HHS Public Access

Author manuscript

Cell Rep. Author manuscript; available in PMC 2021 June 11.

Published in final edited form as:

Cell Rep. 2021 April 13; 35(2): 108975. doi:10.1016/j.celrep.2021.108975.

Merged Affinity Network Association Clustering: Joint multi-omic/clinical clustering to identify disease endotypes

Scott R. Tyler[1], **Yoojin Chun**[1], **Victoria M. Ribeiro**[1], **Galina Grishina**[2], **Alexander Grishin**[2], **Gabriel E. Hoffman**[1], **Anh N. Do**[1], **Supinda Bunyavanich**[1,2,3,*]

[1]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA

[2]Division of Allergy and Immunology, Department of Pediatrics, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA

[3]Lead contact

SUMMARY

Although clinical and laboratory data have long been used to guide medical practice, this information is rarely integrated with multi-omic data to identify endotypes. We present Merged Affinity Network Association Clustering (MANAclust), a coding-free, automated pipeline enabling integration of categorical and numeric data spanning clinical and multi-omic profiles for unsupervised clustering to identify disease subsets. Using simulations and real-world data from The Cancer Genome Atlas, we demonstrate that MANAclust's feature selection algorithms are accurate and outperform competitors. We also apply MANAclust to a clinically and multi-omically phenotyped asthma cohort. MANAclust identifies clinically and molecularly distinct clusters, including heterogeneous groups of "healthy controls" and viral and allergy-driven subsets of asthmatic subjects. We also find that subjects with similar clinical presentations have disparate molecular profiles, highlighting the need for additional testing to uncover asthma endotypes. This work facilitates data-driven personalized medicine through integration of clinical parameters with multi-omics. MANAclust is freely available at https://bitbucket.org/scottyler892/manaclust/src/master/.

Graphical abstract

*Correspondence: supinda@post.harvard.edu.

AUTHOR CONTRIBUTIONS

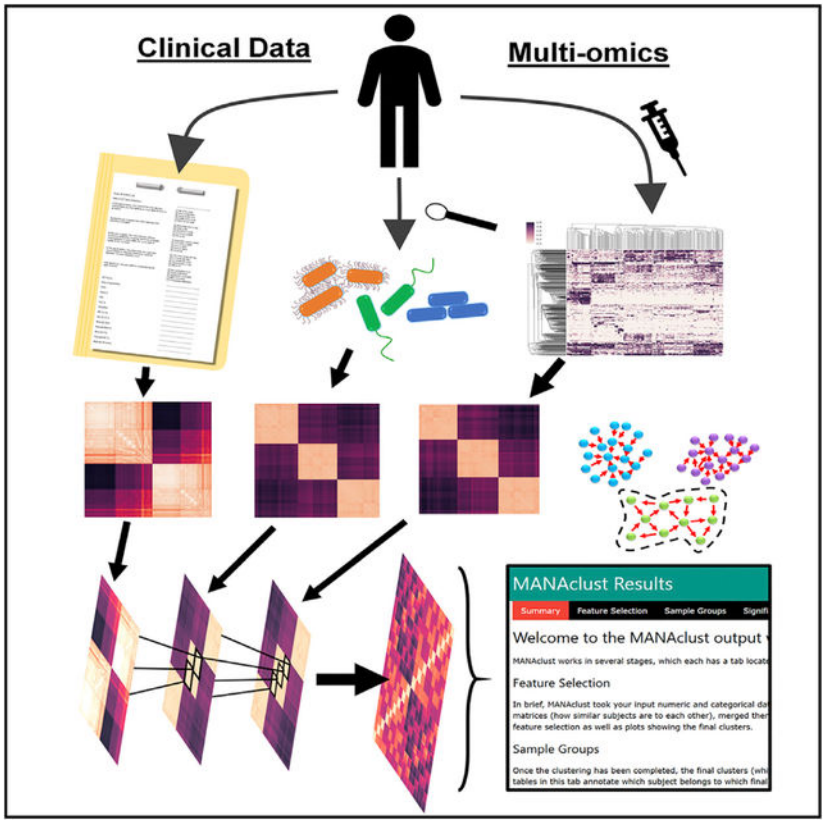
S.R.T. devised the MANAclust pipeline, performed analyses, drafted the manuscript, and developed the web applications. S.B. supervised the research, worked on the manuscript, and procured funding. S.B. and V.M.R. recruited the ARIA cohort and collected samples. S.B., G.G., and A.G. worked on sample processing and sequencing. Y.C., A.N.D., G.E.H., and S.B. processed the data and performed quality control and initial analyses. All authors critically reviewed and edited the manuscript.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.celrep.2021.108975.

DECLARATION OF INTERESTS

The authors declare no competing interests.



In brief

Clinical data commonly used in medical practice are underutilized in multi-omic analyses to identify disease endotypes. Tyler et al. present a python package called Merged Affinity Network Association Clustering (MANAclust) that automatically processes and integrates categorical and numeric data types, facilitating the inclusion of clinical data in multi-omic endotyping efforts.

INTRODUCTION

Complex diseases have multiple molecular etiologies yet result in concordant pathology stemming from multi-dimensional interactions between genetics and the environment (Tyler and Bunyavanich, 2019). Quantifying a single “ome” in a particular tissue will likely yield insufficiently granular details about a complex disease. To address these shortcomings, integrative multi-omics approaches have been developed (Nguyen etal., 2017; Rappoport and Shamir, 2018; Wang et al., 2014; Witten and Tibshirani, 2009), yet few allow for the additional integration of clinical data (e.g., from medical records, laboratory assays), despite the fact that clinical attributes are most often used to distinguish disease subtypes.

One complex disease with disparate underlying causes is asthma. Although airway obstruction and inflammation are shared characteristics of asthma, underlying mechanisms leading to such pathophysiology are non-uniform. A well-understood subtype of asthma is Type 2 asthma, in which genetic susceptibility and allergen exposure lead to skewing toward T-helper 2 (T_H2) cytokine-related inflammation, which upregulates airway inflammation and

airway hyperresponsiveness, causing asthma (Tyler and Bunyavanich, 2019). However, other mechanisms of asthma underlie other subtypes. Subtypes of disease with distinct pathophysiological mechanisms are called endotypes (Tyler and Bunyavanich, 2019).

Multi-omics can simultaneously characterize several molecular domains that may provide orthogonal information on mechanisms of a given endotype (Tyler and Bunyavanich, 2019). However, integrating heterogeneous clinical data with multi-omics presents several challenges, including differences in statistical properties and distributions, and for some clinical attributes, non-numeric features. Unsupervised clustering of categorical data is largely under-researched (Sreenivasulu et al., 2017), yet categorical clinical variables such as comorbidities and medication use are frequently used to subgroup patients for clinical diagnostics and treatment.

Selecting features of a dataset that add structure or patterning is an important step prior to unsupervised clustering. Inclusion of unimportant features adds noise, drowning out meaningful signal. Feature selection for numeric datasets is effective for reducing dimensionality and noise (Andrews and Hemberg, 2018), yet this approach has rarely been applied to categorical datasets for unsupervised clustering. Previous approaches to multi-omic unsupervised clustering often skip feature selection.

Although several algorithms exist for clustering categorical datasets (He et al., 2002; Sharma and Gaud, 2015), few exist to perform clustering on mixed categorical/numeric datasets (Ji et al., 2013), which are common to medicine. Integration of clinical data with multi-omics remains an outstanding challenge. There is need for algorithms designed for integration of multi-omics with clinical-categorical or mixed datasets that natively and automatically accommodate missing data fields and missing data types at the subject level, given how common missing data are in human research. Automated multi-omic approaches that perform unsupervised feature selection for both categorical and numeric datasets prior to clustering would advance the field.

Given the increasing volume of clinical data being generated and exponential availability of omics, we recognized a need to develop an approach that can integrate these rich sources of information via a user-friendly pipeline for identifying disease subtypes. We have created an open-source package called Merged Affinity Network Association Clustering (MANAclust) that: (1) takes numeric or non-numeric datasets as input, allowing for missing subject-level categorical data and missing data types; (2) performs information theory-based categorical feature selection and numeric feature selection; (3) calculates merged affinity network associations across subjects and datasets; (4) identifies final clusters (FCs) using multi-omics data, as well as consensus groups within each ome or dataset; (5) analyzes differences across groups; (6) determines which dataset's groups are significantly concordant with those of other datasets; and (7) creates a collated web display of all results (Figure 1). Using simulations and real-world data, we demonstrate that MANAclust's feature selection algorithms are highly accurate, and that MANAclust outperforms competitors in numeric multi-omics and categorical clustering. We then used MANAclust to identify endotypes in an asthma cohort of subjects who have been clinically phenotyped and undergone multi-omic profiling of their airways (Do et al., 2021), with the majority characterized in all data

dimensions, but with subjects variably missing some data types, as is frequently the case in real-world datasets. MANAclust identified 14 clinically and molecularly distinct clusters of subjects, including heterogeneous groups of “healthy controls,” and viral and allergy-related endotypes. In addition to several algorithmic innovations, these findings lay the groundwork for clinically and molecularly tailored personalized medicine for asthma.

RESULTS

MANAclust

To achieve joint clinical/multi-omics unsupervised clustering, we first created an algorithm for unsupervised feature selection of categorical data. To our knowledge, no automated solution exists for unsupervised categorical feature selection. For our goal—utilizing any type of clinical data (categorical or numeric), along with multi-omics for identifying disease subsets—we began from the premise that if categorical clinical variables are randomly distributed relative to one another, they will have lower relative mutual information compared with if they were non-random relative to one another. MANAclust quantifies the true amount of mutual information across variable pairs, comparing it with shuffled null versions of the dataset (Figures 2A–2C). We then calculate the difference between the observed pairwise feature-feature information in the real dataset compared with the shuffled datasets, calculating the information difference (Figures 2D and 2E). This enables selection of categorical variables that are non-independent and give structure to the dataset (Figure 2F). Additionally, we implemented a feature selection metric for traditional numeric omes (Figure S1; see STAR Methods for details).

To test these approaches, we created synthetic datasets in which subjects belonged to known groups with both categorical and numeric datasets containing features that were either designed to distinguish these groups or were purely random. Our categorical feature selection algorithm captured all real features while filtering out all random features, even with 50% missing data and 50% random noise added (Figures S2A–S2E). To benchmark numeric feature selection, we created synthetic datasets with 200 real features that discriminated groups and 10,000 features that were random noise. Our numeric feature selection algorithm captured an average of 95.8% of real features (i.e., recall), while consistently removing all random features (Figures S2F–S2J). As a classification problem (group-defining versus random features), our categorical feature selection algorithm had high accuracy, precision, recall, specificity, and F1 scores ($T = 54.2$, $P = 1.24e{-12}$, paired t tests relative to baseline; Figures S2A–S2E). Similarly, our feature selection algorithm for numeric datasets demonstrated high performance (P values: accuracy, 8.06e–19; precision, 7.85e–20; recall, 8.06e–19; specificity, 8.06e–19; F1 score 1.92e–20; paired t tests relative to baseline; Figures S2F–S2J). This benchmarking demonstrates that MANAclust’s feature selection algorithms accurately select the meaningful features in both categorical and numeric datasets to increase the signal-to-noise ratio of the inputs.

We hypothesized that these feature selection algorithms would allow for more discretely discernable clusters of subjects. With synthetic data to benchmark, using all features showed little distinguishable patterns in subject-to-subject distance (Figures S3A, S3C, and S3E). However, the feature selected datasets showed the expected number of groups that were

clearly discernable from one another (Figures S3B, S3D, and S3F). This enabled significantly more accurate clustering compared with using the whole datasets as measured by mutual information and cluster purity (Figures S3G and S3H; T = 84.1 and 215; P = 2.4e–14 and 5.12e–18, respectively; n = 10; t tests paired by dataset).

Comparison with other multi-omic clustering approaches using real-world and synthetic datasets

We aimed to benchmark MANAclust against several others designed for numeric multi-omics. We used a previously published benchmarking approach applied to 10 different cancer types from The Cancer Genome Atlas (TCGA) (Rappoport and Shamir, 2018) where performance was assessed by examining differences across clusters in clinical variables and differences in survival rates across clusters. We compared MANAclust with two generic clustering methods: K-means clustering (elbow rule for K-determination) (Lloyd, 1982), spectral clustering (elbow rule for K-determination) (Shi and Malik, 2000), and several multi-omics-specialized clustering approaches, including perturbation clustering for data integration and disease subtyping (PINS) (Nguyen et al., 2017), PINSPlus (Nguyen et al., 2019), Similarity Network Fusion (SNF) (Wang et al., 2014), and sparse multiple canonical correlation analysis (MCCA) (Witten and Tibshirani, 2009) (Figure 3). MANAclust showed the greatest significant differences between the identified clusters, both by difference in survival rate and the number of clinical variables that differed significantly across clusters when clusters were identified based on all multi-omics data (Figure 3A; Table S1). When operating only on the single-best ome per dataset, MANAclust consistently identified either the greatest number of differences between groups clinically or by survival time (Figure 3B). This demonstrates that even operating without missing or categorical data (which the other algorithms cannot accommodate), MANAclust still consistently performs better. Furthermore, the competing methods all require custom programming from the user, whereas MANAclust is an open-source, hyper-parameter-free method of mixed data-type multi-omics unsupervised clustering that accommodates missing data from a command-line interface and generates a web-display walk-through of results for users.

Although these results demonstrate that MANAclust provides good results for numeric datasets, we next sought to benchmark MANAclust in the purely categorical realm. We compared MANAclust with KModes (Huang, 1998) clustering with elbow rule based on within-cluster Hamming distances. Although both MANAclust and KModes clustering with elbow rule worked well in identifying the correct number of clusters, MANAclust significantly outperformed KModes in cluster purity and relative mutual information (F = 445.3, P = 1.39e–22 and F = 53.78, P = 8.75e–9, respectively; one-way ANOVA; Figure 3C). Taken together with the multi-omics benchmark, these results demonstrate that MANAclust works well in both the numeric and categorical domains.

MANAclust discovers clinically and molecularly distinguishable asthma endotypes

Having demonstrated the accuracy of MANAclust, we applied the pipeline to data for a complex human disease known to have multiple mechanistic causes: asthma (Tyler and Bunyavanich, 2019). Many individuals with asthma have Type 2 asthma, a subtype of asthma mediated by Type 2 T helper (Th2) cells or innate lymphoid cells (ILC2s) (Sugita et

al., 2018; Woodruff et al., 2009). Some common clinical characteristics of the Type 2 asthma endotype include: allergies, elevated levels of lung and/or peripheral eosinophils, total serum IgE, and allergen-specific IgE (sIgE) antibodies (Bunyavanich and Schadt, 2015; Liu et al., 2015; Nagasaki et al., 2017; Tyler and Bunyavanich, 2019). However, other less understood subtypes of asthma also exist (Liu et al., 2017; Ricciardolo et al., 2017; Wisniewski et al., 2018). To investigate potential mechanisms of asthma pathology in different subtypes, we applied MANAclust to data generated from the Airway in Asthma (ARIA) cohort (Do et al., 2021) of 316 subjects with asthma and healthy controls who underwent clinical phenotyping and profiling of airway transcriptome, microbiome, and/or methylome.

MANAclust identified 14 different clusters of asthma and health based on clinical and multi-omic data. Following identification of these 14 FCs, MANAclust also identified “consensus groups” for each input data type. Dataset-level consensus groups define which FCs are indistinguishable from one another within the given dataset. For example, two FCs could share a similar transcriptome but two distinct microbiomes; these FCs would have the same transcriptome consensus group but two different microbiome consensus groups, thus separating them into distinct FCs. Each of these FCs can be defined by the unique combination of consensus groups defined by MANAclust (Figures 4A–4F).

To test the efficacy of our approach for finding clinically meaningful differences between FCs, we tested whether the identified clusters differed significantly by treatment (which was excluded from the clinical attributes used for clustering). Indeed, asthma diagnosis ($P = 5.28e{-12}$, χ^2 with Benjamini-Hochberg correction), antibiotic usage ($P = 1.78e{-174}$, χ^2 with Benjamini-Hochberg correction), and oral steroid use ($P = 3.64e{-127}$, χ^2 with Benjamini-Hochberg correction) were significantly different across FCs, despite being held out from clustering. Note that the causal direction of these differences is unknown; for example, the treatments may cause changes in the transcriptome or vice versa. Of the clinical variables that were used for clustering, ranking among the most significant were measures related to atopy, including total serum IgE ($P = 1.61e{-195}$, χ^2 with Benjamini-Hochberg correction) and allergen-specific sensitization (defined by sIgE ≥ 0.10 kU_A/L) to cat and dog dander ($P = 1.70e{-16}$, $1.37e{-17}$, χ^2 with Benjamini-Hochberg correction), mouse urine ($P = 1.52e{-22}$, χ^2 with Benjamini-Hochberg correction), *Blatella germanica* cockroach ($P = 3.11e{-8}$, χ^2 with Benjamini-Hochberg correction), and others (Tables S2A and S2B). Along with total serum IgE levels, each of these specific allergens has previously been implicated in asthma control (Kanchongkittiphon et al., 2015; Rotsides et al., 2010). As we will discuss later, however, some subjects who present with nearly identical symptoms may respond differently to various medications. It is therefore critical to determine which subjects may present the same clinically yet have disparate molecular underpinnings. It is also important to fully understand the heterogeneity among “healthy controls.”

Control groups are highly diverse

There was not a single “control” group of healthy individuals. Instead, we observed five different FCs that corresponded to low prevalence of asthma and high asthma control test (ACT) scores (FCs 3, 7, 8, 9, and 13) (Figure 4F). This finding supports the long-held understanding that there is not one single way to be healthy. Considering such disparate, yet

healthy subjects as a single control group, rather than recognizing them as a collection of different subgroups, may obscure differences that might otherwise be detected if their subgroup memberships were recognized.

Understanding each molecular subtype

Although our main goal is to examine how each molecular data type fits with others and with clinical attributes, it is important to first develop a biologic understanding of each consensus group within each data type. To better understand the differences across consensus groups in the transcriptome and methylome, we performed comparisons of consensus group-enriched genes, co-expression/-methylation network graphs, module analyses, and predicted autocrine-paracrine signaling within each consensus group using PyMINEr (Tyler et al., 2019). In the transcriptome, we observed pathway-level differences across consensus groups relating to ciliated cells and replication (transcriptome consensus group 0) and leukocyte activation (transcriptome consensus group 1) (Figure 4A; Table S2C).

Of interest was the significant concordance between the transcriptome consensus group corresponding to leukocyte activation (transcriptome consensus group 1) and the methylome consensus group characterized by enrichment of the cell adhesion with leukocyte/lymphocyte activation (methylome consensus group 1) ($P \approx 0$; Figures 4B, S4A, and S4B; Table S2D). Also pertinent to asthma, we observed enrichment of T cell activation signaling by pathway analysis in methylome consensus group 0.

At the microbiome level, consensus groups differed primarily by the four most abundant genera (Figure 4C). One microbiome consensus group (2; Figure 4C) had exceptionally high levels of *Moraxella*, with marked reduction in *Corynebacteria* and other genera resulting in low alpha diversity (Figure S4D; one-way ANOVA: Chao1 $F = 52.57$, $P = 9.88e{-}34$; Shannon $F = 44.0$, $P = 3.35e{-}29$; microbiome consensus group 2 significance against all other groups $P \leq 0.01$; Tukey's Honestly Significant Difference (HSD) post hoc test).

Consensus groups within the clinical dataset differed largely based on allergy-related parameters. Additionally, although disease diagnosis and treatments were not included for clustering, these variables segregated across clinical consensus groups (Figure 4D). Type 2 asthma, the best described endotype of asthma, corresponded to clinical consensus groups 0 and 1, while the healthiest group with little asthma and very few allergies corresponded to consensus group 4. Interestingly, subjects in consensus group 5 harbored very few allergies but had the highest rates of asthma. Consensus group 5 had significantly more recent upper respiratory infections (URIs) compared with other groups (mode: URI within past month; $\chi^2 = 303.8$; $P = 2.69e{-}44$, Benjamini-Hochberg corrected; Table S2E). This fits with prior reports of viral infection being a well-characterized trigger of asthma exacerbations (Altman et al., 2019).

Simply characterizing each dataset's consensus group provides limited insights into the multi-factorial nature of a disease, however; multi-omics enables us to observe how these single datasets weave together to form a holistic picture of an individual's pathobiologic type.

Consensus groups combine to define endotypes

Each FC is derived from the combination of consensus groups from each of the input datasets (Figures 4E and 4F). Prior attempts to derive endotypes of disease frequently used only a single data type (Howrylak et al., 2016; Kuo et al., 2017; Liang et al., 2015; Nicodemus-Johnson et al., 2016). A major outstanding question is whether groups identified by one data type are redundant or orthogonal to those identified by another data type. To address this question, MANAclust calculates (1) the most frequent chi-square statistics to determine if consensus groups across data types provide completely orthogonal or partly concordant groups (Figures S4A–S4C), and (2) the Bayesian empirical probability that subjects belong to each consensus group of other datasets given that they were members of a consensus group from a different data type (Figure 4G). Although these Bayesian probabilities have been argued to be related to causality (Williamson, 2009), external intervention in the form of clinical trials would be needed to verify causal relationships (Pearl, 2009).

We observed that some consensus groups did indeed carry redundant information across data types. However, other consensus groups carried orthogonal information that was not concordant with consensus groups of other input data types (Figures 4G and S4A–S4C). For example, clinical consensus group 4, with the lowest prevalence of asthma and allergy markers, was significantly concordant with consensus groups from each other data type (Figure 4G), including microbiome consensus group 0, with highest abundance of *Corynebacterium*, which was previously associated with health and protection from asthma (Zhou et al., 2019). This combination of consensus groups creates FC 8, with the lowest percentage of asthma and highest ACT score (Figure 4F).

Another consensus group that showed strong concordance across data types was the microbiome group with extremely low alpha diversity (Figure S4D) dominated by *Moraxella* (Figure 4G). This group was strongly associated with the methylome profile enriched for lymphocyte and leukocyte adhesion and activation (methylome consensus 1). This methylome profile was further associated with a transcriptional profile related to the immune system and leukocyte activation (transcriptome consensus 1; Figure 4A; Table S2D).

Clinically indistinguishable subjects harbor disparate molecular drivers of asthma

There were many cases in which consensus group membership segregated non-randomly across datasets, indicating non-orthogonality across these datasets, as represented by all edges in Figure 4G. However, two clinical consensus groups shared no significant segregation across different datasets, clinical consensus groups 1 and 3. This indicates that two subjects may present with similar clinical parameters yet have highly divergent underlying molecular signatures. This scenario suggests potential reasons for trial-and-error aspects of some asthma treatment (Tesse et al., 2018).

Lastly, although we have focused on selected FCs and consensus groups, several other FCs remain that were mixed between asthma and control groups. This may point toward well-controlled asthma that phenocopies controls, or that sampling additional tissue such as

bronchial epithelium closer to the site of pathology may provide orthogonal measures that could refine these subgroups more clearly.

Integration of clinical and numeric data improves endotype identification

To assess the impact of the integration of clinical and numeric data in the ARIA cohort, we compared the significance of our full analysis with all data to performances using only the clinical data, all numeric data, or each numeric data type alone. We quantified the maximum significance by −log10(P value) of each clustering run's held-out clinical variables based on either FCs or consensus groups with the noted inputs. Use of all data provided the greatest significance in held-out clinical variables (Figure 4H). This was also true when normalizing for each hold-out variable's lowest level of significance (Figure 4I).

DISCUSSION

We have created an open-source tool called MANAclust that enables automated integration of clinical datasets and omics for unsupervised clustering. The categorical and numeric feature selection algorithms implemented in MANAclust are highly accurate, enabling successful stratification of clinically meaningful groups using real-world data from both TCGA and the ARIA cohort. We used MANAclust to uncover the heterogeneity of "healthy controls," as well as several endotypes of asthma, including those driven by URIs, allergies, and a subset with *Moraxella*-dominated airway microbiomes. We also observed two sets of asthma subjects for whom underlying molecular characteristics were statistically independent from their clinical presentation. MANAclust paves the road to data-driven personalized medicine through the identification of subject clusters not only by multi-omics but also by incorporating clinical parameters that have long been used independently in diagnostics and care.

STAR★METHODS

RESOURCE AVAILABILITY

Lead contact—Further information and requests for resources should be directed to Dr. Supinda Bunyavanich (supinda@post.harvard.edu).

Materials availability—This study did not generate new unique reagents other than those reported in Data and Code Availability.

Data and code availability—TCGA benchmarking datasets are available at Synapse (https://www.synapse.org; https://doi.org/10.7303/syn21301852). Subject-level data for ARIA participants have not been made publicly available because subjects did not consent to public release of their data. Data to generate figures and tables are available from the corresponding author with the appropriate permission from the study team and investigators upon reasonable request and institutional review board approval. All code used for this project is available at the public repository: https://bitbucket.org/scottyler892/manaclust/src/master/.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Human Subjects—All subjects from the AiRway in Asthma (ARIA) study (Do et al., 2021) provided written informed consent for study participation and the study was approved by the Mount Sinai Health System Internal Review Board (IRB). Subjects were recruited at the Icahn School of Medicine at Mount Sinai for phenotyping and nasal sampling. Subjects with severe asthma were diagnosed as having severe persistent asthma by a pulmonologist based on Severe Asthma Research Program criteria (Moore et al., 2007) Subjects with mild/moderate persistent asthma were diagnosed based on ≥ 2 symptoms per week, use of an inhaled bronchodilator ≥ twice weekly or use of daily medication for asthma, and increased airway responsiveness to methacholine (PC20 ≤ 12.5 mg/ml). Non-asthmatic controls had no personal or family history of asthma in first-degree relatives, normal spirometry, and no bronchodilator response). Phenotyping for all subjects included detailed questionnaires on asthma-related symptoms, medication use, and spirometry following American Thoracic Society guidelines (National Asthma Education and Prevention Program, 2007). The ARIA subjects studied comprised individuals age 4 to 60 years, including 158 females and 158 males.

METHOD DETAILS

Handling numeric datasets in MANAclust

Numeric dataset input: Numeric datasets are fed into MANAclust as tab delimited files in which the samples, or subject IDs are organized in columns and the features are organized in rows.

Numeric feature selection: Numeric features to be used for clustering are selected using an anti-correlation-based metric. The first step of this process is determining what qualifies as a ‘significant’ negative correlation. Because each dataset has unique data distribution characteristics with variable power, we use an empirically determined false positive rate generated by a null distribution. This null distribution of Spearman correlations is created through bootstrap shuffling of all values within a gene (or numeric attribute); this unpairs all X-Y relationships in the original dataset, creating an empirical null dataset (Figure S1A). The Spearman correlations are then measured across variables (Figure S1B). Next, the negative correlations are subset from this data, and this vector is concatenated to its own negative, creating a Gaussian. The cutoff for significance is then determined by taking the mean minus 4 standard deviations (Figure S1C). The empirical false positive rate is then measured; because these values were created from a bootstrap shuffled version of the dataset, all measures over the cutoff are false positives. Next the Spearman correlation matrix is calculated on the real, unshuffled dataset. Next, for each variable, the total number of observed negative correlations is calculated (O_t) (Figure S1D). For each feature, the number of negative correlations is quantified, and the expected number of observations less than the negative Rho cutoff (E_{sig}) is calculated from the empiric false positive rate (Figure S1E). A feature is considered significantly anti-correlated if it has more significant anti-correlations than expected by random chance (≥Mult * E_{sig}; Figure S1F). The power in detecting these correlations is dependent on the number of samples; we therefore created a

dynamic cutoff scaling with the log of the number of samples. This provides a cutoff for variable inclusion or exclusion for downstream clustering (Figures S1G and S1H).

Handling categorical datasets in MANAclust

Categorical dataset input: Categorical datasets should be in fed into MANAclust in an m × n matrix where (m) subjects are in rows, while the (n) categorical dataset variables are in columns. The missing values of all variables must be a standard string that is fed into the argument '-missing_data'. In the case of our dataset, we used the 'N/A' string. Note that per-variable missing values only pertain to categorical datasets. It is presumed that individual variables will not be missing from numeric datasets; for example, a transcriptomic dataset would have measures for all genes rather than only for some, and missing values for other genes. In all cases however, MANAclust natively handles subject-level missing datasets regardless of numeric/categorical status. If a subject has missing data for all variables within a dataset, it is treated as if the subject is removed from the dataset entirely given that there are no measures.

Numeric values contained within categorical datasets: If MANAclust finds ≥ 10 numeric values contained within a categorical dataset, these are digitized linearly into bins. The number of bins scales with the number of observed values as follows:

$$bins = max\left\{2, \left\lfloor \frac{\sqrt{\mathrm{n}}}{5} \right\rfloor\right\}$$

Where n = the number of observations for the given variable. In this manner, numeric attributes contained within categorical data are turned into categorical class labels. This enables a mutual information style integration of numeric values alongside categorical values. This equation was used, because to provide meaningful information, at least two bins are necessary, and we sought to scale with $\sqrt{n}$, in a manner similar to variance. Hoping to capture at least 5 samples in each $\sqrt{n}$ scaled bin, we divided by 5, and applied the floor operator.

Categorical label encoding: Categorical variables are one-hot-encoded into binary matrices representing the category membership using the sklearn functions LabelEncoder and OneHotEncoder (Pedregosa et al., 2011). In the case that a subject is missing this feature, no entry is added to this matrix, leaving an empty (zero filled) matrix in its stead.

Categorical feature selection: For feature selection, the mutual information is calculated for both the real and bootstrap shuffled versions of the dataset (Figure 2A). This creates the symmetric mutual information matrix comparing all features to each other for both the real dataset and the randomized bootstrap shuffled datasets (Figures 2B and 2C).

The mutual information matrix from the real dataset is then subtracted from the mutual information calculated from shuffled datasets, creating **R**, an n × n × i matrix, where n = the number of features in the categorical dataset, while i = the number of iterations used for bootstrap shuffling (Figure 2C). The default for the number of iterations is 10. This matrix is filled with the measured mutual information of all feature-feature pairs in the randomly

shuffled datasets. Next the difference between the mutual information in the real dataset and shuffled versions are calculated for each feature-feature pair, creating the information difference matrix **IDM** (n × n × i). Next, the minimum of **IDM** along the third axis (i) is taken to generate the minimum difference matrix **MDM** (n × n). For each feature-feature pair of this matrix, the entry in **MDM** represents the iteration of the randomized matrices, for which the difference between the real dataset and the randomized dataset was the least. In other words, given all of the random iterations – what was the worst-case scenario for this feature-feature pair compared to the randomized versions. The **MDM** matrix is then flattened, using the maximum function along one of the axes of this symmetric matrix, yielding a one-dimensional vector *MMD* (Maximum of Minimum Difference). In *MMD*, each feature's best feature-feature pair from **MDM** is taken, meaning – for each feature, what was the pair in which there was the most information relative to the randomized datasets. Then, a cutoff is applied to include features that have enough information in their best feature-feature pair. The features that pass this metric are then included in the final feature selected dataset for categorical datasets.

Affinity matrix calculation with categorical datasets: First, the average log-loss/ categorical cross entropy comparing all subjects to one another is calculated using the one-hot-encodings described above, including only the variables for which both subjects have non-missing values. The log-loss function is implemented in mana_clust.mana_cat.log_loss, but were computed using the scikit-learn sklearn.metrics.log_loss function (Pedregosa et al., 2011). Next, the negative squared Euclidean distance (affinity matrix) is calculated on this log-loss matrix. In the event that a subject-subject pair has no mutual measures for a categorical dataset, this subject-subject affinity is filled in by the median of all other affinities for this dataset. To prevent invariant measures in the affinity matrix, we perform a smoothing process in which the affinity matrix has Gaussian noise added with mean equivalent to the minimum of the zero masked affinity matrix.

To account for variable distributions in affinity matrices, the affinity matrix is linear normalized to be between 0 and −100 where the sample pairs that are the most different from each other map to −100 and those that are most similar map to 0. Note that self-similarities are not considered in the normalization, thus only the most similar two samples are normalized to 0 rather than the self-comparison diagonal of the affinity matrix mapping to 0. This normalizes the scale of all affinity matrices, enabling their combination with equal weight given to each dataset.

Combining all categorical and numeric affinity matrices in MANAclust—Once all affinity matrices are calculated, they are concatenated along a third axis, followed by taking the missing-value-compatible mean. This process evaluates the average affinity across all omes. If two subjects do not have any overlap in the datasets that they have in common, this location in the final merged affinity matrix is filled in with the median of affinities for all subjects that do have at least one shared dataset.

Clustering in MANAclust—The combined affinity matrix between omes is first fed into the affinity propagation algorithm (Frey and Dueck, 2007). However, under some circumstances, the affinity propagation algorithm may fail to converge on appropriate

exemplars (points central to a cluster, similar to medoids), resulting in singlet clusters. To account for this, we employ a secondary step following affinity propagation in which the affinities from all singlet clusters are subset, and re-clustered using Louvain modularity of a network built from these affinities as follows: A network is created from the subset affinity matrix. The weight of all edges with less affinity than the median of the full affinity matrix is set to zero. The remaining edges are built from inverse negative affinities from the affinity matrices described above; weighted Louvain modularity is then performed on the resultant network.

Consensus group identification in MANAclust—To define consensus groups, MANAclust takes the members of each of the final clusters and compares their calculated affinities for each dataset with the goal of determining whether two final clusters are derived from the same distribution of affinities within each input dataset. This is done by sub-setting each affinity matrix of an input dataset for each final cluster pair (we will use A and B for illustrative reasons). Next, the within dataset affinities for subject-subject pairs within FC A and within FC B are taken; these are concatenated into a single vector. Next, the affinities across these two FCs are taken (FC A to FC B affinities and reverse); these are then flattened into a single vector. These two vectors are compared to each other by t test. All pairwise comparisons of each FC to all other FCs are performed, creating a symmetric probability matrix that represents the probability that each FC-FC pair came from the same distribution. These P values are then Benjamini-Hochberg corrected for multiple comparisons. Lastly, these probabilities are modularized through Louvain community detection. The communities identified are the consensus groups for each dataset.

Synthetic datasets for benchmarking MANAclust—To test the efficacy of MANAclust, we created synthetic datasets that contained one categorical dataset (two groups), and two numeric datasets (three groups each). Each dataset's individual clusters were created independently from one another in which each dataset provided orthogonal information; this yielded 2*3*3 = 18 FCs that were unique from each other. Categorical datasets harbored 15 of the 165 features of this dataset discriminated the two groups from each while all others were variables that subjects were randomly assigned a category. 50% of the entirety of these datasets was also assigned to a random category to introduce noise to the real variables. Lastly, 50% of this dataset was masked with missing values. Numeric datasets contained a total of 10,200 features, 200 of which were real features that discriminated groups while the remaining variables were random, drawing from a Gaussian distribution. Source code for dataset generation and comparisons is contained in the simulate_datasets.py and run_simulation_study.py files of the MANAclust distribution package.

For benchmarking purely categorical data clustering, we compared MANAclust to the KModes (as implemented in the python 'kmodes' pip package) using the elbow rule based on sum of the within-group Hamming distances as implemented in SciPy's scipy.spatial.distance.hamming function. In this implementation of the elbow rule, we used the greatest value from the second derivative of the sum of within-group distances versus group number curve. 50 real features, and 100 random features were simulated to create 5,

10, 15, and 20 groups from a total of 1000 simulated subjects, with 50% noise added to the data. Each scenario was simulated 5 times. Significance for differences in method purity and relative mutual information was determined by the stats.f_oneway function from the SciPy package.

The Cancer Genome Atlas (TCGA) benchmarking of MANAclust—Our code was adapted from a previously published pipeline with several minor differences (Rappoport and Shamir, 2018). There were several algorithms that were not included in our analysis either because they required a MATLAB license (MultiNMF), were not open source (LRAcluster and rMKL-LP), or took prohibitively long to run (iClusterBayes). We allowed each algorithm to perform clustering (and any embedded dimension reduction or feature selection within each method).

Clinical Datasets

Allergen sensitization: Allergen sensitization as reported in Figure 4D was calculated as the percent of allergens (that were tested) with specific sIgE levels (sIgE) ≥ 0.10 kU_A/L. Specific sIgE levels were measured for the following allergens: *Dermatophagoides pteronyssinus*, *Dermatophagoides farinae*, cat dander, dog dander, mouse, *Blatella germanica* (cockroach), grass mix, mold yeast mix, tree pollen mix, weed pollen mix (Pham et al., 2019). For the 2 subjects who did not have all ten sIgE levels measured, the percent of the tested allergens that were above the 0.10 threshold are reported in Figure 4D.

Asthma percentage in consensus groups: The % Asthma reported in Figures 4D and 4F was the percentage of subjects from the respective cluster or consensus group who answered yes to “Has a doctor ever diagnosed you with asthma”?. This information was not used for clustering.

Nasal RNaseq processing and consensus group analysis—Nasal samples were collected using a cytology brush and placed in RNAlater (Do et al., 2021). RNA was isolated using the QIAGEN RNeasy mini kit. Libraries were prepared with the TruSeq RNA Sample Prep Kit v2. Samples were then sequenced on the Illumina HiSeq 2500 platform. Samples were aligned to the Ensembl human genome version 38 using the Salmon 0.13.1 aligner and sum collated to the gene level (Patro et al., 2017; Zerbino et al., 2018). For all analyses, log2(TPM+1) units were used.

Methylation data processing and consensus group analysis—Nasal samples were collected from subjects and DNA was isolated using the QIAGEN DNeasy Mini Kit. Methylation was quantified using the Illumina Infinium MethylationEPIC v1.0 array with processing of the methylation data using the ENmix package (version 1.18.1) in R version 3.5.1. Two samples were excluded due to with low quality CpG calls (> 10% of loci with CpG detection P value > 0.01). CpGs on sex chromosomes were also excluded for clustering. Low-quality CpGs, as defined by those with low confidence calls in > 10% of samples, or CpGs with distinctly multi-modal distributions were removed. CpGs with known SNPs were also removed, as array hybridization could be affected by subject-specific mismatches at these loci. Samples were quantile normalized with ENmix (Xu et al., 2016).

For clustering, percent methylated values were used, while for pathway analyses, percent unmethylated was used to provide a presumed correlate with gene expression. For analysis with MANAclust, overdispersed loci were used as defined by PyMINEr,(Tyler et al., 2019) using a Z-score of > = 2 from the residual from the Lowess fit of mean to variance relationship.

For analysis of consensus groups with PyMINEr, only the subset of loci with “promoter associated” annotations were used. This matrix was then converted to percent unmethylated by subtracting percent methylated from 1; this was done to provide a correlate with presumed gene expression where less promoter methylation thought to track with higher transcription. Lastly, this matrix was filtered for loci mapping to a gene symbol, then the average of each gene symbol was taken.

Microbiome processing and consensus group analysis—The V3V4 16S amplicons were amplified from nasal DNA samples using the Illumina “16S Metagenomic Sequencing Library Preparation” protocol (Illumina, 2013). Library prep was performed using the Nextera XT Index Kit v2. Illumina MiSeq platform with 2×250bp paired-end reads were used for sequencing. SEQTK 1.2 was used to trim primer regions and low quality bases from the ends of reads. Reads were demultiplexed using Qiime version 1.9.1 and OTUs were picked with similarity 97% using the Greengenes database version 13.8 reference (Caporaso et al., 2010). To account for differences in read-depth, samples were rarefied to 2048 counts total for each sample, thus normalizing for variance in total depth across samples (Chun et al., 2020; Weiss et al., 2017). This OTU matrix was then log2 transformed prior to analysis: log2(OTU+1).

MicrobiomeAnalyst was used to analyze MANAclust’s defined consensus groups for differential abundance and alpha diversity measures as indicated in Figure 4C (Dhariwal et al., 2017). Data shown in Figure 4D represents genus level quantifications to show a high-level view of the microbiome, while the input dataset for clustering was OTU level quantifications (Dhariwal et al., 2017). Filtered Chao1 and Shannon alpha diversity quantifications (as seen in Figure S4D) were performed on filtered and normalized OTU quantifications. Filtering and normalization was performed as follows: Data were filtered for abundance for those OTUs that had a minimum total count of 4, or less than 10% total abundance in all samples using the ApplyAbundanceFilter function. The OTUs in the lowest 5% for variance based on the inner-quartile range were then removed using the ApplyVarianceFilter function. Lastly, the relative log expression normalization was performed PerformNormalization function in MicrobiomeAnalyst. The two most diverse, two middle, and least diverse consensus groups are annotated as such in Figure 4C.

PyMINEr analysis of methylome and transcriptome consensus groups—For transcriptome and methylome analysis, we used PyMINEr’s -manual_sample_groups argument feeding in the consensus group IDs that were identified by MANAclust (Tyler et al., 2019).

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical differences between groups with categorical datasets—MANAclust first determines global statistical differences using chi-square tests; for those that were significant after FDR correction, pairwise post-hocs are performed. Chi-square analyses are performed to assess for non-independent segregation of categorical variables across the final clusters or consensus groups. Global statistics were FDR corrected using the Benjamini-Hochberg correction. Chi-square analyses were then performed as post-hocs assessing each pair of groups for non-independent segregation across groups when the global statistic was significant (alpha = 0.001 for FDR q-values). Chi-square analyses are performed using the stats.chi2_contingency function in SciPy. To calculate adjusted residuals and convert these to cell-wise P values for the contingency table, we implemented a previously published approach that had not been previously implemented in python (García-pérez and Núñez-antón, 2003). The approach is implemented in MANAclust's mana_annotate_results.cont_to_p_independent function. MANAclust does not automatically check if each contingency table meets the assumption of n = 5 for each cell, as is assumed by χ^2 test-of-independence analyses.

Statistical differences between groups with numeric datasets—Following final cluster and consensus group identification, global statistical differences for each variable in the input datasets are assessed by 1-way ANOVA comparing all groups to each other. P values are FDR adjusted using the Benjamini-Hochberg method. Pairwise differences between all groups for all variables is performed by first filtering only for significant differences by adjusted 1-way ANOVA q-values, for those features that are significantly different globally, t tests are performed pairwise between all final clusters and consensus groups. These are also Benjamini-Hochberg corrected. ANOVAs are performed using the function stats.f_oneway from the SciPy package. t tests are performed using the stats.ttest_ind function from SciPy. MANAclust does not automatically check for within-group normality, as is assumed by these tests.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

This study was funded by the National Institutes of Health R01AI118833, U19AI136053, R01AI147028, T32CA078207, and K99HG011270.

REFERENCES

Altman MC, Gill MA, Whalen E, Babineau DC, Shao B, Liu AH, Jepson B, Gruchalla RS, O'Connor GT, Pongracic JA, et al. (2019). Transcriptome networks identify mechanisms of viral and nonviral asthma exacerbations in children. Nat. Immunol 20, 637–651. [PubMed: 30962590]

Andrews TS, and Hemberg M (2018). Identifying cell populations with scRNASeq. Mol. Aspects Med 59, 114–122. [PubMed: 28712804]

Bunyavanich S, and Schadt EE (2015). Systems biology of asthma and allergic diseases: a multiscale approach. J. Allergy Clin. Immunol 135, 31–42. [PubMed: 25468194]

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI, et al. (2010). QIIME allows analysis of high-throughput community sequencing data. Nat. Methods 7, 335–336. [PubMed: 20383131]

Chun Y, Do A, Grishina G, Grishin A, Fang G, Rose S, Spencer C, Vicencio A, Schadt E, and Bunyavanich S (2020). Integrative study of the upper and lower airway microbiome and transcriptome in asthma. JCI Insight 5, e133707.

Dhariwal A, Chong J, Habib S, King IL, Agellon LB, and Xia J (2017). MicrobiomeAnalyst: a web-based tool for comprehensive statistical, visual and meta-analysis of microbiome data. Nucleic Acids Res. 45 (W1), W180–W188. [PubMed: 28449106]

Do AN, Chun Y, Grishina G, Grishin A, Rogers AJ, Raby BA, Weiss ST, Vicencio A, Schadt EE, and Bunyavanich S (2021). Network study of nasal transcriptome profiles reveals master regulator genes of asthma. J. Allergy Clin. Immunol 147, 879–893. [PubMed: 32828590]

Frey BJ, and Dueck D (2007). Clustering by passing messages between data points. Science 315, 972–976. [PubMed: 17218491]

García-pérez MA, and Núñez-antón V (2003). Cellwise Residual Analysis in Two-Way Contingency Tables. Educ. Psychol. Meas 63, 825–839.

He Z, Xu X, and Deng S (2002). Squeezer: An efficient algorithm for clustering categorical data. J. Comput. Sci. Technol 17, 611–624.

Howrylak JA, Moll M, Weiss ST, Raby BA, Wu W, and Xing EP (2016). Gene expression profiling of asthma phenotypes demonstrates molecular signatures of atopy and asthma control. J. Allergy Clin. Immunol 137, 1390–1397.e6. [PubMed: 26792209]

Huang Z (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Min. Knowl. Discov 2, 283–304.

Illumina (2013). 16S Metagenomic Sequencing Library Preparation. 10.1016/emea.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/16s/16s-metagenomic-library-prep-guide-15044223-b.pdf.

Ji J, Bai T, Zhou C, Ma C, and Wang Z (2013). An improved k-prototypes clustering algorithm for mixed numeric and categorical data. Neurocomputing 120, 590–596.

Kanchongkittiphon W, Mendell MJ, Gaffin JM, Wang G, and Phipatanakul W (2015). Indoor environmental exposures and exacerbation of asthma: an update to the 2000 review by the Institute of Medicine. Environ. Health Perspect 123, 6–20. [PubMed: 25303775]

Kuo CS, Pavlidis S, Loza M, Baribaud F, Rowe A, Pandis I, Sousa A, Corfield J, Djukanovic R, Lutter R, et al.; U-BIOPRED Study Group (2017). T-helper cell Type 2 (Th2) and non-Th2 molecular phenotypes of asthma using sputum transcriptomics in U-BIOPRED. Eur. Respir. J 49, 1602135. [PubMed: 28179442]

Liang L, Willis-Owen SAG, Laprise C, Wong KCC, Davies GA, Hudson TJ, Binia A, Hopkin JM, Yang IV, Grundberg E, et al. (2015). An epigenome-wide association study of total serum immunoglobulin E concentration. Nature 520, 670–674. [PubMed: 25707804]

Liu T, Wu J, Zhao J, Wang J, Zhang Y, Liu L, Cao L, Liu Y, and Dong L (2015). Type 2 innate lymphoid cells: A novel biomarker of eosinophilic airway inflammation in patients with mild to moderate asthma. Respir. Med 109, 1391–1396. [PubMed: 26459159]

Liu W, Liu S, Verma M, Zafar I, Good JT, Rollins D, Groshong S, Gorska MM, Martin RJ, and Alam R (2017). Mechanism of TH2/TH17-pre-dominant and neutrophilic TH2/TH17-low subtypes of asthma. J. Allergy Clin. Immunol 139, 1548–1558.e4. [PubMed: 27702673]

Lloyd S (1982). Least squares quantization in PCM. IEEE Trans. Inf. Theory 28, 129–137.

Moore WC, Bleecker ER, Curran-Everett D, Erzurum SC, Ameredes BT, Bacharier L, Calhoun WJ, Castro M, Chung KF, Clark MP, et al.; National Heart, Lung, Blood Institute's Severe Asthma Research Program (2007). Characterization of the severe asthma phenotype by the National Heart, Lung, and Blood Institute's Severe Asthma Research Program. J. Allergy Clin. Immunol 119, 405–413. [PubMed: 17291857]

Nagasaki T, Matsumoto H, and Izuhara K; KiHAC Respiratory Medicine Group (2017). Utility of serum periostin in combination with exhaled nitric oxide in the management of asthma. Allergol. Int 66, 404–410. [PubMed: 28256388]

National Asthma Education and Prevention Program (2007). Expert Panel Report 3: Guidelines for the Diagnosis and Management of Asthma. Clinical Practice Guidelines (National Heart, Lung, and Blood Institute).

Nguyen T, Tagett R, Diaz D, and Draghici S (2017). A novel approach for data integration and disease subtyping. Genome Res. 27, 2025–2039. [PubMed: 29066617]

Nguyen H, Shrestha S, Draghici S, and Nguyen T (2019). PINSPlus: a tool for tumor subtype discovery in integrated genomic data. Bioinformatics 35, 2843–2846. [PubMed: 30590381]

Nicodemus-Johnson J, Myers RA, Sakabe NJ, Sobreira DR, Hogarth DK, Naureckas ET, Sperling AI, Solway J, White SR, Nobrega MA, et al. (2016). DNA methylation in lung cells is associated with asthma endotypes and genetic risk. JCI Insight 1, e90151. [PubMed: 27942592]

Patro R, Duggal G, Love MI, Irizarry RA, and Kingsford C (2017). Salmon provides fast and bias-aware quantification of transcript expression. Nat. Methods 14, 417–419. [PubMed: 28263959]

Pearl J (2009). Causal inference in statistics: An overview. Stat. Surv 3, 96–146.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, and Dubourg V (2011). Scikit-learn: Machine learning in Python. J. Mach. Learn. Res 12, 2825–2830.

Pham MN, Andrade J, Mishoe M, Chun Y, and Bunyavanich S (2019). Perceived Versus Actual Aeroallergen Sensitization in Urban Children. J. Allergy Clin. Immunol. Pract 7, 1591–1598.e4. [PubMed: 30654198]

Rappoport N, and Shamir R (2018). Multi-omic and multi-view clustering algorithms: review and cancer benchmark. Nucleic Acids Res. 46, 10546–10562. [PubMed: 30295871]

Reimand J, Kull M, Peterson H, Hansen J, and Vilo J (2007). g:Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments. Nucleic Acids Res. 35, W193–W200. [PubMed: 17478515]

Ricciardolo FLM, Sorbello V, Folino A, Gallo F, Massaglia GM, Favatà G, Conticello S, Vallese D, Gani F, Malerba M, et al. (2017). Identification of IL-17F/frequent exacerbator endotype in asthma. J. Allergy Clin. Immunol 140, 395–406. [PubMed: 27931975]

Rotsides DZ, Goldstein IF, Canfield SM, Perzanowski M, Mellins RB, Hoepner L, Ashby-Thompson M, and Jacobson JS (2010). Asthma, allergy, and IgE levels in NYC head start children. Respir. Med 104, 345–355. [PubMed: 19913396]

Sharma N, and Gaud N (2015). K-modes Clustering Algorithm for Categorical Data. Int. J. Comput. Appl 127, 46.

Shi J, and Malik J (2000). Normalized cuts and image segmentation. IEEE Trans. Pattern Anal. Mach. Intell 22, 888–905.

Sreenivasulu G, Raju SV, and Rao NS (2017). Review of Clustering Techniques (Springer Singapore).

Sugita K, Steer CA, Martinez-Gonzalez I, Altunbulakli C, Morita H, Castro-Giner F, Kubo T, Wawrzyniak P, Rückert B, Sudo K, et al. (2018). Type 2 innate lymphoid cells disrupt bronchial epithelial barrier integrity by targeting tight junctions through IL-13 in asthmatic patients. J. Allergy Clin. Immunol 141, 300–310.e11. [PubMed: 28392332]

Tesse R, Borrelli G, Mongelli G, Mastrorilli V, and Cardinale F (2018). Treating Pediatric Asthma According Guidelines. Front Pediatr. 6, 234. [PubMed: 30191146]

Tyler SR, and Bunyavanich S (2019). Leveraging -omics for asthma endotyping. J. Allergy Clin. Immunol 144, 13–23. [PubMed: 31277743]

Tyler SR, Rotti PG, Sun X, Yi Y, Xie W, Winter MC, Flamme-Wiese MJ, Tucker BA, Mullins RF, Norris AW, et al. (2019). PyMINEr Finds Gene and Autocrine-Paracrine Networks from Human Islet scRNA-Seq. Cell Rep. 26, 1951–1964.e8. [PubMed: 30759402]

Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, Haibe-Kains B, and Goldenberg A (2014). Similarity network fusion for aggregating data types on a genomic scale. Nat. Methods 11, 333–337. [PubMed: 24464287]

Weiss S, Xu ZZ, Peddada S, Amir A, Bittinger K, Gonzalez A, Lozupone C, Zaneveld JR, Vázquez-Baeza Y, Birmingham A, et al. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. Microbiome 5, 27. [PubMed: 28253908]

Williamson J (2009). Probabilistic theories of causality. In The Oxford Handbook of Causation, Beebee H, Hitchcock C, and Menzies P, eds. (Oxford Handbooks Online), pp. 185–212.

Wisniewski JA, Muehling LM, Eccles JD, Capaldo BJ, Agrawal R, Shirley D-A, Patrie JT, Workman LJ, Schuyler AJ, Lawrence MG, et al. (2018). TH1 signatures are present in the lower airways of children with severe asthma, regardless of allergic status. J. Allergy Clin. Immunol 141, 2048–2060.e13. [PubMed: 28939412]

Witten DM, and Tibshirani RJ (2009). Extensions of sparse canonical correlation analysis with applications to genomic data. Stat. Appl. Genet. Mol. Biol 8, Article 28.

Woodruff PG, Modrek B, Choy DF, Jia G, Abbas AR, Ellwanger A, Koth LL, Arron JR, and Fahy JV (2009). T-helper Type 2-driven inflammation defines major subphenotypes of asthma. Am. J. Respir. Crit. Care Med 180, 388–395. [PubMed: 19483109]

Xu Z, Niu L, Li L, and Taylor JA (2016). ENmix: a novel background correction method for Illumina HumanMethylation450 BeadChip. Nucleic Acids Res. 44, e20. [PubMed: 26384415]

Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Girón CG, et al. (2018). Ensembl 2018. Nucleic Acids Res. 46 (D1), D754–D761. [PubMed: 29155950]

Zhou Y, Jackson D, Bacharier LB, Mauger D, Boushey H, Castro M, Durack J, Huang Y, Lemanske RF Jr., Storch GA, et al. (2019). The upper-airway microbiota and loss of asthma control among asthmatic children. Nat. Commun 10, 5714. [PubMed: 31844063]

Highlights

- MANAclust enables integrated analysis of clinical and multi-omics data
- Inter-variable relative information provides accurate categorical feature selection
- MANAclust outperforms competing approaches for multi-omic analysis
- MANAclust identifies clinically and molecularly distinct asthma clusters

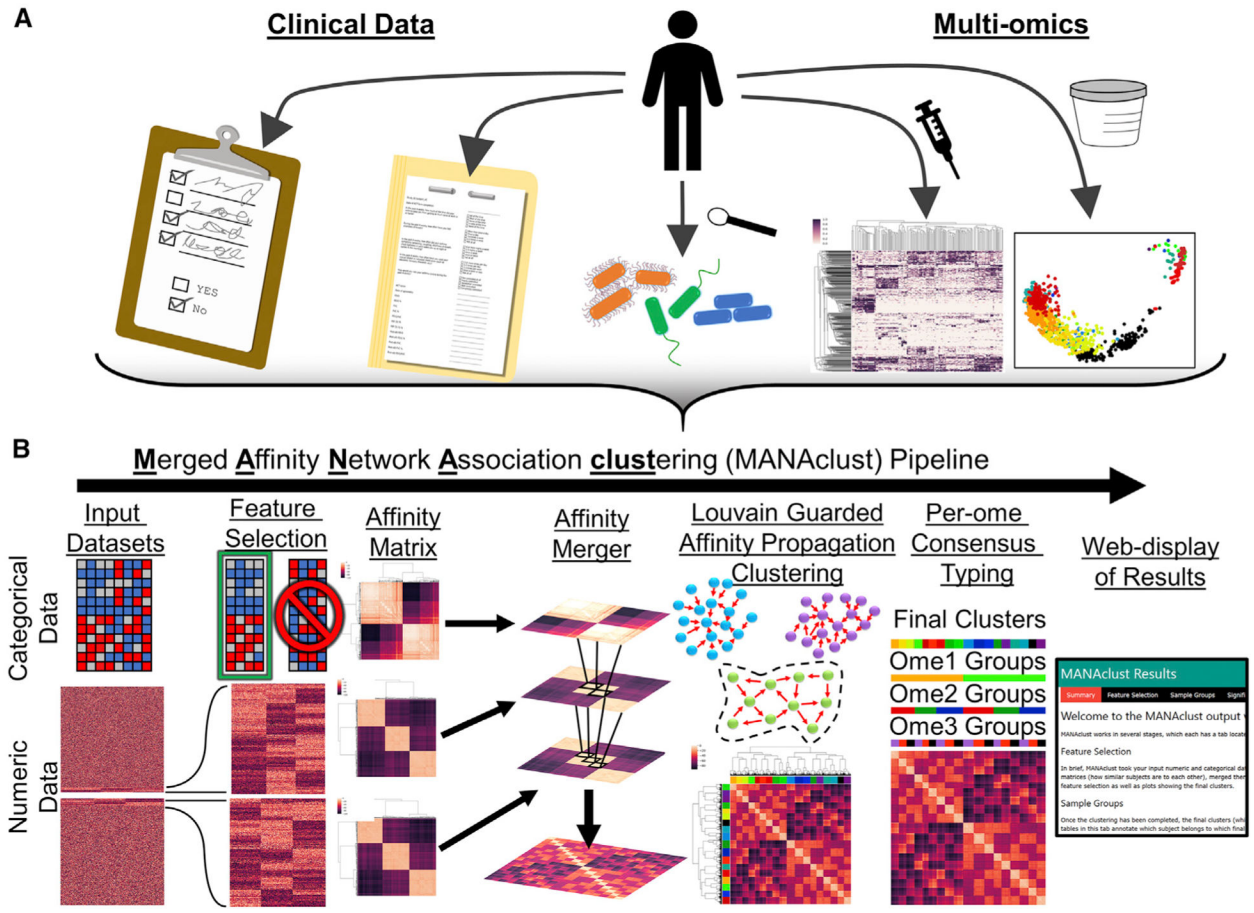


Figure 1. MANAclust pipeline

(A) Clinical data hold key information for defining diseases, yet they are often ignored by traditional (multi-)omics studies. MANAclust enables feature selection and merged analyses of categorical and continuous clinical data along with traditional multi-omics to enable discovery of disease subtypes.

(B) Categorical and numeric datasets are fed into the program, after which feature selection is performed (see STAR Methods for details on algorithms). Normalized affinity matrices are calculated across all omes; these affinity matrices are then merged by taking their missing value compatible average. We combined the strengths of Louvain modularity and affinity propagation into a new clustering algorithm that is then used for final cluster (FC) assignment on the combined affinity matrix. FCs are then examined within each input dataset to determine whether the FCs differ within each given dataset to identify the consensus groups. Post-clustering analyses are also performed to identify the significant differences across each FC and consensus group for all datasets. Analyses are then collated and displayed in a webpage format.

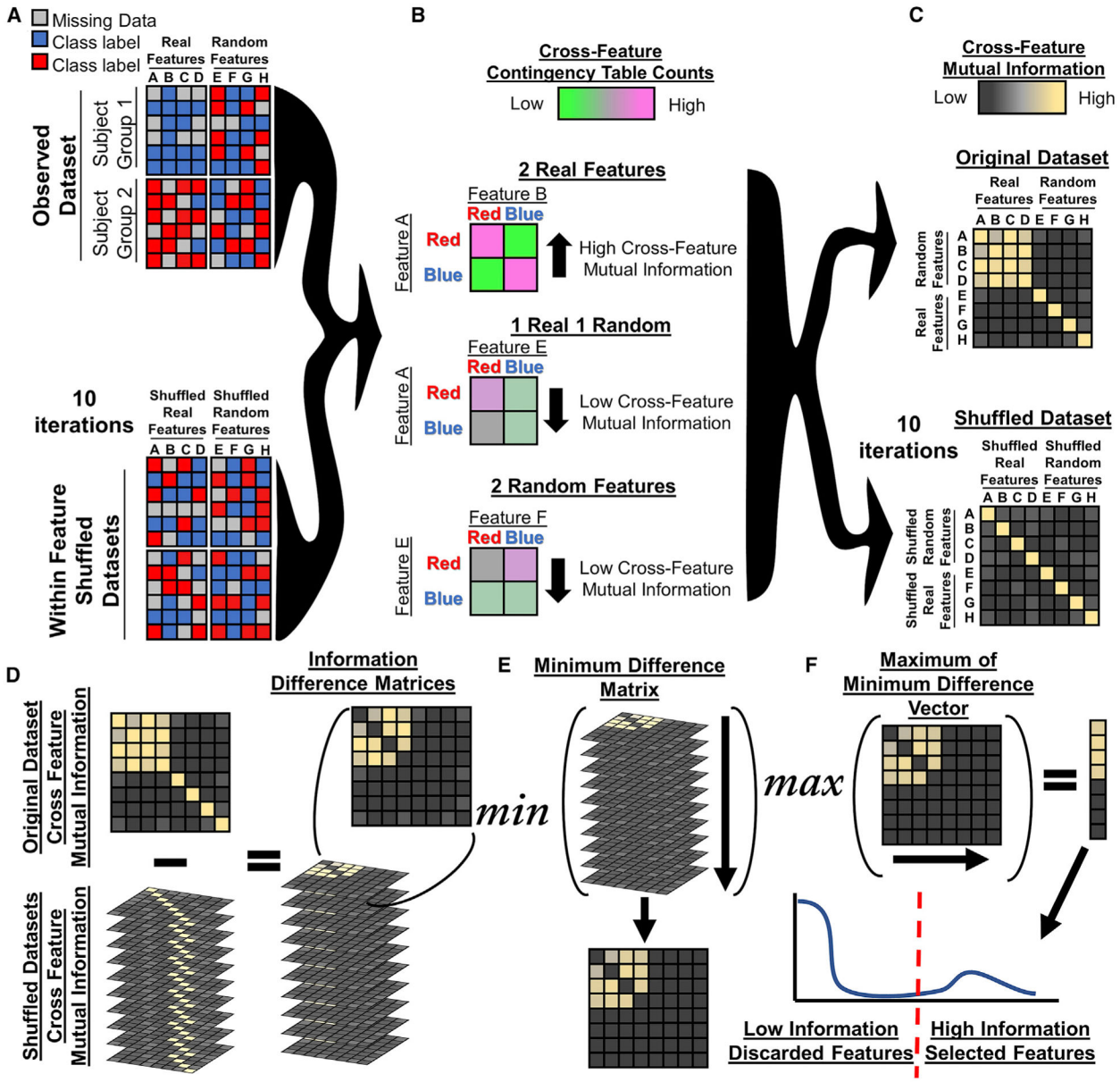


Figure 2. Categorical feature selection

(A) The observed dataset, which will often consist of both structured, meaningful, real features and randomly distributed features, is shuffled for 10 iterations to generate a background null distribution for feature selection.

(B) To select real features, each feature is compared with all other features, creating a contingency table of all observations for each feature pair.

(C) The mutual information for each feature pair is then calculated. This process is performed both with the observed dataset and its iteratively shuffled versions to create an accurate background for comparison.

(D) We then calculated a background-corrected stack of information difference matrices, having subtracted the null backgrounds from the observed cross-feature mutual information.

(E) This difference matrix stack is then flattened to a single two-dimensional matrix of feature-pairwise mutual information difference by taking the minimum for each feature pair. This is essentially the worst-case scenario in which there was the least amount of difference between the original dataset and one of the shuffled datasets.

(F) To select meaningful features, we calculated the row-wise maximum from the minimum difference matrix. This results in a one-dimensional vector corresponding to the maximum amount of relative information contained in all pairwise comparisons for each individual feature. A cutoff is then applied to select the high- and low-information features, retaining the high-information features for downstream log-loss and affinity matrix calculations.

See Figure S1 for a summary of the mathematical approach to feature selection on numeric datasets. See also Figure S2 for a summary of the accuracy of these feature selection methods.

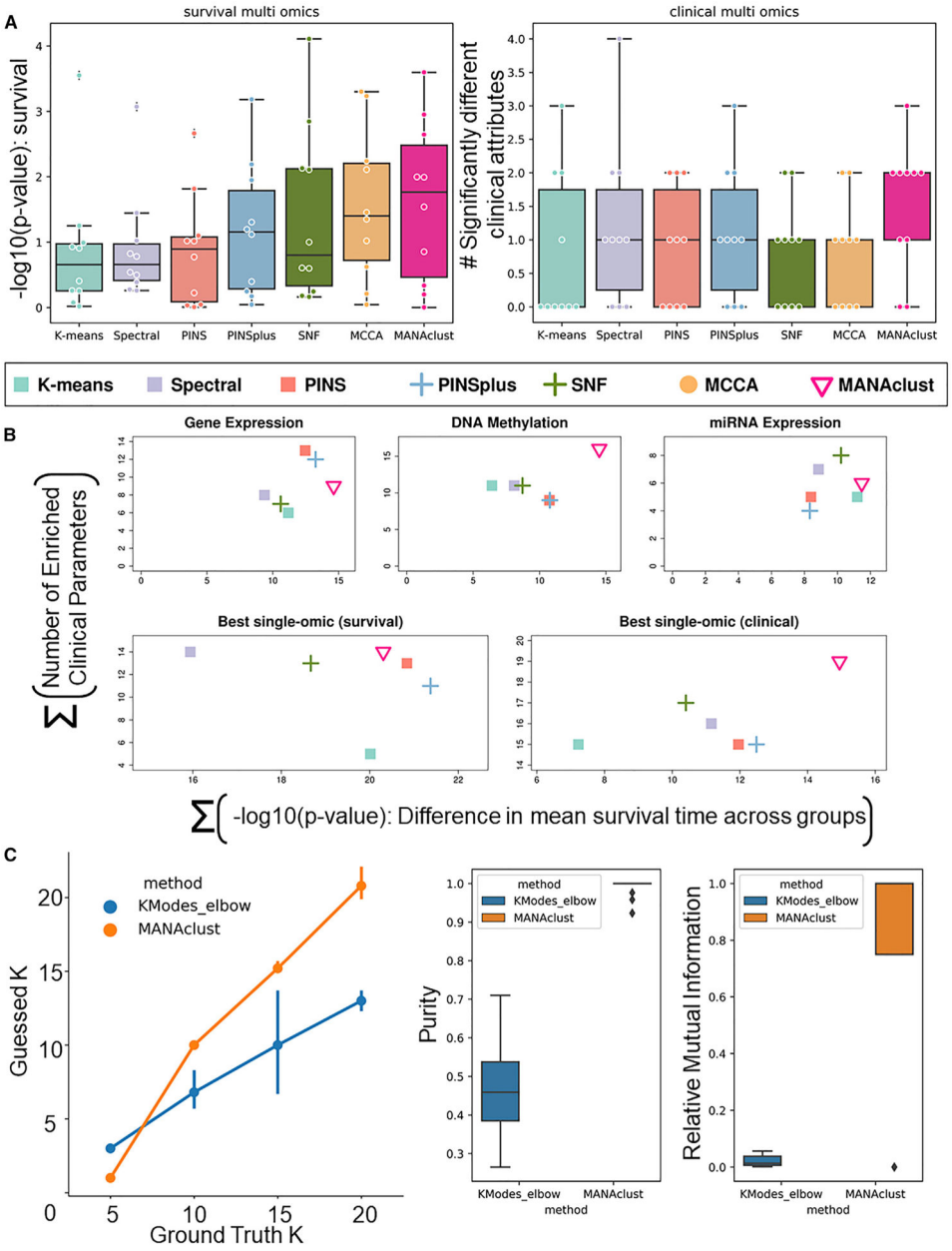


Figure 3. Comparison of MANAclust with other multi-omic clustering algorithms on TCGA data and simulated categorical data

We compared MANAclust's ability to identify clinically distinct clusters with five other existing algorithms: two non-multi-omics-specialized algorithms, K-means or spectral clustering on concatenated datasets; and four specialized multi-omics algorithms, PINS (Nguyen et al., 2017), PINSPlus (Nguyen et al., 2019), Similarity Network Fusion (SNF) (Wang et al., 2014), and sparse multiple canonical correlation analysis (MCCA) (Witten and Tibshirani, 2009).

(A) MANAclust finds the most significant differences in identified FCs in 10 different cancer types taken in aggregate. The clusters identified by each method were compared with each other for significant differences for survival rate and group segregation based on

clinical attributes. Displayed are the sum of the −log10(P values) for each cancer type and the sum of the number of statistically significant differences in clinical attributes for each method. See Table S1 for all tabulated results.

(B) All algorithms were also benchmarked using single-omes rather than in a multi-omic manner. In the top row of plots, each ome type (gene expression, DNA methylation, miRNA expression) denotes the sum across all cancer types for the −log10(P values) for significance in survival differences across groups (x axis) and the sum of the total number of enriched clinical parameters (y axis) for all cancer types combined. Note that MCCA works only with multi-omics data and is therefore not included in the single-omics comparison.

(C) To assess the ability of MANAclust to accurately incorporate categorical data, we performed a synthetic dataset benchmark comparing MANAclust's categorical clustering (orange) with KModes clustering using elbow rule on within-group sum of Hamming distances. MANAclust was slightly more accurate than KModes elbow rule in selecting the appropriate number of clusters. However, MANAclust significantly outperformed KModes in clustering purity ($F = 445$, $P = 1.38e{-22}$, one-way ANOVA) and relative mutual information ($F = 53.7$, $P = 8.75e{-9}$, one-way ANOVA). Categorical simulations included 5, 10, 15, and 20 groups with 1,000 subjects per simulation; each scenario was simulated five times.

See Figure S3 for benchmarking of MANAclust's feature selection algorithms.

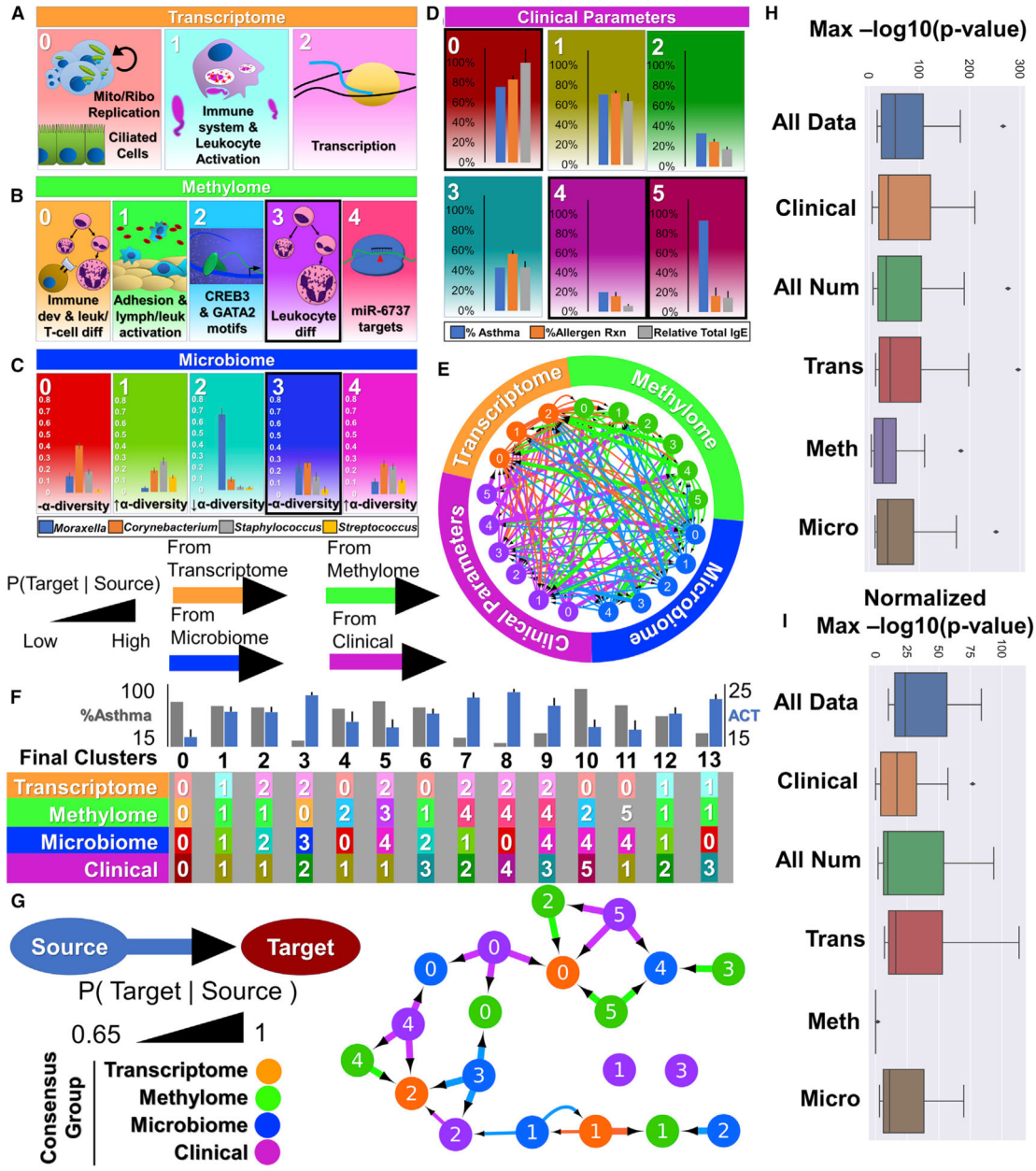


Figure 4. Clinical, methylation, transcriptome, and microbiome consensus groups

(A–D) For each input dataset, the large-scale take-home message from the top enriched pathways or variables that characterized the given consensus groups is shown. Each consensus group has a unique color that maps consistently throughout the figure. A thick border is drawn around consensus groups that uniquely map to a single consensus group of all other datasets. In other words, if one can determine this given ome, one can universally determine the subject's group membership across other omes.

(A) Pathway analysis was performed for the transcriptome-level consensus groups using PyMINEr; selected pathways were prioritized using PyMINEr's individual class importance metric (Table S2C) (Reimand et al., 2007; Tyler et al., 2019).

(B) For methylome pathway analyses, methylation loci were filtered only for those with promoter annotations, using the given promoter as genes for analysis by PyMINEr (Table S2D) (Reimand et al., 2007; Tyler et al., 2019).
(C) Relative abundance (by percentage) for the four most prevalent bacterial genera in each of the microbiome consensus groups. A note on the consensus group alpha diversity is also made beneath genus level quantifications; alpha diversity measures are shown directly in Figure S4D.
(D) Each consensus group shows the percent abundance of asthma, proportion of allergen-specific IgE (a marker of sensitization to specific allergens), and serum total IgE, normalized to mean total IgE of the highest group. See STAR Methods for details on quantification of the allergen-specific IgE calculation. Bars indicate means and errors are standard error.
(E) A circular graph network showing each data type's consensus groups and their connections to other consensus groups in different data types. All the subjects from within that consensus group of a given data type are examined and if at least one subject within the given consensus group maps to a consensus group from a different data type, those consensus groups are connected. The edges are weighted by the Bayesian probability of consensus group membership for the target given the source of the arrow. See Figure S4 for more details.
(F) The FCs are characterized by their membership in their given consensus groups for each data type. Colors within the table are consistent with those in (A)–(E). For each FC, the relative percentage with asthma is shown (gray bars) along with asthma control test (ACT) scores (blue bars), a measure of asthma severity. Bars indicate mean ± standard error.
(G) The significant subset of (E); only edges that showed significant concordance between the two consensus groups are shown. Edge thickness corresponds to the probability of membership in the target consensus group, given that a subject is a member of the source consensus group of a different dataset. Consensus group nodes are colorized based on their data type only.
(H) A boxplot showing the maximum level of significance (with FCs or consensus groups) using all data, clinical data only, all numeric data, or each numeric data type on its own. Using all data resulted in the greatest average significance of held-out variables.
(I) This finding was consistent, including when subtracting each variable's instance of least significance to normalize for the different baselines within a variable.

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
Human nasal brushings and fecal samples	Icahn School of Medicine at Mount Sinai	N/A
Chemicals, peptides, and recombinant proteins		
RNAlater	ThermoFisher Scientific	AM7020
Critical commercial assays		
TruSeq RNA Sample Prep Kit v2	Illumina	RS-122–2001 RS-122–2002
RNeasy mini kit	QIAGEN	74106
DNeasy Mini Kit	QIAGEN	69506
Infinium MethylationEPIC v1.0	Illumina	WG-317–1003
TG Nextera® XT Index Kit v2 Set A/B	Illumina	TG-131–2001 TG-131–2002
Deposited data		
Benchmarking datasets	The Cancer Genome Atlas (TCGA)	syn21301852
Software and algorithms		
PyMINEr	https://www.sciencescott.com/pyminer	N/A
MANAclust	This work; https://bitbucket.org/scottyler892/manaclust/src/master/	N/A
Qiime	http://qiime.org/	N/A
MicrobiomeAnalyst	https://www.microbiomeanalyst.ca/	N/A
ENmix	https://bioconductor.org/packages/release/bioc/html/ENmix.html	N/A