNA sequencing with the minimum pre-processing. The major requirement is that gene expression profile should follow the normal distribution.

Supplementary Results

Software implementation

The PCAGenomicSignatures class inherits SummarizedExperiment data structure and stores RAVindex, metadata, and annotation, which we collectively refer to as the RAVmodel (Supplementary Fig. 1B). Functions and S4 methods for the PCAGenomicSignatures class to access components of the RAVmodel, visualize analyses, and interpret new datasets are implemented in the GenomicSuperSignature R/Bioconductor package (Supplementary Table 1). We provide different versions of RAVmodels based on gene sets used for GSEA annotation and they are readily available (as .rds format files) to download from the internet through the `getModel` function or `wget` (Supplementary Table 4).

Two key functions, `validate` and `calculateScore`, allow interpretation of new datasets at the study level and at the individual sample level, respectively. The validate function calculates Pearson correlation coefficients between the top 8 PCs of a dataset and all RAVs (Supplementary Methods), from which the highest value is assigned as a 'validation score' of the corresponding RAV. Validation score provides a quantitative representation of the relevance between a new dataset and RAV. In general, the higher validation score implies that the RAV

explains a more significant feature of a dataset. Validation outputs can be visualized as a heatmap table (Fig. 2A and 2B) and an interactive plot (Supplementary Fig. 7), through heatmapTable and plotValidate functions, respectively. Average silhouette width of each cluster is available as a reference for quality control and as an additional filtering option to find significant RAVs. The calculateScore function calculates a RAV-assigned 'sample score' to each sample, which is the matrix multiplication result (B^t, red) between the 'samples x genes' matrix (Y^t, grey) and RAVindex (Z, blue) (Supplementary Fig. 1C). Similar to validation score, sample score provides a quantitative representation of the relevance between samples and the given RAV.

In addition to the study-level validation scores and the sample scores acquired through gene expression profile, we can access the knowledge graph created by GenomicSuperSignature through various entry points, such as metadata, MeSH term, and keywords, because RAVmodel maintains the link between RAV and it's source data. (Fig. 5).

Guides for interpretation

We applied RAVmodel on a human B cell dataset (GSE2350, from bcellViber package[37]), which consists of 211 normal and tumor human B-cell, and checked how many of the blood cell related pathways were found in the top and bottom 10 enriched pathways of the selected RAVs (data not shown). GSEA annotation of RAVmodel used the ranked gene list from RAV, which is not directly based on the gene expression level but based on the variation of the gene expression. We suspect both top and bottom enriched pathways can be equally informative, so we analyzed both of them. From this analysis, we confirmed that the high validation score serves as a major criterion for the best RAV and the larger average silhouette width can be a secondary standard if the validation score is similar. So we recommend using both validation score and average silhouette width to find the RAV best explaining your dataset. Also, any RAV with a single element is, by definition, not a replicative signal and requires an extra care on interpretation. The current RAVmodel from 536 training datasets contains 1,378 single-element RAVs; only one of them is PC1, none is from PC2, and the vast majority are later PCs. We suspect most of these one- and two-element clusters are not representing any significant information.

RAVs for colorectal cancer characterization

To evaluate the performance of RAVs compared to PCSSs, we searched for colon cancer associated RAVs in three different ways. First, we ran Kruskal-Wallis rank sum test between CMS subtypes and RAV-assigned scores. Two RAVs with the highest chi-square, RAV834 and RAV833, were selected for further testing. Second, we identified two RAVs, RAV1575 and RAV834, with the highest Pearson correlation coefficients with PCSS1 and PCSS2, respectively (0.59 and 0.56). Last, we calculated validation scores for 18 colon cancer datasets from curatedCRCData[36] and collected top 10 validated RAVs from each dataset. We summarized the frequency of different RAVs validating each dataset without any additional filtering criteria and selected the top 2 most frequently validated RAVs, RAV188 and RAV832, which were captured 14 and 10 times, respectively (Supplementary Table 5). In spite of the major difference in

training datasets, RAV834/833 showed a comparable performance on colon cancer subtyping to PCSS1/2 (Fig. 3A). Notably, RAVs identified by CMS metadata (RAV834/833) performed better at CRC subtyping than the validated RAVs (RAV188/832), suggesting that the most prominent feature shared by 18 CRC datasets is not their disease subtypes (Fig.3A and Supplementary Fig. 5D).

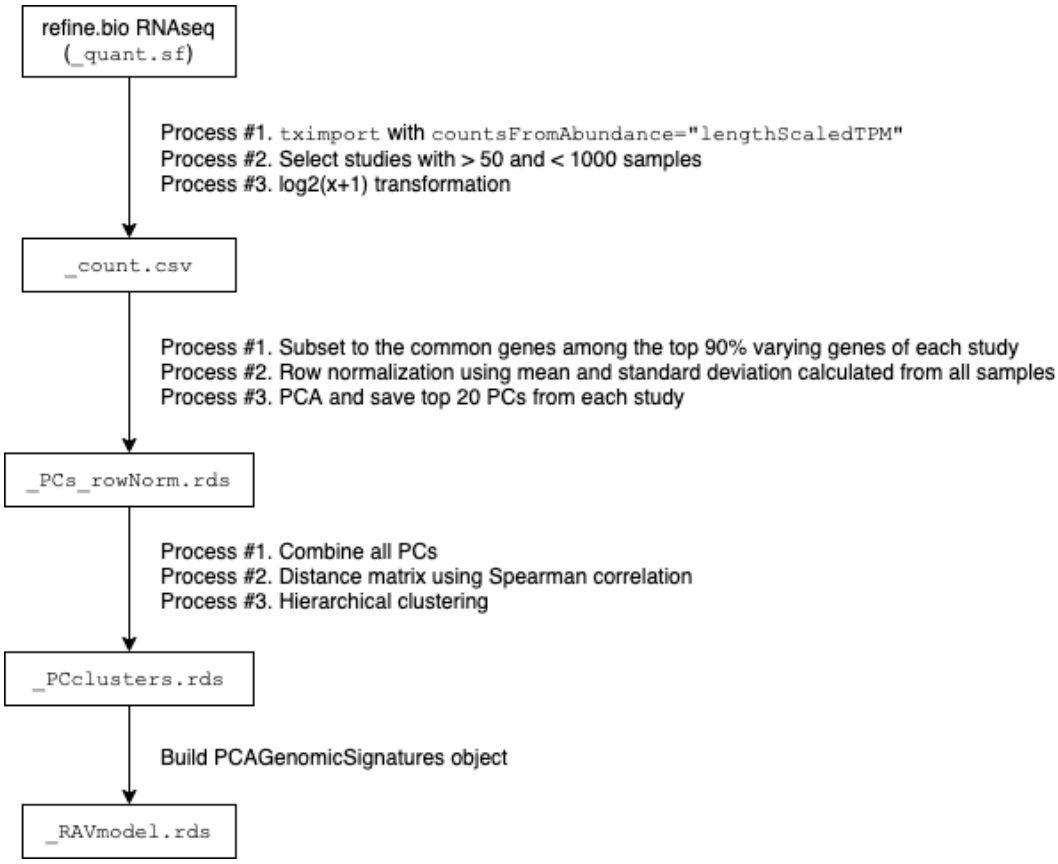
PCA plot annotated with pre-calculated GSEA

One of the widely used exploratory data analysis methods is PCA and a PCA plot can provide a quick overview of sample composition and distribution. However, the interpretation of different PCs is not readily available in the conventional PCA. We couple PCs from new data with GSEA annotation of RAVmodel, and enable the instant interpretation of PCA results. We showed this example using a microarray dataset from isolated immune cells (E-MTAB-2452)[38] and RAVmodel annotated with three priors from the PLIER package (RAVmodel_PLIERpriors, Supplementary Table 4). GenomicSuperSignature performs PCA on a centered, but not scaled, dataset and identifies the most relevant RAV for each PC. GSEA annotation of these matched RAVs can be summarized in a table (Supplementary Fig. 8A, `annotatePC` function). If you want to draw a PCA plot - currently any pair of top 8 PCs is supported, GSEA annotation will be displayed as a linked table (Supplementary Fig. 8B, `plotAnnotatedPCA` function).

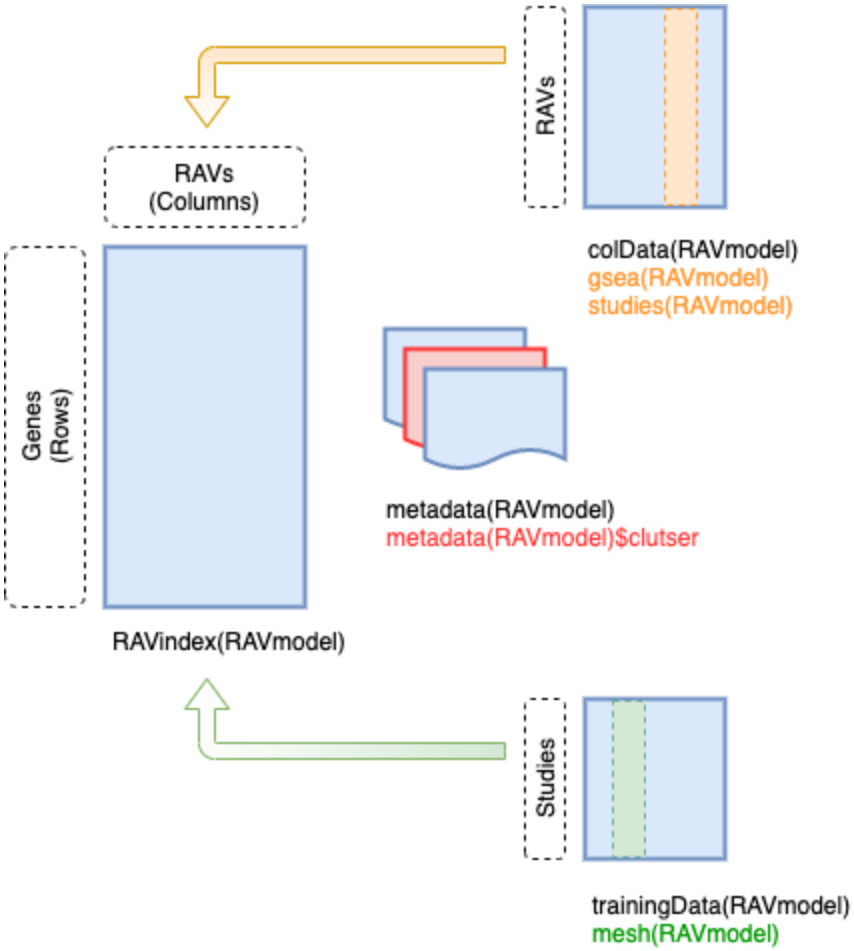
Supplementary Figures

Sup.Fig 1. Overview of RAVmodel building

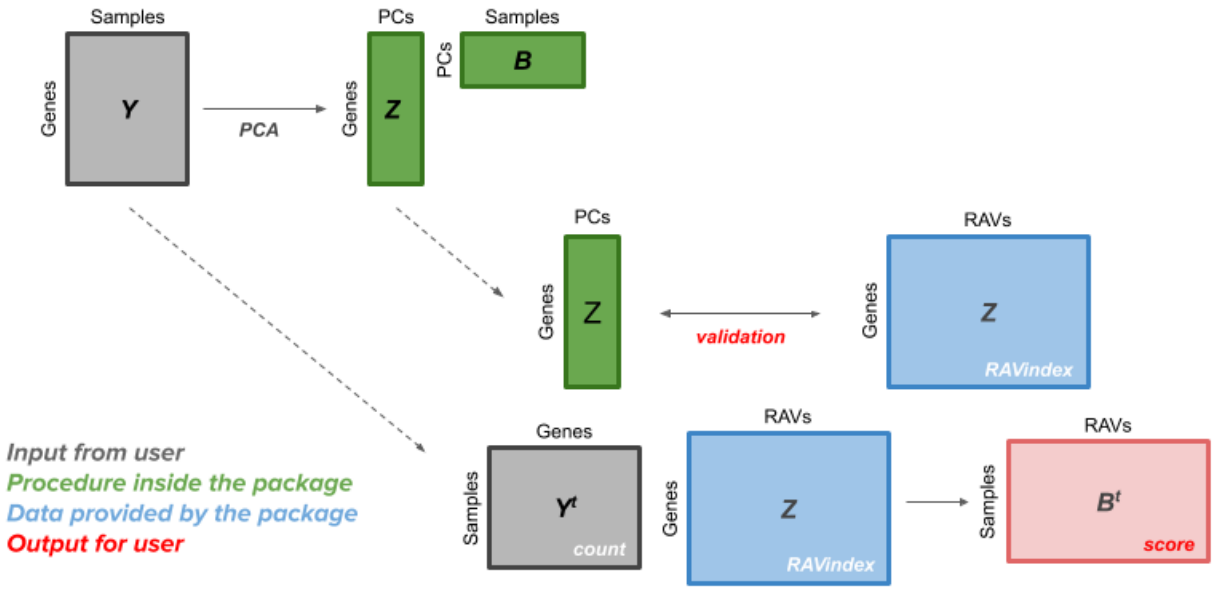
A) We downloaded human `quant.sf` files from refinebio RNA-seq Sample Compendia. We subset studies with more than 50 and less than 1,000 samples from 6,460 studies available at the time of the snapshot. Some RAVmodels retain additional filtering criteria on their training datasets. For example, the current version of RAVmodel predominantly used in this study further excludes datasets potentially from single cell analysis. Selected datasets were imported through `tximport`, followed by log2 transformation to bring them close to normal distribution. We used the common genes among the top 90% varying genes of each study, which was 13,934 genes for the current RAVmodel, and did row normalization using mean and standard deviation calculated from all 44,890 samples. PCA was done on a row-normalized expression matrix at the study level and top 20 PCs from each study were collected, ending up with 10,720 PCs. Distance matrix between these PCs was calculated using Spearman correlation and hierarchical clustering was applied with the pre-defined optimum number of clusters. Weighted MeSH terms and GSEA on each RAV, along with RAVindex and other metadata, were assembled into PCAGenomicSuperSignatures object, named as RAVmodel. In the below workflow diagram, boxes represent the intermediate files we created during the model building process.



B) Schematic of PCAGenomicSuperSignatures object. RAVindex is a ‘genes x RAVs’ matrix. The colData provides information on RAVs, such as studies contributing to each RAV and GSEA results from each RAV. The metadata stores details on the RAVmodel itself, such as cluster memberships of PCs and the size of each cluster. The trainingData provides information on studies used for the model training, which includes MeSH terms assigned to each study and PCA summary of each study.

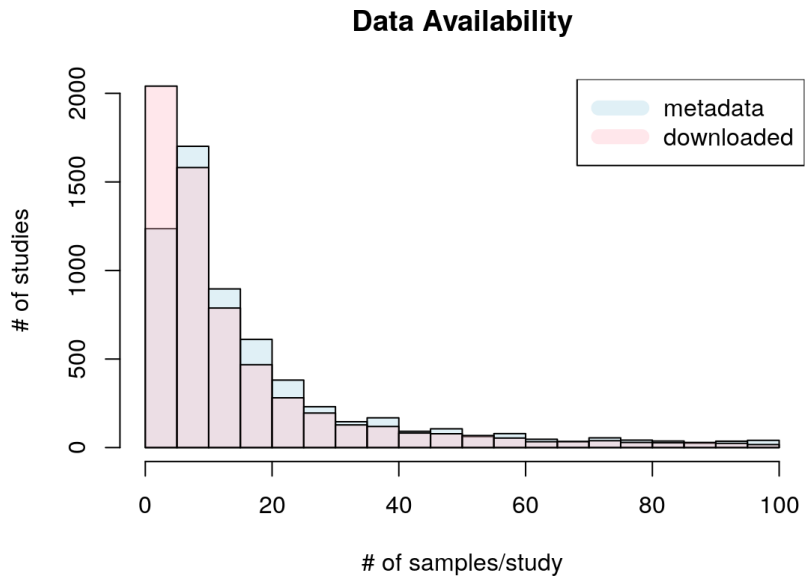


C) User's perspective. The GenomicSuperSignature package allows users to access a RAVmodel (Z matrix, blue) and annotation information on each RAV. From a gene expression matrix (Y matrix, grey), users can calculate dataset-level validation score or sample score matrix (B matrix, red). Through the RAV of your interest, additional information such as related studies, GSEA, and MeSH terms can be easily extracted.



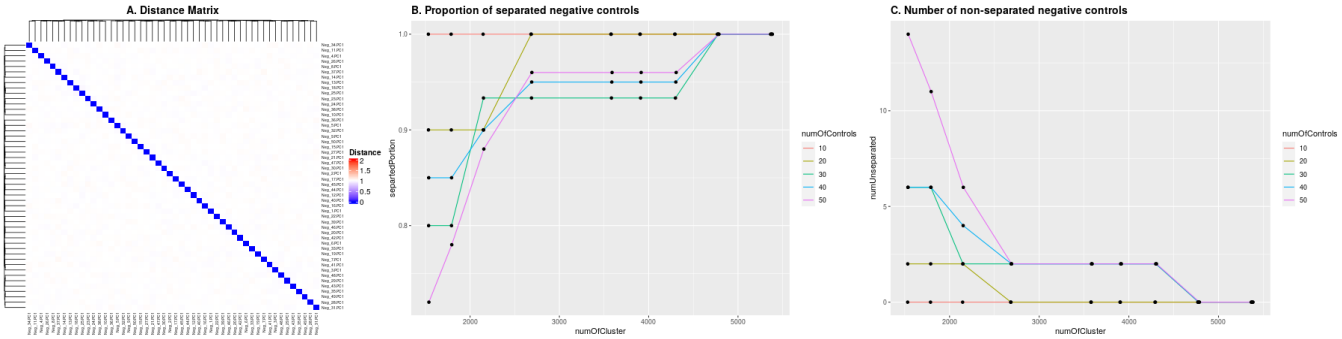
Sup.Fig 2. Data availability from refine.bio

The refinebio database is actively updated and our current RAVmodel is based on the snapshot on April 10th, 2020. Metadata bar (light blue) shows the number of studies with the given ranges of study sizes based on the metadata. Downloaded bar (pink) represents the number of studies with the given ranges of study sizes that were successfully downloaded and imported through `tximport`. Based on metadata, there were studies with more than 100 samples, but at the time of snapshot, only up to 100 samples were available. Thus, the plot displays only up to 100 samples/study cases. Due to the unavailability of certain samples, more studies belong to 0-5 samples/study bracket than metadata suggests.



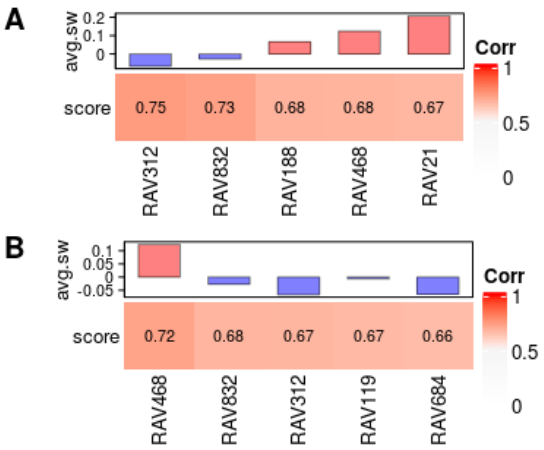
Sup.Fig 3. The optimum number of clusters

We used PC1s from synthetic datasets, designed as negative controls, to decide the number of clusters for hierarchical clustering. Training datasets, top 20 PCs from 536 studies (`RAVmodel_536` column in Supplementary Table 2), were combined with PC1s from the different numbers of synthetic datasets (10, 20, 30, 40, and 50) (Supplementary Methods). A) Heatmap of the distance matrix between 50 negative controls. Distance was calculated based on Spearman's correlation. B) Proportion of the negative controls that were separated with the given cluster number. `numOfControls` is the number of negative controls added to the training datasets. `numOfCluster` is the round of the total PCs (from training datasets and negative controls) divided by 7, 6, 5, 4, 3, 2.75, 2.5, 2.25, and 2. Different numbers of negative controls were completely separated when we used the cluster number `k = round((the number of PCs)/2.25)`. C) Number of negative controls that were not separated.



Sup.Fig 4. Colon and rectal cancer associated RAV

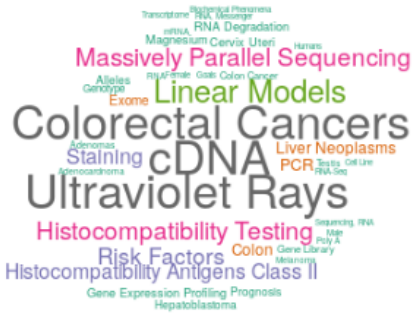
Based on Fig. 2A, RAV832 seems to be associated with TCGA-COAD and TCGA-READ. Top validation results of A) TCGA-COAD and B) TCGA-READ include RAV832 with the negative average silhouette width. C) MeSH terms associated with RAV832. D) Studies contributing to RAV832. E) MSigDB C2 gene sets enriched in RAV832.



D

studyName	title
SRP010181	Derivation of HLA types from shotgun sequence datasets
SRP029880	Gene expression profiling study by RNA-seq in colorectal cancer
SRP068591	Gene signature in sessile serrated polyps identifies colon cancer subtype
SRP073267	Impact of RNA degradation on fusion detection by RNA-seq
SRP123604	Immune Profiling of Premalignant Lesions in Patients with Lynch Syndrome

C

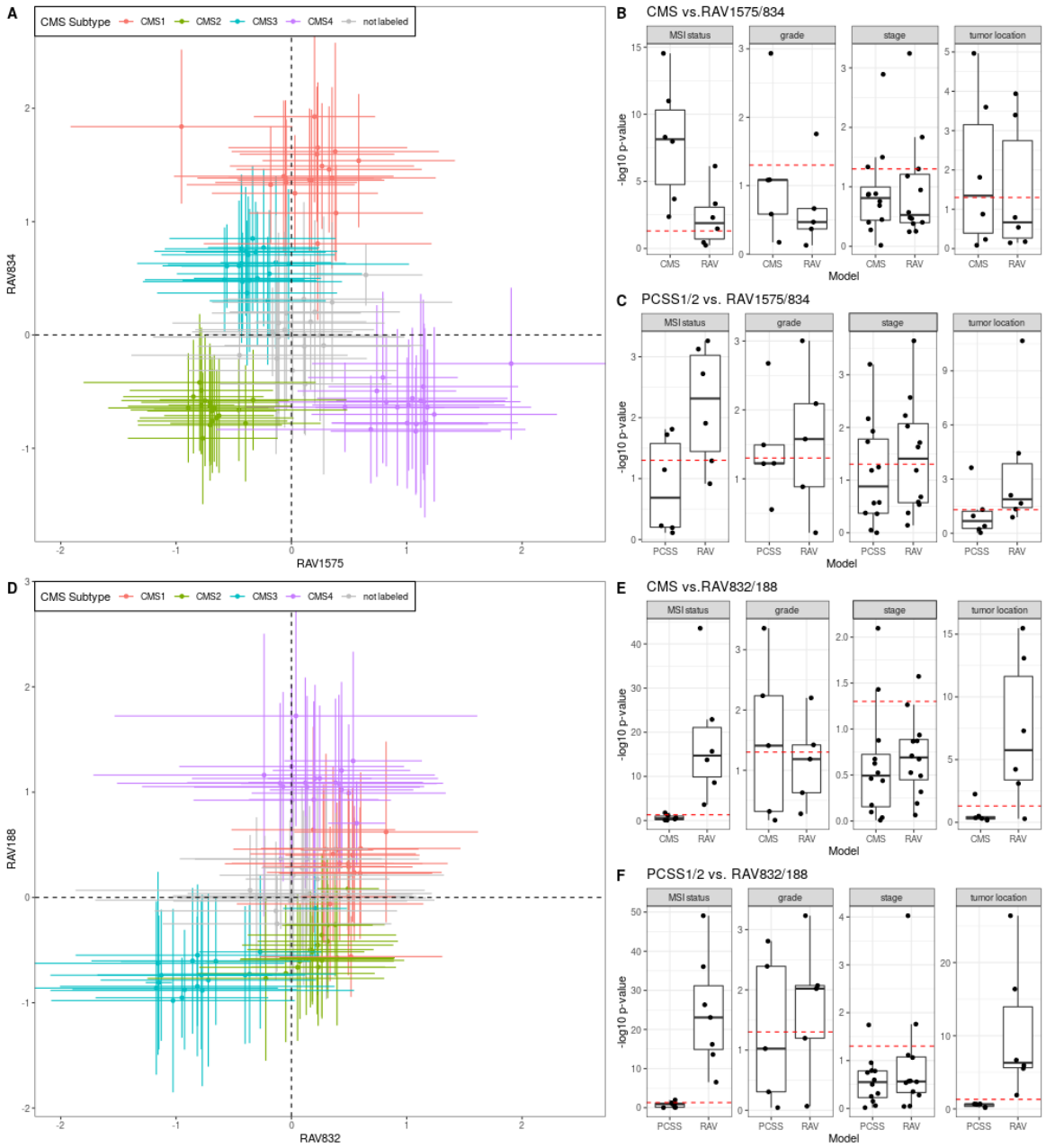


E

RAV832
SABATES_COLORECTAL_ADENOMA_UP
HSIAO_LIVER_SPECIFIC_GENES
GRADE_COLON_AND_RECTAL_CANCER_UP
REACTOME_EUKARYOTIC_TRANSLATION_ELONGATION
REACTOME_SRP_DEPENDENT_COTRANSLATIONAL_PROTEIN_TARGETING_TO_MEMBRANE
KEGG_RIBOSOME
LEE_LIVER_CANCER_ACOX1_DN
REACTOME_SELENOAMINO_ACID_METABOLISM
KOBAYASHI_EGFR_SIGNALING_24HR_DN
REACTOME_REGULATION_OF_EXPRESSION_OF_SLITS_AND_ROBOS

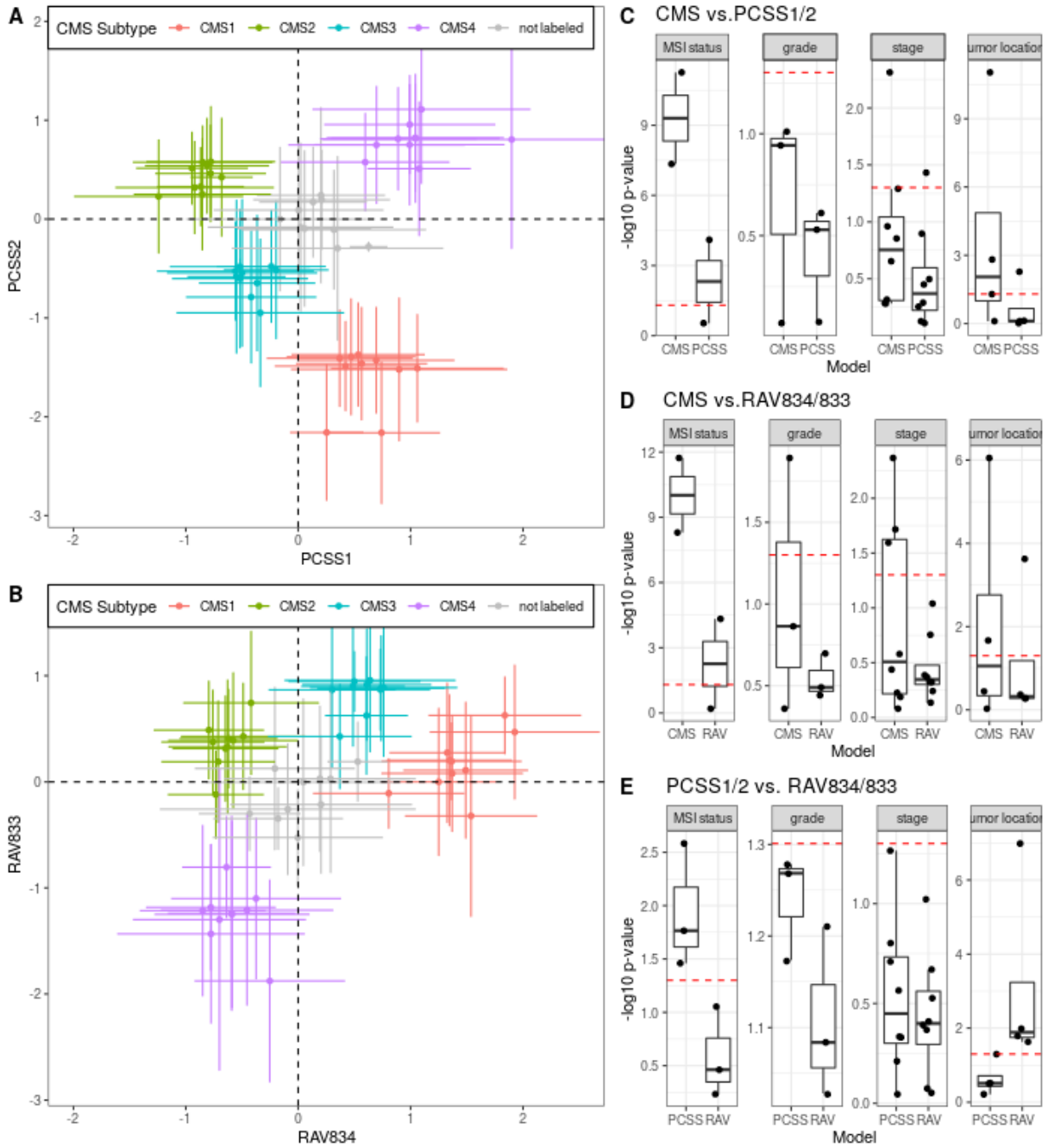
Sup.Fig 5. CRC characterization with different RAVs

In the results, we described two additional pairs of RAVs, RAV1575/834 and RAV188/832, that are potentially useful for CRC characterization. We applied the same analysis procedure on 18 CRC datasets as in Fig. 3 using those two pairs of RAVs. (A-C) CRC characterization with RAV1575/834. RAV1575 and RAV834 were identified based on their similarity to PCSS1 and PCSS2, respectively. (D-F) CRC characterization with RAV188/832. RAV188 and RAV832 were most frequently found among the top 10 validated RAVs.



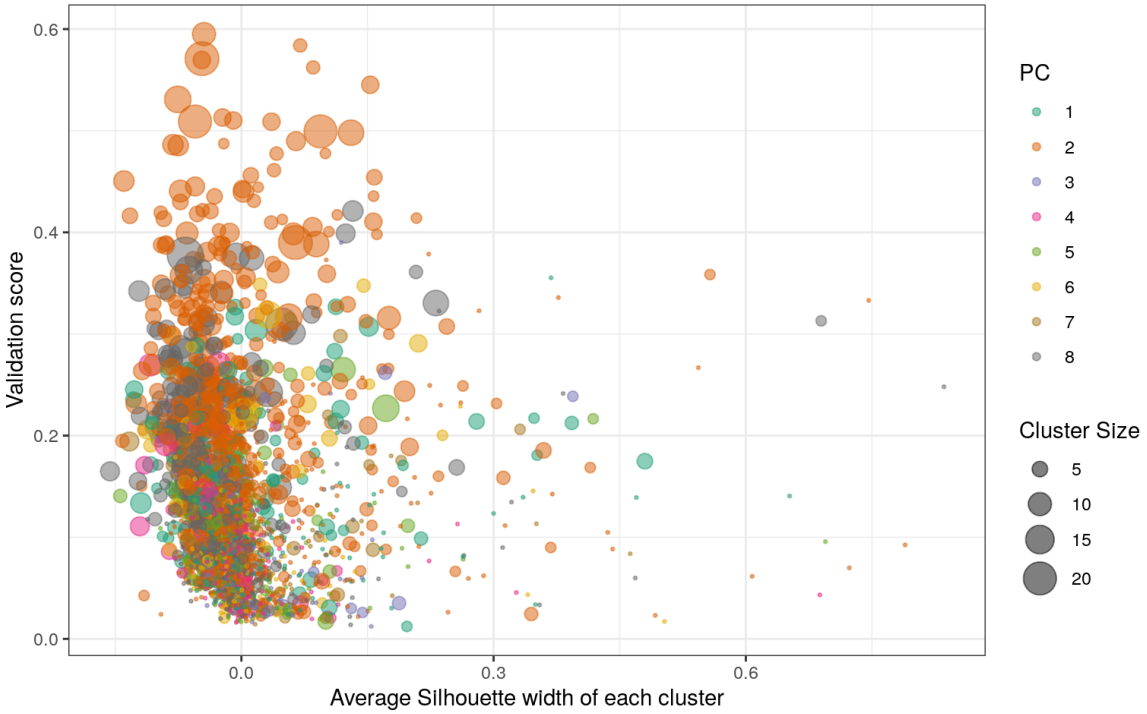
Sup.Fig 6. CRC characterization with 10 validation datasets

Analyses in Fig. 3 were repeated with only 10 CRC datasets, excluding 8 datasets used to train PCSSs. A) Subtype- and study-specific mean of PCSS1 and PCSS2 scores are plotted as points while the error bars represent standard deviation. B) The same plotting scheme as A was applied on RAV834 and RAV833-assigned scores. C-E) LRTs compare the full model to a simplified model containing only C) CMS subtypes or PCSS1/2, D) CMS subtypes or RAV834/833, and E) PCSS1/2 or RAV834/833.



Sup.Fig 7. Overview validation results via an interactive plot

In Fig 2B, we used a table format to display the validation results. To understand the overall validation pattern for each PCs of new data, we provide an interactive plot as one of the visualization options. Here, we plotted the validation plot of Human B-cell expression dataset (GSE2350) generated from microarray. X-axis represents the average silhouette width and y-axis represents the validation score. Each point represents RAV, where the color shows the PC with the highest validation score for a given RAV. The point size reflects the cluster size, the number of PCs contributing to a given RAV. In general, we interpret that the points toward the upper right corner with the intermediate sizes are more relevant to new data than the others. Interactive form of this graph is available with the argument `interactive=TRUE`, where you can hover each data point for more information, such as cluster number and exact cluster size.



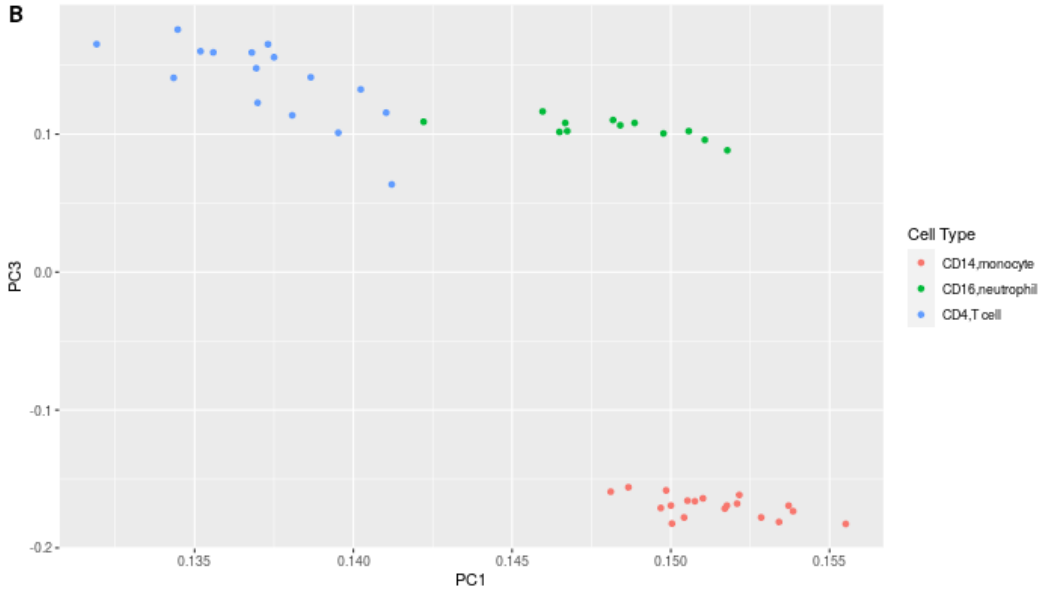
Sup.Fig 8. PCA with GSEA annotation

PCA result of leukocyte gene expression data (E-MTAB-2452) is displayed in A) a table or B) a scatter plot. PCA is done on a centered, but not scaled, input dataset by default. Different cutoff parameters for GSEA annotation, such as minimum validation score or NES, can be set.

A

PC1-RAV23	PC2-RAV1552	PC3-RAV1387	PC4-RAV684
SVM T cells CD8	IRIS_Monocyte-Day0	MIPS_55S_RIBOSOME_MITOCHONDRIAL	REACTOME_CELL_CYCLE
SVM T cells CD4 naive	IRIS_DendriticCell-Control	REACTOME_RESPIRATORY_ELECTRON_TRANSPORT_ATP_S...	REACTOME_CELL_CYCLE_MITOTIC
SVM T cells follicular helper	DMAP_MONO2	MIPS_39S_RIBOSOMAL_SUBUNIT_MITOCHONDRIAL	NA
SVM T cells regulatory (Tregs)	IRIS_Monocyte-Day7	REACTOME_RESPIRATORY_ELECTRON_TRANSPORT	NA
SVM T cells gamma delta	SVM Monocytes	REACTOME_TCA_CYCLE_AND_RESPIRATORY_ELECTRON_T...	NA

B



PC1-RAV23	PC3-RAV1387
SVM T cells CD8	MIPS_55S_RIBOSOME_MITOCHONDRIAL
SVM T cells CD4 naive	REACTOME_RESPIRATORY_ELECTRON_TRANSPORT_ATP_S...
SVM T cells follicular helper	MIPS_39S_RIBOSOMAL_SUBUNIT_MITOCHONDRIAL
SVM T cells regulatory (Tregs)	REACTOME_RESPIRATORY_ELECTRON_TRANSPORT
SVM T cells gamma delta	REACTOME_TCA_CYCLE_AND_RESPIRATORY_ELECTRON_T...

Supplementary Tables

Sup.Table 1. Summary of new terms

Terms	Description
GenomicSuperSignature	Name of the R/Bioconductor package that contains all the functions to apply RAVmodel to new data, serving as a 'toolbox'. RAVmodels stored in Google Bucket are downloadable using the `getModel` function of the package.
GenomicSignatures-object	Data structure inherited from SummarizedExperiment
PCAGenomicSignatures-object	Data structure inherited from GenomicSignatures
RAV (Replicable Axes of Variation)	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	PCAGenomicSignatures-object. It 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 'samples x genes' 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.

Sup.Table 2. Training datasets used in this study

https://github.com/shbrief/GenomicSuperSignature/blob/master/inst/extdata/studyMeta.tsv

The study accession, the number of samples, and the title of 536 training datasets used to construct the current version of RAVmodel (`RAVmodel_536` column). The number of samples for each study based on the metadata (`metadata` column) and the number of samples actually used (`downloaded` column) are different due to the data availability at the time of download.

Sup.Table 3. Source types of training datasets

https://github.com/shbrief/GenomicSuperSignaturePaper/blob/master/inst/extdata/source_name_annotated.tsv

We obtained the source name for 435 studies (~81.2% of all training datasets) from OmicIDX[39] and did a manual curation based on the source name (`source_name` column) to understand the types of training datasets. Curation covered four categories: 1) whether the dataset is cancer or

not (`cancer` column), 2) whether the dataset is blood or not (`blood` column), 3) whether the dataset is cell line or not (`cell_line` column), 4) what is the origin of samples (`origin` column).

Sup.Table 4. Available RAVmodels

https://github.com/shbrief/GenomicSuperSignaturePaper/blob/master/inst/extdata/SupTable3_RAVmodels.csv

This is the list of currently available RAVmodels that are different in 1) the size of training datasets, 2) the number of top PCs collected from each study, 3) the number of clusters for hierarchical clustering, and 4) gene sets used for GSEA annotation. Two RAVmodels used in this work, RAVmodel_C2 and RAVmodel_PLIERpriors, are available for download using the `getModel` function and the others are available upon request.

Sup.Table 5. Summary of top 10 validated RAVs for 18 colorectal cancer datasets

https://github.com/shbrief/GenomicSuperSignaturePaper/blob/master/Results/CRC/outputs/CRC_top10_validated_ind.tsv

References

1. Myers, C. L., Barrett, D. R., Hibbs, M. A., Huttenhower, C. & Troyanskaya, O. G. Finding function: evaluation methods for functional genomic data. *BMC Genomics* **7**, 187 (2006).
2. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12**, 453–457 (2015).
3. Leek, J. T. *et al.* Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* **11**, 733–739 (2010).
4. Meng, C. *et al.* Dimension reduction techniques for the integrative analysis of multi-omics data. *Brief. Bioinform.* **17**, 628–641 (2016).
5. Stein-O'Brien, G. L. *et al.* Decomposing Cell Identity for Transfer Learning across Cellular Measurements, Platforms, Tissues, and Species. *Cell Syst* **8**, 395–411.e8 (2019).
6. Sharma, G., Colantuoni, C., Goff, L. A., Fertig, E. J. & Stein-O'Brien, G. projectR: An R/Bioconductor package for transfer learning via PCA, NMF, correlation, and clustering. *Bioinformatics* (2020) doi:10.1093/bioinformatics/btaa183.

7. Mao, W., Zaslavsky, E., Hartmann, B. M., Sealfon, S. C. & Chikina, M. Pathway-Level Information ExtractoR (PLIER) for gene expression data. doi:10.1101/116061.

8. Taroni, J. N. *et al.* MultiPLIER: A Transfer Learning Framework for Transcriptomics Reveals Systemic Features of Rare Disease. *Cell Syst* **8**, 380–394.e4 (2019).

9. Sastry, A. V. *et al.* Matrix factorization recovers consistent regulatory signals from disparate datasets. *bioRxiv* 2020.04.26.061978 (2020) doi:10.1101/2020.04.26.061978.

10. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods* vol. 16 1289–1296 (2019).

11. Lee, J., Oh, S. & Sael, L. GIFT: Guided and Interpretable Factorization for Tensors with an application to large-scale multi-platform cancer analysis. *Bioinformatics* **34**, 4151–4158 (2018).

12. Peng, M., Li, Y., Wamsley, B., Wei, Y. & Roeder, K. cFIT: Integration and transfer learning of single cell transcriptomes, illustrated by fetal brain cell development. *bioRxiv* (2020).

13. Pan, S. J. & Yang, Q. A Survey on Transfer Learning. *IEEE Trans. Knowl. Data Eng.* **22**, 1345–1359 (2010).

14. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).

15. refine.bio. *Refine.bio* https://www.refine.bio.

16. Mao, W., Zaslavsky, E., Hartmann, B. M., Sealfon, S. C. & Chikina, M. Pathway-level information extractor (PLIER) for gene expression data. *Nat. Methods* **16**, 607–610 (2019).

17. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

18. Ma, S. *et al.* Continuity of transcriptomes among colorectal cancer subtypes based on meta-analysis. *Genome Biol.* **19**, 142 (2018).

19. Guinney, J. *et al.* The consensus molecular subtypes of colorectal cancer. *Nat. Med.* **21**, 1350–1356 (2015).
20. Nojadeh, J. N., Behrouz Sharif, S. & Sakhinia, E. Microsatellite instability in colorectal cancer. *EXCLI J.* **17**, 159–168 (2018).
21. Ogino, S. & Goel, A. Molecular classification and correlates in colorectal cancer. *J. Mol. Diagn.* **10**, 13–27 (2008).
22. Banchereau, R., Hong, S., Cantarel, B. & Baldwin, N. Personalized immunomonitoring uncovers molecular networks that stratify lupus patients. *Cell* (2016).
23. Grayson, P. C. *et al.* Brief Report: Defining the Nasal Transcriptome in Granulomatosis With Polyangiitis (Wegener's). *Arthritis & Rheumatology* vol. 67 2233–2239 (2015).
24. Becht, E. *et al.* Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biol.* **17**, 218 (2016).
25. Brubaker, D. K. *et al.* An interspecies translation model implicates integrin signaling in infliximab-resistant inflammatory bowel disease. *Sci. Signal.* **13**, (2020).
26. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).
27. Soneson, C., Love, M. I. & Robinson, M. D. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Res.* **4**, 1521 (2015).
28. Pages, H., Carlson, M., Falcon, S. & Li, N. AnnotationDbi: annotation database interface. *R package version 1. 4* (2008).
29. Lee, A. J., Park, Y., Doing, G., Hogan, D. A. & Greene, C. S. Correcting for experiment-specific variability in expression compendia can remove underlying signals. doi:10.1101/2020.05.03.066597.
30. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R Package for Comparing

Biological Themes Among Gene Clusters. *OMICS* **16**, 284–287 (2012).

31. Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* **1**, 417–425 (2015).
32. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* vol. 102 15545–15550 (2005).
33. Baumann, N. How to use the medical subject headings (MeSH). *Int. J. Clin. Pract.* **70**, 171–174 (2016).
34. Mork, J., Aronson, A. & Demner-Fushman, D. 12 years on - Is the NLM medical text indexer still useful and relevant? *J. Biomed. Semantics* **8**, 8 (2017).
35. Geistlinger, L. *et al.* Toward a gold standard for benchmarking gene set enrichment analysis. *Brief. Bioinform.* (2020) doi:10.1093/bib/bbz158.
36. Parsana, P. & Riester, M. Waldron L. curatedCRCData: clinically annotated data for the colorectal Cancer transcriptome. Bioconductor.
37. Alvarez, M. J., Giorgi, F. & Califano, A. bcellViper, a data package required for the examples and vignette of the viper package. *Interactions* **172240**, 1 (2014).
38. McKinney, E. F., Lee, J. C., Jayne, D. R. W., Lyons, P. A. & Smith, K. G. C. T-cell exhaustion, co-stimulation and clinical outcome in autoimmunity and infection. *Nature* **523**, 612–616 (2015).
39. OmicIDX. http://omicidx.cancerdatasci.org/.