LaGrACE: Estimating gene program dysregulation using latent gene regulatory network for biomedical discovery

Minxue Jia[1,2], Haiyi Mao[1,2], Mengli Zhou[1,4], Yu-Chih Chen[1,2,3], Panayiotis V. Benos[1,2,4,*]

[1]Department of Computational and Systems Biology, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania, USA; [2]Joint CMU-Pitt PhD Program in Computational Biology, Pittsburgh, Pennsylvania, USA; [3]UPMC Hillman Cancer Center, Pittsburgh, Pennsylvania, USA; [4]University of Florida, Gainesville, Florida, USA

Abstract

Gene expression programs that establish and maintain specific cellular states are orchestrated through a regulatory network composed of transcription factors, cofactors, and chromatin regulators. Dysregulation of this network can lead to a broad range of diseases by altering gene programs. This article presents LaGrACE, a novel method designed to estimate dysregulation of gene programs utilizing omics data with clinical information. This approach facilitates grouping of samples exhibiting similar patterns of gene program dysregulation, thereby enhancing the discovery of underlying molecular mechanisms. We rigorously evaluated LaGrACE's performance using synthetic data, bulk RNA-seq clinical datasets (breast cancer, chronic obstructive pulmonary disease (COPD)), and single-cell RNA-seq drug perturbation datasets. Our findings demonstrate that LaGrACE is exceptionally robust in identifying biologically meaningful and prognostic molecular subtypes. Additionally, it effectively discerns drug-response signals at a single-cell resolution. Moreover, the COPD analysis revealed a new association between LEF1 and COPD molecular mechanisms and mortality. Collectively, these results underscore the utility of LaGrACE as a valuable tool for elucidating underlying mechanisms of diseases.

Introduction

Gene programs are sets of genes that facilitate transcriptional regulation mechanisms and collectively contribute to a specific biological function under particular conditions or phenotypes (Harris et al. 2021). Gene programs can interact and work cooperatively through complex mechanisms. There are a number of factors that affect the gene programs and the complex mechanisms they participate, like transcription factor (TF) activities, protein signaling, chromatin accessibility or microRNA expression (Lambert et al. 2018). In most typical datasets, multiple of these modalities are not observed. Furthermore, the control that these factors exert on the gene programs and their interactions creates collinearity in the expression of member genes. Patient phenotypes may be the result of differences in these gene programs and their connections, reflecting their roles in clinical outcomes, responses to treatments, or molecular disease subtypes. Although these factors are not directly observable, they provide a more comprehensive picture of the genetic mechanism of disease and facilitate the development of effective therapeutic strategies.

To harness this potential, significant advances have been made in methods that quantify gene program activity on single patient sample or individual cell through unsupervised factorization approaches, such as non-negative matrix factorization (NMF). But most methods are not generally identifiable (Pan and Doshi-Velez 2016), which hampers interpretability of the identified factors. Moreover, some orthogonal factor models overlook the fact that gene programs are organized in networks that collectively control the observed expression of their member genes. Conversely, considerable progress has been made in methods that quantify gene program activity by leveraging existing information from curated databases. Examples of such methods include Spectra (Kunes et al. 2023), expiMap (Lotfollahi et al. 2023), Pathifier (Yotam, Michal, and Eytan 2013) and ssGSEA (Barbie et al. 2009). Nevertheless, these approaches strongly depend on the curated gene programs, which are derived from multiple biological contexts, and may not correspond to the specific context under study.

Those limitations become particularly evident when considering the dynamic nature of gene regulation in response to treatment or disease (Lee and Young 2013). Such alterations can lead to divergent changes in functionally relevant gene programs if shared regulatory elements (e.g., bound TFs and mediators) are modified across conditions (**Figure 1A**). Measuring alterations in the gene programs regulatory network is challenging. A primary challenge in biomedical research is the inherent inability to observe identical samples simultaneously in both treatment and control states. Motivated by our understanding of gene regulatory networks, we introduce LaGrACE (**La**tent **Gr**aph-based individu**A**l **C**ausal **E**ffect Estimation). LaGrACE is a novel approach designed to estimate regulatory network-based pseudo control outcome to characterize gene program dysregulation for samples within treatment (or disease) group. This method enables grouping of samples with similar patterns of gene program dysregulation, thereby facilitating discovery of underlying molecular mechanisms induced by treatment or disease.

Results

LaGrACE infers divergence score for gene program dysregulation

The gene programs, that collectively engage in the same biological function, often interact with each other (Kunes et al. 2023). Perturbation in those gene programs and their interactions can lead to a cascade of dysregulation, affecting gene expression, pathways, and other cellular processes. A significant challenge arises in modeling alterations in gene programs within a single sample, especially when a biomedical sample can typically only be observed in one condition at a time, either the control or the non-control (treatment) state. This is a common issue in biomedical studies.

We propose a novel approach to estimate single-sample changes in gene program interactions compared to a reference gene program network. First, we start by assuming that there is a Bayesian network, delineating gene program interactions (Friedman 2004; Segal et al. 2003) present in all control samples (“reference network”). In any given disease (or treatment) sample, we expect that some of the gene programs and their interactions would change. To assess these changes in this sample, we do the following. In the “reference network", we build a predictor of a gene program activity by using the variables in its Markov blanket (i.e., the parents, children, and spouses; see **Methods**). Second, we measure the difference between the expected value of each gene program based on the “reference network” and the observed values in the (treatment or disease) sample. Specifically, we assume that the general structure of this “reference network” for the most part is maintained in samples from the disease (or treatment) group, although certain part of this network may be altered in the disease (or treatment) samples. The altered parts will depend on the particular characteristics of the patient. This approach allows for the assessment of sample-level dysregulation of the (unobserved) gene programs in disease (or treatment) samples, based on omics profiling and clinical features (see **Figure 1B**). But this raises a question: if a given disease sample could also be observed in a control state, will gene programs with analogous biological functions exhibit comparable behavior?

In a clinical study, we can only measure transcriptomic profiles in a sample once: either in the control or in the disease (or treatment) state, leaving the other side unobserved. Let W (1: treated; 0: untreated) represent the treatment label assignment variable for an individual j, where only a portion of the variables can be directly observed. Let V_{ij} denote the observed feature i (gene program or clinical variable) from individual j:

$$V_{ij}^{0} = \begin{cases} V_{ij}, W_j = 0 \\ ?, W_j = 1 \end{cases}, V_{ij}^{1} = \begin{cases} ?, W_j = 0 \\ V_{ij}, W_j = 1 \end{cases}$$

To estimate control state ($W_j = 0$) of a disease (or treated) sample V_{ij}^1, we calculate the pseudo control outcome $\hat{V}_{ij}^0$, leveraging Markov blanket of V_i derived from the "reference network" (**Figure 1C**). This approach assumes that gene programs with analogous biological functions in control samples maintain similar behavior in disease samples.

Before estimating the pseudo-control outcome, we first discuss how to estimate gene programs. We infer gene programs from high-dimensional gene expression data. Gene programs are typically estimated using factor analysis or NMF methods. However, these methods often face issues with identifiability and impose orthogonality constraints. Here, we utilize the LOVE algorithm (Bing et al. 2019), which addresses these limitations by providing (1) theoretical guarantees regarding the unique identifiability of latent factors without relying on assumptions about the data-generating mechanisms, and (2) the ability to identify non-orthogonal latent factors without restrictive assumptions regarding orthogonality.

Then, we infer a reference network, represented as a Bayesian network, to delineate interactions among gene programs (Z) and between gene programs and clinical variables (U). With such a reference network inferred from reference (or control) samples, the value of a gene program Z_i can be approximated based on its Markov Blanket $MB(Z_i)$ using a regression model. A set of Markov Blanket regression models, trained on reference samples, represents reference gene program interactions. The pseudo control outcome of treatment samples, $\hat{Z}_i^0$, can then be estimated using those trained regression models based on reference samples. If Markov Blanket of a gene program is altered under treatment condition, this alteration can be quantified by calculating expectation divergence, which is the difference between observed value and the pseudo control outcome in treatment samples.

$$\text{Divergence Score: } V_{ij}^1 - \hat{V}_{ij}^0$$

These divergence scores can then be used to group samples based on patterns of dysregulation by evaluating alterations within the reference network and their magnitudes.

Changes in the gene program regulation network, induced by disease or treatment, result in phenotypic changes including mortality and morphological transformations. Notably, identical phenotype or outcomes associated with a disease or treatment can be observed as a result of changes in different parts of this network. For instance, patients with Alzheimer's disease or cancer can be stratified into molecular subtypes characterized by distinct dysregulated pathways (Neff et al. 2021; Rosario et al. 2018; Yang et al. 2024). In this work, we evaluate LaGrACE based on the ability to identify such subtypes.

LaGrACE effectively recovers network subtypes in low- and high-dimensional synthetic data

Before applying LaGrACE to high-dimensional, collinear datasets for subtype identification, we conducted initial evaluations on low-dimensional synthetic datasets. Each synthetic dataset consisted of a reference network alongside three perturbed networks, each subjected to random edge alterations (both deletions and additions). These modifications simulate variations corresponding to disease subtypes, treatment categories, or clinical interventions (**see Methods**). All simulated networks consisted of 50 continuous and 25 categorical variables that resemble the variable types observed in biomedical datasets. Each network was used to generate 1,000 samples. LaGrACE learned the reference network from the reference network samples. Subsequently, we calculated LaGrACE divergence scores for each "disease" sample, generated by one of the perturbed networks. We clustered the samples based on the divergence scores (**see Methods)**. The resulting clusters demonstrated improved

separation and better Adjusted Rand Index (ARI) and Adjusted Mutual Information (AMI) scores compared to observed continuous variables from synthetic data (**Figure 2A & B**). These results suggest LaGrACE's superior performance in subtype delineation. This experiment was repeated with smaller reference sample sizes, 100 and 500 samples randomly drawn from 1000 original samples. Even with fewer reference samples, LaGrACE's divergence estimation method continued to accurately identify distinct clusters (**Figure 2E**).

Next, we evaluated LaGrACE on high-dimensional synthetic datasets, comprising 25 discrete and 2500 continuous variables generated from simulated latent factors (**see Methods**), which represent clinical features (discrete) and gene expression (continuous). The perturbed groups delineated by LaGrACE divergence score demonstrated better separation comparing to observed continuous variables' (**Figure 2C**). We clustered the samples using (1) observed continuous variables, (2) ground truth latent factors, (3) LaGrACE divergence with LOVE-recovered latent factors (LOVE LF) and (4) LaGrACE divergence with ground truth latent factors (**Figure 2D**). Notably, LaGrACE divergence score achieved best performance. Also, LaGrACE based on LOVE LF exhibited performance comparable to LaGrACE with ground truth latent factors. This indicates that LaGrACE is robust for subtyping tasks in high-dimensional and collinear datasets.

LaGrACE demonstrates precise drug response detection at single-cell level

We also tested LaGrACE's performance in a number of real-life datasets. scRNA data were collected from A549 lung adenocarcinoma cells treated with dexamethasone over intervals of 0, 1, and 3 hours (Cao et al. 2018). Addressing scRNA-seq data sparsity, we utilized imputed RNA features estimated by scVI (Lopez et al. 2018) as input for LaGrACE. The vehicle group served as a reference for network inference, Markov blanket estimators, and divergence score computation. LaGrACE divergence score facilitated clear stratification of cells by treatment duration in UMAP embedding, surpassing the classification achieved using (unimputed & imputed) RNA features and gene set scores inferred using Spectra and expiMap (**Figure 3A & 3B**). Higher ARI, AMI and treatment time purity scores in LaGrACE clusters underscored its capacity at processing single-cell level data (**Figure 3C**). Association among gene expression features, regulated by LF1 and its Markov Blanket LF2, varied at difference treatment conditions (**Figure 3F**), leading to changes in gene programs. Strikingly, divergence scores for LOVE latent factors exhibited higher Spearman correlation with treatment duration compared to the latent factors themselves (**Figure 3D**). This indicates that LaGrACE effectively discerned gene program dysregulation.

Three latent factors with divergence scores exhibiting significant correlations with treatment duration were identified. These factors were significantly enriched in pathways related to E2F targets, the p53 pathway, and the complement system (**Figure 3E**). Such findings align with known dexamethasone effects on p53 phosphorylation, E2F transcriptional activity and synthesis of complement proteins in A549 cells (Greenberg et al. 2002; Hill et al. 1993; Urban et al. 2003).

Further application of LaGrACE to an additional scRNA-seq dataset, A549 cells treated with small molecules (BMS-345541, nutlin-3A, vorinostat) over a 24-hour period and across various doses (Srivatsan et al. 2020), showed that our approach adeptly captured dose-responsive signals. This was evident from UMAP embeddings, clustering accuracy (ARI and AMI values), and cell type purity scores inferred from drug dose information, surpassing the performance of gene expression features and prior knowledge-based gene set scores (**Supplementary Figures S1-S3**).

New Breast Cancer molecular subtypes identified and validated by LaGrACE

Application of LaGrACE to the METABRIC breast cancer dataset uncovered potential new molecular subtypes. In the absence of normal tissue samples in METABRIC, basal subtype samples, which is significantly different from all other molecular phenotypes, was used as the reference group. Utilizing 1368 latent factors from RNA-seq data, along with clinical features (i.e., ER/HER2/PR status, positive lymph nodes, age at diagnosis, and tumor size), LaGrACE identified six clusters (**Figure 4A – 4H and Table 1**) significantly associated with molecular subtypes (p-value < 2.2e-16), disease-specific survival (p-value = 1.31e-22), distant (p-value = 1.91e-21) and loco-regional relapse (p-value=1.45e-5), and clinical features (**Supplementary Table S1**). Cluster 0 notably grouped 82% of HER2+ and 29% of luminal B samples (**Table 1**). Clusters 1, 2, and 3 were mainly luminal samples, with higher proportions of luminal A in cluster 1 and luminal B in cluster 3. Claudin-low samples were primarily found in Clusters 4 and 5. Clusters 1, 2, and 3, all luminal, showed varying survival rates, metastasis rates, and other clinical features (**Supplementary Table S1 and Supplementary Figure S4)**. with Cluster 1 showing better survival and lower metastasis rates compared to Clusters 2 and 3. Cluster 2 was differentiated from Cluster 3 by menopausal states, tumor cellularity, and the Nottingham Prognostic Index. The Claudin-low Clusters 4 and 5 exhibited variations in cancer histology and menopausal states, highlighting our method's ability to distinguish different clinical phenotypes within similar molecular subtypes (**Supplementary Figure S5)**.

For validation, cluster memberships were assigned to new samples from an external breast cancer dataset (SCAN-B) based on their divergence scores. The clusters assigned to SCAN-B samples were significantly associated with molecular subtypes (p-value<2.2e-16), survival (p-value=1.61e-07) and clinical features (**Figures 4I, 4J, Supplementary Table S1**). Following the same trend as in METABRIC, most HER+ samples were in cluster 0 (**Table 1**). Cluster 1 had the highest survival rates, and cluster 2 exhibited better histological scores than cluster 3 (**Supplementary Figures S4-S5 and Supplementary Table S2**). No samples were assigned to cluster 5. Notably, the METABRIC patients were enrolled before trastuzumab availability (anti-HER2 treatment), leading to a higher hazard ratio in Cluster 0 compared to that observed in the SCAN-B dataset.

Comparison with prior knowledge-based methods (pathway information) yielded similar results for molecular subtypes within the METABRIC dataset (**Supplementary Figure S6**). Single-sample Gene Set Enrichment Analysis (ssGSEA) and Pathifier identified 7 and 9 clusters, respectively, which were significantly associated with molecular subtype, survival, and metastasis (p-values: <2.2e-16, 4.22e-16, 6.33e-14 for ssGSEA; <2.2e-16, 7.06e-9, 1.02e-7 for Pathifier). Employing the linear predictor derived from the Cox proportional hazard model, we evaluated the discriminatory power of identified clusters using Harrell's C-index and the concordance probability estimate (CPE). LaGrACE outperformed both ssGSEA and Pathifier in prognostic value for survival and recurrence (Supplementary Table S3). Time-dependent AUC analysis also showed LaGrACE's superior discrimination ability in metastasis, survival, and local relapse in both METABRIC and SCAN-B datasets (**Figure 4K**). Overall, LaGrACE achieved great performance in classifying tumor samples by molecular subtype, survival, and disease progression.

Identification of Biological Networks Driving Breast Cancer Subtypes Networks

LaGrACE clusters were instrumental in analyzing molecular alterations impacting clinical phenotype and progression of breast cancer. Epithelial cell differentiation score (Prat et al. 2010), proliferation index (Nielsen et al. 2010), and immune cytolytic activity (Rooney et al. 2015) were independently associated with LaGrACE subtypes (**Supplementary Figures S7A-C**). The luminal clusters 1, 2, 3 demonstrated varying scores: Cluster 1 exhibited lower differentiation and proliferation rates but higher cytolytic activity. Conversely, clusters 2 and 3 presented higher differentiation and lower cytolytic activity, with cluster 3 exhibiting increased proliferation rates. Claudin-low clusters 4 and 5 were characterized by a

high inferred immune cellular fraction (**Supplementary Figure S8**), with Cluster 4 expressing high levels of immune exhaustion markers PD-1 and LAG3 (**Supplementary Figure S7F**). This indicates an immunosuppressive environment and early tumor progression (Davidson et al. 2020; Rye et al. 2022).

LaGrACE subtypes allow us to develop subtype-specific networks, integrating clinical variables and latent factors to elucidate disease-associated molecular mechanisms. We used FCI-Stable algorithm to infer networks for each luminal cluster (**Supplementary Figure S9**). Focusing on tumor size-related variables, all luminal cluster networks found lymph node number to be directly linked to tumor size, but different latent factors were associated with tumor size in each subtype. In cluster 1, a latent factor linked to tumor size is related to apical junction, which stabilizes cell-cell adhesion and cell polarity. Metastasis is characterized by loss of cell-cell adhesion and cell polarity (González-Mariscal et al. 2020; Heikenwalder and Lorentzen 2019). This may explain the low metastasis rate in cluster 1 due to its significantly higher pathway enrichment score in junction-associated pathways, compared to clusters 2 and 3 (**Supplementary Figure S7D**).

For clusters 2 and 3, tumor size was linked to latent factors associated with by E2F transcriptional factors, which have been linked to the aggressiveness of breast cancer (Johnson et al. 2016; Oshi et al. 2020). Cluster 1 exhibited downregulation in E2F target genes, including FOXM1, AURKB, and PLK1 (**Supplementary Figure S7E**). Those genes promote the WNT/β-catenin signaling pathway, crucial in breast cancer progression (Chen et al. 2016; Mancini et al. 2019; Wang et al. 2020; Xu et al. 2020; Zhang et al. 2011). We validated this finding by inhibiting FOXM1 in breast cancer cell lines, and we observed reduced cell motility (**Supplementary Figures S10**). This further underscores LaGrACE's ability to uncover important disease-relevant molecular subtypes with distinct disease mechanisms.

Identification by LaGrACE of LEF1 and as an Important COPD Mortality Factor via LaGrACE

We further applied LaGrACE to blood-derived RNA-sequencing gene expression measurements from 1750 subjects from COPDGene study, which is a longitudinal study that aims to discover genetic basis of COPD disease mechanism. The reference group was meticulously selected based on following criteria: participants (a) had clinical data from both phase 1 (baseline) and phase 2 (5-yr follow-up) visits; (b) had no missing data for forced expiratory volume in 1 second (FEV1) and forced vital capacity (FVC); (c) had no spirometric abnormalities (i.e., GOLD 0) in phase 2 and phase 3 (10-yr follow-up) visits; (d) were not current smokers; (e) exhibited less than 5% percent emphysema in both phase 2 and phase 3 visits; (6) had less than 5% decrease in FEV1 %predicted between phase 2 and phase 3 visits. This filtering procedure resulted in 139 reference and 1611 COPD subjects (former and current smokers).

Utilizing 602 latent factors from RNA-seq data and gender information, LaGrACE identified four clusters. Some clinical variables were significantly different between clusters, including spirometry, chest CT scan, symptom questionnaires, white blood cell differentials as well as smoking status (**Figure 5A & 5C, Table 2**). Cluster 0 and 3 represented the most impaired groups, showing poorer spirometry results (FEV1/FVC ratio), lower six-min walking distance, more symptom (measured by COPD Assessment Test and SGRQ St. George's Respiratory Questionnaire) and dyspnea (measured by Medical Research Council Dyspnea Scale). The neutrophil to lymphocyte ratio (NLR), a valuable predictor of acute exacerbations of COPD and mortality (Paliogiannis et al. 2018), displayed distinctive patterns in cluster 3, which is a cluster primarily consisting of younger African American samples with more current smoking participants, higher lymphocyte and lower neutrophile level. This variation may be attributed to racial differences in leukocyte subpopulations and immune response (Freedman et al. 1997; Nédélec et al. 2016) Cluster 1 and 2 showed comparable FEV1/FVC ratio and 6-min walking distance (**Supplementary Figure S11**). Cluster 1 had lower neutrophile and CD4 naïve T cell percentage, but higher lymphocyte and CD4 resting

memory T cell percentage compared to cluster 2 (**Supplementary Figure S12**). Additionally, the identified clusters were significantly associated with mortality (**Figure 5B**), with cluster 1 maintaining the best survival probability and cluster 0 exhibiting the worse survival probability. CD4 naïve and CD4 resting memory T cells, exhibiting significant difference across the four clusters, were prognostic for white patients but not for African American patients (**Supplementary Figures S13A-B**).

Two latent factors (**Figure 5F & 5G**), enriched in lymphocyte percentage and CD4-specific genes, correlated significantly with lymphocyte parentage (measured from complete blood count test) and CD4 T cells (sum of CD4 naïve T cell and memory T cellular fractions). Their LaGrACE divergence scores were significantly correlated with FEV1/FVC ratio and FEV1 %predicted (**Table 3**). Unlike cluster 0 and 3, Cluster 1 showed no dysregulation on CD4 T and lymphocyte latent factors (**Figure 5E**). Based on the inferred reference network, the two latent factors were linked to LEF1 motif (**Figure 5D**). The transcriptional factor LEF1, which binds to β-catenin (CTNNB1) and regulates lymphopoiesis via canonical WNT signaling, also promotes commitment to the CD4 lineage through induction of Th-POK (Steinke et al. 2014; van de Wetering, de Lau, and Clevers 2002). Thus, we evaluated the association between LEF1 with COPD mortality. LEF1 transcriptome emerged as a risk factor for overall survival, as demonstrated through multivariate Cox regression analysis that adjusted for other clinical factors (**Figure 5H**). Also, LEF1 gene expression was found to be prognostic for both white and African American patients (**Supplementary Figures S13C**).

Further investigation into the association between LEF1 and COPD disease progression revealed that LEF1 gene expression correlated significantly with age, FEV1 %predicted, and FEV1/FVC ratio (**Table 3**). Importantly, the positive correlation between LEF1 and both spirometric measures (FEV1 % predicted, FEV1/FVC) persisted after controlling for age. To validate the association between LEF1 and pulmonary function, blood-derived gene expression microarray data from the ECLIPSE study were employed. This dataset captured expression levels of two LEF1 gene regions: the full-length LEF1 and a partial region. Only the full-length LEF1 exhibited a positively correlation with FEV1 %predicted and FEV1/FVC ratio. This association was not observed for the partial region lacking the β-catenin-binding domain. Consistently, CD4 and CD8 T cells from COPD lung tissue, mirroring the transcriptional state of CD4 memory T cells from blood samples of COPD patients, exhibited lower LEF1 regulon scores (**Supplementary Figures S14**).

Discussion

In this paper, we introduce LaGrACE, a novel method to quantify sample-level gene program dysregulation in observational biomedical datasets. Transcriptomic profile of samples can only be observed either in the control or the non-control state. Given a set of control (reference) samples, our method estimates how gene program regulation differs in each non-control sample (e.g., disease or treatment). In simulation studies, LaGrACE demonstrates robust performance in characterizing alteration of regulatory network across varying reference sample sizes and dimensionality. LaGrACE also outperforms existing single cell gene program methods without *prior knowledge* and effectively discerned drug treatment signals across varying doses and exposure times at a single-cell resolution.

LaGrACE effectively identifies biologically significant and prognostic clusters in breast cancer data (METABRIC). The stability of these identified clusters was affirmed through validation in the SCAN-B cohort, illustrating our method's reliability in patient stratification. Notably, these clusters provide deeper prognostic insight than molecular subtypes and clusters deduced from *prior knowledge*-based methods. Upon constructing cluster-specific networks, we observed significant associations between

disease progression in ER-positive breast cancer patients and key signaling pathways, including apical junction system and E2F transcription factor. Our findings suggest that LaGrACE is a potent tool for elucidating the underlying mechanisms of diseases characterized by complex and heterogeneous pathophysiological profiles.

We also identify a novel association between the full-length LEF1 isoform and COPD molecular mechanisms and mortality. This association suggests that differential expression of LEF1 isoforms in COPD patients may perturb canonical WNT signaling and lymphopoiesis, consequently altering immune response (Feder et al. 2020). Furthermore, the phenomenon of cellular senescence is known to correlate with several hallmarks of COPD pathogenesis (Araya & Kuwano, 2022). Declines in LEF1 expression levels and the ratio of full-length to truncated LEF1 isoforms with aging have been implicated in contributing to cellular senescence, thereby exacerbating COPD pathogenesis (Elyahu et al. 2019; Jia et al. 2023).

LaGrACE has some limitations. First, its underlying formulation may deviate from the biological nature of the data due to assumptions about linearity between observed variables (Xs) and latent factors (Zs), as well as between clinical features (Ys) and latent factors. Other limitations include the assumption of *acyclicity* between latent factors and the *Gaussianity* assumption for Xs and Ys. However, the linearity assumption between Xs and Zs is a reasonable approximation in many cases, as reflected in our results with real-life datasets (breast cancer, COPD, and single-cell analysis). In the future, we plan to extend LaGrACE to non-Gaussian and non-linear models.

Materials and Methods

LaGrACE Algorithm

LaGrACE is an algorithm that evaluates gene program dysregulation through estimating pseudo control outcome on a gene program or a clinical variable. LaGrACE comprises four main steps, which are summarized below and elaborated in the following sections.

(1) **Gene Program Discovery**: infer biologically meaningful latent factors (gene programs) from gene expression profiles of reference samples and then calculate factor matrix for treatment group using loading matrix inferred based on reference samples.
(2) **Reference Network Learning**: learn a Bayesian network encapsulated interactions among inferred gene programs and clinical variables of interest.
(3) **Pseudo Control Outcome Estimation:** estimate pseudo control outcome to approximate pre-treatment state on a gene program or a clinical measurement of disease (or treated) samples by utilizing its Markov blanket derived from the reference network.
(4) **Divergence Estimation**: calculate difference between observed value and pseudo-control outcome for treatment samples.

In most biomedical study, there are omics (gene expression) data and clinical variables related to a specific treatment (or disease). Each sample in the dataset is labeled to denote whether it has received this treatment. Assume that the treatment is well-defined, and feature interactions are invariant across reference samples, we denote an individual sample in treatment state by W = 1 and the reference one by W = 0.

As a notation, let $X \in R^{N_{genes} \times N_{samples}}$ represent gene expression matrix, and $Z \in R^{k \times N_{samples}}$ represent k gene programs inferred from gene expression profile X, where $k < N_{genes}$. The matrix $A \in R^{N_{genes} \times K}$ represents the membership matrix assigning observed gene expression features to k latent factors. Since biomedical data usually include clinical measurements, we denote $U \in R^{N_{variables} \times N_{samples}}$ represent clinical variable matrix, where clinical variable may be either continuous or discrete.

The reference network, representing conditional independence relations among variables, delineates interactions among gene programs inferred from gene expression profiles and clinical variables. Let $V = [Z, U]$ denote random variables in the network, and V_{ij}^0 be value of feature i (could be gene program Z or clinical variable of interest U) in a sample j if the value was observed from an individual with a treatment label W = 0. For an individual j from treatment group, V_{ij}^1 represents observed value of feature i, but V_{ij}^0 is unknown. We infer pseudo control outcome $\hat{V}_{ij}^0$ to approximate V_{ij}^0. We focus on understanding its impact without the complication of multiple treatments or interference effects.

Gene Program Discovery

The (soft) clustering approach LOVE (Bing et al. 2019), a robust and scalable distribution-free latent model-based statistical method, is capable of inferring biologically meaningful and non-orthogonal latent factors. The latent factor model is formularized as

$$X = AZ + E$$

The model identifiability requires a mild condition that, each latent factor needs to have at least two primary variables assigned uniquely to it. Unlike other factorization models, this method allows a slight departure from independence among latent factors that are inferred from observed gene expression features. Hyperparameters (delta) were selected using Akaike Information Criterion (AIC) method.

Reference Network Learning

Once gene programs are identified, interactions among those gene programs and clinical variables of interest are modelled using the Fast Greedy Equivalent Search (FGES) algorithm (Ramsey et al. 2017) on reference samples. FGES constructs a partially Directed Acyclic Graph (pDAG), which contains both directed and undirected edges and can be used to represent a Markov equivalence class of the data-generating directed acyclic graph (DAG) G, over all input variables. FGES holds the Causal Markov assumption, stating that each variable in a graphical model is conditionally independent of every other variable (that is not affect by the variable), conditional on the variable's direct causes. FGES starts its search with an empty graph, then adds a single edge increasing DG score. Finally, it removes single edges while the score increases.

To accommodate the potential for discrete features within clinical data, FEGS was performed by maximizing degenerate Gaussian (DG) score (Andrews, Ramsey, and Cooper 2019). Let $V = (V_1, \cdots, V_p)$ be a set of p random variables with n instances, where V_i may be either continuous (**C**) or discrete (**D**). The DG score relies on standard transformation (One-hot Encoding) of variables:

$$T_i = \begin{cases} V_i & if\ i \in C \\ l_1(V_i), \dots, l_{k-1}(V_i) & if\ i \in D \end{cases}$$

Where $l_k(V_i)$ is an indicator function such that $l_k(V_i) = 1$ if $V_i = k$, otherwise $l_k(V_i) = 0$ for $V_i \in D$. The One-hot Encoding transformation embeds the discrete variables into a degenerate distribution. Thus, the transformed data is further treated as jointly Gaussian. The DG score is sum of local components (its nodes and their parents):

$$S(G,T) = \sum_{i=1}^{p} s\left(T_i, T_{Pa_i^G}\right)$$

where $T_{Pa_i^G}$ represents the set of parents of T_i, and

$$s\left(T_i, T_{Pa_i^G}\right) = l\left(\hat{\theta}_{mle} \middle| T_{\{i\} \cup Pa_i^g}\right) - l\left(\hat{\theta}_{mle} \middle| T_{Pa_i^G}\right) - \frac{PD}{2} |T_i| \left|T_{Pa_i^G}\right| log\ (n)$$

PD is a penalty value (penalty discount) that controls the sparsity of graph. $l\left(\hat{\theta}_{mle} \middle| T_{\{i\} \cup Pa_i^g}\right)$ and $l\left(\hat{\theta}_{mle} \middle| T_{Pa_i^G}\right)$ are Gaussian log-likelihood of the transformed data T given the causal DAG G and its parameters (θ).

Pseudo Control Outcome Estimation

The pseudo control outcome approach is designed to approximate pre-treatment state on a gene program or a clinical measurement by utilizing its Markov blanket based on the reference network. The Markov blanket of a variable V_i, MB(V_i), consists of the parents, children, and spouses, which are connected by adjacent directed or undirected edges under the causal Markov assumption. The variable

of interest is considered independent of all other variables, conditioned on its Markov blanket. Once the reference network is estimated, all discrete variables are transformed into binary indicator variables, representing categories of the original discrete variable. A set of Markov blanket-based regression models are then built for each variable in the reference network using reference samples.

$$V_i{}^0 = \sum_{V_k{}^0 \in MB(V_i{}^0)} \beta_{k,i}{}^0 V_k{}^0 + \beta_{0,i}{}^0 + \varepsilon$$

Given the optimized coefficients, we predict pseudo control outcome for each sample from treatment group.

$$\hat{V}_i^0 = \sum_{V_k{}^0 \in MB(V_i{}^0)} \beta_{k,i}{}^0 V_k{}^1 + \beta_{0,i}{}^0 + \varepsilon$$

Divergence Estimation

Causal reasoning involves comparing the outcome when a treatment is administrated (V^1) against when it is not (V^0). A causal effect on the outcome is inferred if these outcomes differ ($V^1 \neq V^0$). Individual causal effects are defined as contrasts between counterfactual outcomes, but for each individual, only one of these outcomes can be observed. To approximate the unobserved counterfactual outcome, we estimate a pseudo control outcome based on our understanding of the gene program regulatory network. The divergence score, which quantifies the dysregulation of a feature i within reference network or the causal effect induced by disease or treatment, is calculated by comparing the pseudo control outcome with observed feature values in samples from the treatment group.

$$\text{Divergence Score: } V_i^1 - \hat{V}_i^0 = V_i^1 - \left(\sum_{k \in MB(V_i)} \beta_{k,i}{}^0 V_k{}^1 + \beta_{0,i}{}^0\right)$$

Specifically, gene program divergence score can be defined as $Z_i^1 - \hat{Z}_i^0$.

The divergence score across all variables in the reference network can also be further coupled with *k*-fold cross-validation on the reference samples. This approach assists in selecting the sparsity penalty value of FGES that gives the most informative reference network.

Sample Clustering

Using the *Seurat* package (Hao et al. 2021), we conducted single-sample clustering, comparing clusters based on LaGrACE divergence score and other methods. Firstly, we applied linear dimensional reduction, projecting samples into principal component space. The optimal number of principal components was determined by the JackStraw method or variation between consecutive PCs (Harvard Chan Bioinformatics Core n.d.). The *clustree* package was used to determine appropriate clustering resolution for stable clusters (Zappia and Oshlack 2018). A graph-based clustering was performed with constructed KNN graph based on the Euclidean distance in Principal Component Analysis (PCA) space and then optimized the standard modularity function. Finally, non-linear dimensional reduction (UMAP) was performed to visualize the dataset in a low-dimensional space. For subtyping on synthetic datasets, we used 20 principal components for clustering (to match the *Seurat* default value). In synthetic dataset subtyping, we used 20 principal components with a minimal clustering resolution, ensuring at least three clusters and each cluster accounting for less than 60% of total perturbed samples.

Synthetic Dataset Evaluation

To assess LaGrACE's efficacy in identifying subtypes from mixed-type data, comprising both continuous and discrete variables, we utilized the Lee and Hastie (LH) model (Lee and Hastie 2015). This model delineates the joint distribution over datasets with mixed variables. We first simulated a reference network with a vertex degree of 2 and then created three perturbed networks through random edge modifications (deletion and addition). Each network contained 50 continuous and 25 discrete variables, the latter with four categorical levels. Datasets of 1000 samples were simulated for each network based on the LH model, with continuous variables drawn from $\mathcal{N}(\mu, \sigma)$, where $\mu \sim \mathrm{U}(-1,1)$ and $\sigma \sim \mathrm{U}(0.5,1)$. Edge weights, W_{X1X2}, were assigned from a uniform distribution $\mathrm{U}(0.5,1.5)$. We assumed linear interactions for discrete-continuous and discrete-discrete edges, with parameters modeled according to (Sedgewick et al. 2016). For a comprehensive evaluation, we generated 10 datasets, each with 1000 samples from the reference network and 3000 from the three perturbed networks.

Biologically Inspired Synthetic High-Dimensional Dataset

In the analysis of high-dimensional and collinear omics datasets, such as transcriptomics profiles, we modeled gene expression data using LH data simulated from one reference network and three perturbed networks. This model incorporated 50 continuous variables representing latent factors (Z) (e.g., transcriptional factor (TF) protein activities), and 25 discrete variables, analogous to clinical features. We then simulated 2500 observed variables (X), representing "gene expression", from the 50 latent factors (Z) and their loadings (A) using

$$X = AZ + E$$

where $E \sim \mathcal{N}(0,1)$ denotes independent noise. Each Z factor directly influenced approximately 50 observed features, and A was simulated with weights from a Gaussian distribution $\mathcal{N}(0,1)$. Only 1% of the values in the loading matrix with the largest absolute values were retained, while the rest were set to zero. We generated 10 datasets, each comprising 1000 "normal" samples from the reference network and 3000 "treatment" samples from the three perturbed networks.

Application to Breast Cancer and COPD

Cancer Datasets and Preprocessing

We applied LaGrACE to two cancer datasets containing RNA-seq and clinical data: the METABRIC Dataset (Curtis et al. 2012; Rueda et al. 2019) and the SCAN-B Dataset (Saal et al. 2015). METABRIC data were downloaded from cBioPortal and METABRIC study (Rueda et al. 2019), and SCAN-B data from NCBI GEO (GSE96058). METABRIC dataset was split into several molecular subtypes: basal, claudin-low, HER2+, luminal A, luminal B, and normal breast-like. SCAN-B dataset was comprised of basal group, HER2+ group, luminal A group, and luminal B group. The normal-like group of primary specimens is usually composed of actual normal breast samples, but it also includes a small number of tumor samples. Several studies questioned the existence of this subtype and suspected that the tumor samples in the normal breast-like group may come from (adjacent) normal breast tissue. Thus, the normal-like samples are excluded from downstream analysis (Bernard et al. 2009; Prat et al. 2010; Prat and Perou 2011). Besides, samples from subjects that died of other causes were also excluded.

A common set of 16917 genes was used for both datasets, with Z-score normalization to mitigate batch effects. Six clinical features (breast carcinoma estrogen receptor status, breast carcinoma progesterone receptor status, her2 immunohistochemistry receptor status, age at diagnosis, lymph nodes examined

positive, and tumor size) were included for mixed-type data analysis, and samples with missing values were excluded.

COPD Datasets and Preprocessing

LaGrACE analysis was conducted based on clinical and blood-derived biological features (white blood cell differential percentages, whole blood gene expression) from COPD Genetic Epidemiology (COPDGene) Study (Ghosh et al. 2022). The ECLIPSE study (Vestbo et al. 2008) and scRNA-seq data of human lung (Adams et al. 2020) was used for external validation of our findings.

Data Analysis

Kaplan-Meier survival curves and recurrence events were analyzed using *survival* and *survminer* packages (Kassambara A, Kosinski M, Biecek P 2021; Therneau 2022). Time-dependent AUC curve plots were generated using timeROC package (Kamarudin, Cox, and Kolamunnage-Dona 2017). Discriminatory power of identified clusters was accessed by Harrell's concordance index (C-index), Gönen & Heller's concordance probability estimate (CPE), time-dependent ROC curve (Gönen and Heller 2005; Harrell, Lee, and Mark 1996). Differences in clinical variables among clusters were evaluated using Chi-squared tests for categorical variables and Kruskal Wallis or Wilcoxon rank-sum tests for continuous/ordinal variables. The FCI-stable algorithm was performed using the *rCausalMGM* package (Lovelace, Benos, *in preparation*).

We calculated the epithelial cell differentiation score (Prat et al. 2010), proliferation index (Nielsen et al. 2010), and immune cytolytic activity (Rooney et al. 2015) on all tumor samples. The differentiation score, as a measure of mammary development, was computed using a centroid-based predictor with information from approximately 20,000 gene expression features. The proliferation signature index was derived from 11 genes (BIRC5, CCNB1, CDC20, NUF2, CEP55, NDC80, MKI67, PTTG1, RRM2, TYMS, and UBE2C), and immune cytolytic activity from the geometric mean of GZMA and PRF1 expression.

Pathway enrichment was assessed through ssGSEA (Barbie et al. 2009), MSigBD gene sets (Liberzon et al. 2015), and *GSVA* package (Hänzelmann, Castelo, and Guinney 2013). Transcriptional factor regulon scores were inferred using *decoupleR* and *CollecTRI* packages (Badia-i-Mompel et al. 2022; Müller-Dott et al. 2023). Cell type fractions in bulk transcriptome profiles were deconvoluted using CIBERSORTx (Newman et al. 2019) based on immune population gene signature matrix (LM22) and breast cancer single-cell profiles (Wu et al. 2021). Enrichr (Xie et al. 2021) was employed to find overlaps between genes assigned to latent factor of interest and precompiled ontologies.

COPD module scores on T lymphocyte from human lung were computed based on differentially expressed genes associated with COPD, identified in memory CD4 T cells from blood using *AddModuleScore* function from *Seurat* package. To control for effect of aging, the comparison is restricted to samples from individuals aged 55 years and older.

Assignment of Cluster Membership to New Samples

To assign cluster labels to the validation dataset based on LaGrACE features, we trained a random forest classifier with LaGrACE features from the METABRIC dataset using *randomForest* package (Wiener 2002). Parameters were tuned using 10-fold cross-validation with the *caret* and *mlbench* package (Dimitriadou 2021; Kuhn 2008). To minimize batch effects, we calibrated the model learned from the METABRIC

dataset with basal samples (reference group) from the SCAN-B dataset, then calculated LaGrACE features for samples from the SCAN-B dataset.

Application to Single Cell Drug Response Dataset

To evaluate LaGrACE suitability for single-cell data, we used two public scRNA-seq datasets: GSE117089, featuring a time series of dexamethasone treatment on A549 human lung adenocarcinoma cells (Cao et al. 2018); and GSE139944, containing data on A549 cells treated with increasing doses of small molecules (BMS345541, nutlin-3A, SAHA, or DMSO vehicle control) for 24 hours (Srivatsan et al. 2020).

Data Analysis

Both scRNA-seq datasets were processed using the *Seurat* pipeline. The deep generative model scVI (Lopez et al. 2018) was applied to impute the top 3000 highly variable genes from each dataset. In the LaGrACE analysis, the vehicle group was selected as the reference group.

Method Comparison

Pathway Activity Quantification Method

Pathifier and ssGSEA were used for bulk RNA-seq datasets. We used the MSigBD C2 gene set collection for ssGSEA. For Pathifier, we applied KEGG pathway information from *KEGGREST* (Tenenbaum D 2022) and the *pathifier* package for pathway deregulation scoring.

Single cell Gene Program Score Method

To run Spectra, expression data was normalized using sran. We used the default parameters with 151 default Spectra global genesets. For expiMap, we used default parameters with Reactome genesets used in expiMap tutorial.

Evaluation metrics

Clustering Evaluation

Clustering performance was assessed using the Adjusted Rand Index (ARI) and Adjusted Mutual Information (AMI) via the *aricode* package (Julien Chiquet, Guillem Rigaill, Martina Sundqvist 2020).

Single Cell Neighborhood Evaluation

For performance evaluation on scRNA-seq datasets without relying on clustering, Milo (Dann et al. 2022) was used to model cell states as neighborhoods on a KNN graph. This graph was constructed using PCA based on input features, such as RNA or LaGrACE features. Cell type purity for each neighborhood, defining it as the frequency of the most prevalent cell type, was calculated utilizing dose information or treatment time.

Cell Culture

We cultured MDA-MB-231 cells in Dulbecco's modified eagle medium (DMEM, Gibco 11995) supplemented with 10% fetal bovine serum (FBS, Gibco 16000), 1% GlutaMax (Gibco 35050), 1% penicillin/streptomycin (pen/strep, Gibco 15070), and 0.1% of plasmocin (InvivoGen ant-mpp). We cultured SUM149 and SUM159 cells in F-12 (Gibco 11765) media supplemented with 5% FBS (Gibco 16000), 1% pen/strep (Gibco 15070), 1% GlutaMax (Gibco 35050), 1 µg mL−1 hydrocortisone (Sigma H4001), and 5 µg mL−1 insulin (Sigma I6634), and 0.1% of plasmocin (InvivoGen, ant-mpp). These breast

cancer cells were obtained from Dr. Gary Luker's lab at the University of Michigan. We maintained all cells at 37 °C in a humidified incubator with 5% CO2. All the cells were cultured and passaged when the cells reached over 80% confluency in the dish.

Cell Transfection

We transfected breast cancer cells using Xfect™ Transfection Reagent (Takara 631317) with 5 µg of pEF1alpha-tdTomato Vector plasmid (Takara 631975). The cells labelled with red fluorescent proteins can be tracked for migration. The transfected cells were selected using G418 (Takara 631307) treatment and sorted by flow cytometry for red fluorescence.

High-Throughput Single-Cell Migration Assay by Microfluidics

The cell migration assays were performed using a microfluidic single-cell migration platform modified from our previous work (Chen et al. 2018, 2019; Zhou et al. 2023). The migration channel is designed to be 5 µm in height and 10 µm in width. The device was primed by a collagen solution (1.45mL Collagen (Collagen Type 1, 354236, BD Biosciences) and 0.1mL acetic acid in 50mL DI Water) to enhance cell adhesion. Right before the migration experiment, the devices were rinsed by cell culture media. Breast cancer Cells were trypsinized, centrifuged, and re-suspended to a concentration of 4 × 105 cells/mL for loading onto the devices. The devices were incubated for 30 min to enhance cell adhesion. FOXM1 inhibitors (Thiostrepton, 10 µM (Cat#: 19200), FDI-6, 10 µM (HY-112721), RCM-1, 10 µM (HY-19979)) were used to inhibit cells motility.

The microfluidic devices were incubated for 24 hours, and the migration distance was measured based on the final cell position. The images were analyzed by a custom MATLAB code [3]. Cells were identified based on their fluorescence, and debris were ignored by their small size. In this work, 2 independent replicates (~200 migration channels) were performed. The Violin graphs were plotted using Prism (v.9.3).

As cell migration results do not follow normal distribution, a non-parametric Mann-Whitney U test was used for comparisons of cell motility with a significance level of 0.05 considered statistically significant. * refers to $P < 0.05$, ** refers to $P < 0.01$, and *** refers to $P < 0.001$.

Data availability

The following publicly available datasets were used in the manuscript: the METABRIC dataset (https://www.cbioportal.org/study/summary?id=brca_metabric), the SCAN-B dataset (GEO: GSE96058), breast cancer single-cell dataset (https://doi.org/10.5281/zenodo.4739739), single cell drug response dataset (GEO: GSE117089 & GSE139944), the COPD dataset (GEO: GSE76705 & http://www.copdgene.org).

Code availability

LaGrACE is available as a R Package at https://github.com/Lobbii/LaGrACE.

REFERENCES

Adams, Taylor S., Jonas C. Schupp, Sergio Poli, Ehab A. Ayaub, Nir Neumark, Farida Ahangari, Sarah G. Chu, Benjamin A. Raby, Giuseppe DeIuliis, Michael Januszyk, Qiaonan Duan, Heather A. Arnett, Asim Siddiqui, George R. Washko, Robert Homer, Xiting Yan, Ivan O. Rosas, and Naftali Kaminski. 2020. 'Single-Cell RNA-Seq Reveals Ectopic and Aberrant Lung-Resident Cell Populations in Idiopathic Pulmonary Fibrosis'. *Science Advances* 6(28). doi: 10.1126/sciadv.aba1983.

Andrews, Bryan, Joseph Ramsey, and Gregory F. Cooper. 2019. 'Learning High-Dimensional Directed Acyclic Graphs with Mixed Data-Types'. *Proc Mach Learn Res* 104(4–21).

Badia-i-Mompel, Pau, Jesús Vélez Santiago, Jana Braunger, Celina Geiss, Daniel Dimitrov, Sophia Müller-Dott, Petr Taus, Aurelien Dugourd, Christian H. Holland, Ricardo O. Ramirez Flores, and Julio Saez-Rodriguez. 2022. 'DecoupleR: Ensemble of Computational Methods to Infer Biological Activities from Omics Data'. *Bioinformatics Advances* 2(1):vbac016. doi: 10.1093/bioadv/vbac016.

Barbie, David A., Pablo Tamayo, Jesse S. Boehm, So Young Kim, Susan E. Moody, Ian F. Dunn, Anna C. Schinzel, Peter Sandy, Etienne Meylan, Claudia Scholl, Stefan Fröhling, Edmond M. Chan, Martin L. Sos, Kathrin Michel, Craig Mermel, Serena J. Silver, Barbara A. Weir, Jan H. Reiling, Qing Sheng, Piyush B. Gupta, Raymond C. Wadlow, Hanh Le, Sebastian Hoersch, Ben S. Wittner, Sridhar Ramaswamy, David M. Livingston, David M. Sabatini, Matthew Meyerson, Roman K. Thomas, Eric S. Lander, Jill P. Mesirov, David E. Root, D. Gary Gilliland, Tyler Jacks, and William C. Hahn. 2009. 'Systematic RNA Interference Reveals That Oncogenic KRAS-Driven Cancers Require TBK1'. *Nature* 462(7269):108–12. doi: 10.1038/nature08460.

Bernard, Philip S., Joel S. Parker, Michael Mullins, Maggie C. U. Cheung, Samuel Leung, David Voduc, Tammi Vickery, Sherri Davies, Christiane Fauron, Xiaping He, Zhiyuan Hu, John F. Quackenhush, Inge J. Stijleman, Juan Palazzo, J. S. Matron, Andrew B. Nobel, Elaine Mardis, Torsten O. Nielsen, Matthew J. Ellis, and Charles M. Perou. 2009. 'Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes'. *Journal of Clinical Oncology*. doi: 10.1200/JCO.2008.18.1370.

Bing, Xin, Florentina Bunea, Martin Royer, and Jishnu Das. 2019. 'Latent Model-Based Clustering for Biological Discovery.' *IScience* 14:125–35. doi: 10.1016/j.isci.2019.03.018.

Cao, Junyue, Darren A. Cusanovich, Vijay Ramani, Delasa Aghamirzaie, Hannah A. Pliner, Andrew J. Hill, Riza M. Daza, Jose L. McFaline-Figueroa, Jonathan S. Packer, Lena Christiansen, Frank J. Steemers, Andrew C. Adey, Cole Trapnell, and Jay Shendure. 2018. 'Joint Profiling of Chromatin Accessibility and Gene Expression in Thousands of Single Cells'. *Science* 361(6409):1380–85. doi: 10.1126/science.aau0730.

Chen, Yaohui, Yu Li, Jianfei Xue, Aihua Gong, Guanzhen Yu, Aidong Zhou, Kangyu Lin, Sicong Zhang, Nu Zhang, Cara J. Gottardi, and Suyun Huang. 2016. 'Wnt-Induced Deubiquitination FoxM1 Ensures Nucleus β-Catenin Transactivation.' *The EMBO Journal* 35(6):668–84. doi: 10.15252/embj.201592810.

Chen, Yu-Chih, Brock Humphries, Riley Brien, Anne E. Gibbons, Yu-Ting Chen, Tonela Qyli, Henry R. Haley, Matthew E. Pirone, Benjamin Chiang, Annie Xiao, Yu-Heng Cheng, Yi Luan, Zhixiong Zhang, Jason Cong, Kathryn E. Luker, Gary D. Luker, and Euisik Yoon. 2018. 'Functional Isolation of Tumor-Initiating Cells Using Microfluidic-Based Migration Identifies Phosphatidylserine Decarboxylase as a Key Regulator.' *Scientific Reports* 8(1):244. doi: 10.1038/s41598-017-18610-5.

Chen, Yu-Chih, Saswat Sahoo, Riley Brien, Seungwon Jung, Brock Humphries, Woncheol Lee, Yu-Heng Cheng, Zhixiong Zhang, Kathryn E. Luker, Max S. Wicha, Gary D. Luker, and Euisik Yoon. 2019. 'Single-Cell RNA-Sequencing of Migratory Breast Cancer Cells: Discovering Genes Associated with Cancer Metastasis.' *The Analyst* 144(24):7296–7309. doi: 10.1039/c9an01358j.

Curtis, Christina, Sohrab P. Shah, Suet-Feung Chin, Gulisa Turashvili, Oscar M. Rueda, Mark J. Dunning, Doug Speed, Andy G. Lynch, Shamith Samarajiwa, Yinyin Yuan, Stefan Gräf, Gavin Ha, Gholamreza Haffari, Ali Bashashati, Roslin Russell, Steven McKinney, Carlos Caldas, Samuel Aparicio, Christina Curtis†, Sohrab P. Shah, Carlos Caldas, Samuel Aparicio, James D. Brenton, Ian Ellis, David Huntsman, Sarah Pinder, Arnie Purushotham, Leigh Murphy, Carlos Caldas, Samuel Aparicio, Carlos Caldas, Helen Bardwell, Suet-Feung Chin, Christina Curtis, Zhihao Ding, Stefan Gräf, Linda Jones, Bin Liu, Andy G. Lynch, Irene Papatheodorou, Stephen J. Sammut, Gordon Wishart, Samuel Aparicio, Steven Chia, Karen Gelmon, David Huntsman, Steven McKinney,

Caroline Speers, Gulisa Turashvili, Peter Watson, Ian Ellis, Roger Blamey, Andrew Green, Douglas Macmillan, Emad Rakha, Arnie Purushotham, Cheryl Gillett, Anita Grigoriadis, Sarah Pinder, Emanuele de Rinaldis, Andy Tutt, Leigh Murphy, Michelle Parisien, Sandra Troup, Carlos Caldas, Suet-Feung Chin, Derek Chan, Claire Fielding, Ana-Teresa Maia, Sarah McGuire, Michelle Osborne, Sara M. Sayalero, Inmaculada Spiteri, James Hadfield, Samuel Aparicio, Gulisa Turashvili, Lynda Bell, Katie Chow, Nadia Gale, David Huntsman, Maria Kovalik, Ying Ng, Leah Prentice, Carlos Caldas, Simon Tavaré, Christina Curtis, Mark J. Dunning, Stefan Gräf, Andy G. Lynch, Oscar M. Rueda, Roslin Russell, Shamith Samarajiwa, Doug Speed, Florian Markowetz, Yinyin Yuan, James D. Brenton, Samuel Aparicio, Sohrab P. Shah, Ali Bashashati, Gavin Ha, Gholamreza Haffari, Steven McKinney, Anita Langerød, Andrew Green, Elena Provenzano, Gordon Wishart, Sarah Pinder, Peter Watson, Florian Markowetz, Leigh Murphy, Ian Ellis, Arnie Purushotham, Anne-Lise Børresen-Dale, James D. Brenton, Simon Tavaré, Carlos Caldas, Samuel Aparicio, METABRIC Group, Co-chairs, Writing committee, Steering committee, Tissue and clinical data source sites:, University of Cambridge/Cancer Research U. K. Cambridge Research Institute, British Columbia Cancer Agency, University of Nottingham, King's College London, Manitoba Institute of Cell Biology, Cancer genome/transcriptome characterization centres:, and Data analysis subgroup: 2012. 'The Genomic and Transcriptomic Architecture of 2,000 Breast Tumours Reveals Novel Subgroups'. *Nature* 486(7403):346–52. doi: 10.1038/nature10983.

Dann, Emma, Neil C. Henderson, Sarah A. Teichmann, Michael D. Morgan, and John C. Marioni. 2022. 'Differential Abundance Testing on Single-Cell Data Using k-Nearest Neighbor Graphs'. *Nature Biotechnology* 40(2):245–53. doi: 10.1038/s41587-021-01033-z.

Davidson, Sarah, Mirjana Efremova, Angela Riedel, Bidesh Mahata, Jhuma Pramanik, Jani Huuhtanen, Gozde Kar, Roser Vento-Tormo, Tzachi Hagai, Xi Chen, Muzlifah A. Haniffa, Jacqueline D. Shields, and Sarah A. Teichmann. 2020. 'Single-Cell RNA Sequencing Reveals a Dynamic Stromal Niche That Supports Tumor Growth.' *Cell Reports* 31(7):107628. doi: 10.1016/j.celrep.2020.107628.

Dimitriadou, Friedrich Leisch and Evgenia. 2021. 'Mlbench: Machine Learning Benchmark Problems'.

Elyahu, Yehezqel, Idan Hekselman, Inbal Eizenberg-Magar, Omer Berner, Itai Strominger, Maya Schiller, Kritika Mittal, Anna Nemirovsky, Ekaterina Eremenko, Assaf Vital, Eyal Simonovsky, Vered Chalifa-Caspi, Nir Friedman, Esti Yeger-Lotem, and Alon Monsonego. 2019. 'Aging Promotes Reorganization of the CD4 T Cell Landscape toward Extreme Regulatory and Effector Phenotypes'. *Science Advances* 5(8). doi: 10.1126/sciadv.aaw8330.

Feder, Kristin, Katrin Edmaier-Schröger, Vijay P. S. Rawat, Nicole Kirsten, Klaus Metzeler, Johann M. Kraus, Konstanze Döhner, Hartmut Döhner, Hans A. Kestler, Michaela Feuring-Buske, and Christian Buske. 2020. 'Differences in Expression and Function of LEF1 Isoforms in Normal versus Leukemic Hematopoiesis'. *Leukemia* 34(4):1027–37. doi: 10.1038/s41375-019-0635-1.

Freedman, D., L. Gates, W. Flanders, O. Van Assendelft, J. Barboriak, M. Joesoef, and T. Byers. 1997. 'Black/White Differences in Leukocyte Subpopulations in Men'. *International Journal of Epidemiology* 26(4):757–64. doi: 10.1093/ije/26.4.757.

Friedman, Nir. 2004. 'Inferring Cellular Networks Using Probabilistic Graphical Models'. *Science* 303(5659):799–805. doi: 10.1126/science.1094068.

Ghosh, Auyon J., Aabida Saferali, Sool Lee, Robert Chase, Matthew Moll, Jarrett Morrow, Jeong Yun, Peter J. Castaldi, and Craig P. Hersh. 2022. 'Blood RNA Sequencing Shows Overlapping Gene Expression across COPD Phenotype Domains'. *Thorax* 77(2):115–22. doi: 10.1136/thoraxjnl-2020-216401.

Gönen, Mithat, and Glenn Heller. 2005. 'Concordance Probability and Discriminatory Power in Proportional Hazards Regression'. *Biometrika* 92(4):965–70. doi: 10.1093/biomet/92.4.965.

González-Mariscal, Lorenza, Jael Miranda, Helios Gallego-Gutiérrez, Misael Cano-Cortina, and Elida Amaya. 2020. 'Relationship between Apical Junction Proteins, Gene Expression and Cancer'. *Biochimica et Biophysica Acta (BBA) - Biomembranes* 1862(9):183278. doi: 10.1016/j.bbamem.2020.183278.

Greenberg, Alissa K., Jing Hu, Sharmila Basu, John Hay, Joan Reibman, Ting-an Yie, Kam Meng Tchou-Wong, William N. Rom, and Theodore C. Lee. 2002. 'Glucocorticoids Inhibit Lung Cancer Cell Growth through Both the Extracellular Signal-Related Kinase Pathway and Cell Cycle Regulators'. *American Journal of Respiratory Cell and Molecular Biology* 27(3):320–28. doi: 10.1165/rcmb.4710.

Hänzelmann, Sonja, Robert Castelo, and Justin Guinney. 2013. 'GSVA: Gene Set Variation Analysis for Microarray and RNA-Seq Data'. *BMC Bioinformatics*. doi: 10.1186/1471-2105-14-7.

Hao, Yuhan, Stephanie Hao, Erica Andersen-Nissen, William M. Mauck, Shiwei Zheng, Andrew Butler, Maddie J. Lee, Aaron J. Wilk, Charlotte Darby, Michael Zager, Paul Hoffman, Marlon Stoeckius, Efthymia Papalexi, Eleni P. Mimitou, Jaison Jain, Avi Srivastava, Tim Stuart, Lamar M. Fleming, Bertrand Yeung, Angela J. Rogers, Juliana M. McElrath, Catherine A. Blish, Raphael Gottardo, Peter Smibert, and Rahul Satija. 2021. 'Integrated Analysis of Multimodal Single-Cell Data'. *Cell* 184(13):3573-3587.e29. doi: https://doi.org/10.1016/j.cell.2021.04.048.

Harrell, F. E. Jr, K. L. Lee, and D. B. Mark. 1996. 'Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors.' *Statistics in Medicine* 15(4):361–87. doi: 10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4.

Harris, Benjamin D., Megan Crow, Stephan Fischer, and Jesse Gillis. 2021. 'Single-Cell Co-Expression Analysis Reveals That Transcriptional Modules Are Shared across Cell Types in the Brain'. *Cell Systems* 12(7):748-756.e3. doi: 10.1016/j.cels.2021.04.010.

Harvard Chan Bioinformatics Core. n.d. 'Single-Cell RNA-Seq Analysis Workshop'. Retrieved (https://hbctraining.github.io/scRNA-seq/lessons/elbow_plot_metric.html).

Heikenwalder, Mathias, and Anna Lorentzen. 2019. 'The Role of Polarisation of Circulating Tumour Cells in Cancer Metastasis'. *Cellular and Molecular Life Sciences* 76(19):3765–81. doi: 10.1007/s00018-019-03169-3.

Hill, L. D., L. Sun, M. P. Leuschen, and T. L. Zach. 1993. 'C3 Synthesis by A549 Alveolar Epithelial Cells Is Increased by Interferon-Gamma and Dexamethasone.' *Immunology* 79(2):236–40.

Jia, Minxue, Khaled Sayed, Maria G. Kapetanaki, William Dion, Lorena Rosas, Saad Irfan, Eleanor Valenzi, Ana L. Mora, Robert A. Lafyatis, Mauricio Rojas, Bokai Zhu, and Panayiotis V. Benos. 2023. '<scp>LEF1</Scp> Isoforms Regulate Cellular Senescence and Aging'. *Aging Cell* 22(12). doi: 10.1111/acel.14024.

Johnson, J., B. Thijssen, U. McDermott, M. Garnett, L. F. A. Wessels, and R. Bernards. 2016. 'Targeting the RB-E2F Pathway in Breast Cancer'. *Oncogene* 35(37):4829–35. doi: 10.1038/onc.2016.32.

Julien Chiquet, Guillem Rigaill, Martina Sundqvist, Valentin Dervieux. 2020. 'Aricode: Efficient Computations of Standard Clustering Comparison Measures'.

Kamarudin, Adina Najwa, Trevor Cox, and Ruwanthi Kolamunnage-Dona. 2017. 'Time-Dependent ROC Curve Analysis in Medical Research: Current Methods and Applications'. *BMC Medical Research Methodology* 17(1):53. doi: 10.1186/s12874-017-0332-6.

Kassambara A, Kosinski M, Biecek P, Fabian S. 2021. 'Survminer: Drawing Survival Curves Using 'ggplot2''.

Kuhn, Max. 2008. 'Building Predictive Models in R Using the Caret Package'. *Journal of Statistical Software* 28(5 SE-Articles):1–26. doi: 10.18637/jss.v028.i05.

Kunes, Russell Z., Thomas Walle, Max Land, Tal Nawy, and Dana Pe'er. 2023. 'Supervised Discovery of Interpretable Gene Programs from Single-Cell Data'. *Nature Biotechnology*. doi: 10.1038/s41587-023-01940-3.

Lambert, Samuel A., Arttu Jolma, Laura F. Campitelli, Pratyush K. Das, Yimeng Yin, Mihai Albu, Xiaoting Chen, Jussi Taipale, Timothy R. Hughes, and Matthew T. Weirauch. 2018. 'The Human Transcription Factors'. *Cell* 172(4):650–65. doi: 10.1016/j.cell.2018.01.029.

Lee, Jason D., and Trevor J. Hastie. 2015. 'Learning the Structure of Mixed Graphical Models'. *Journal of Computational and Graphical Statistics*. doi: 10.1080/10618600.2014.900500.

Lee, Tong Ihn, and Richard A. Young. 2013. 'Transcriptional Regulation and Its Misregulation in Disease'. *Cell* 152(6):1237–51. doi: 10.1016/j.cell.2013.02.014.

Liberzon, Arthur, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P. Mesirov, and Pablo Tamayo. 2015. 'The Molecular Signatures Database Hallmark Gene Set Collection'. *Cell Systems*. doi: 10.1016/j.cels.2015.12.004.

Liu, Han, Kathryn Roeder, and Larry Wasserman. 2010. 'Stability Approach to Regularization Selection (StARS) for High Dimensional Graphical Models'. in *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010, NIPS 2010*.

Lopez, Romain, Jeffrey Regier, Michael B. Cole, Michael I. Jordan, and Nir Yosef. 2018. 'Deep Generative Modeling for Single-Cell Transcriptomics'. *Nature Methods* 15(12):1053–58. doi: 10.1038/s41592-018-0229-2.

Lotfollahi, Mohammad, Sergei Rybakov, Karin Hrovatin, Soroor Hediyeh-zadeh, Carlos Talavera-López, Alexander V. Misharin, and Fabian J. Theis. 2023. 'Biologically Informed Deep Learning to Query Gene Programs in Single-Cell Atlases'. *Nature Cell Biology*. doi: 10.1038/s41556-022-01072-x.

Mancini, M., S. De Santis, C. Monaldi, L. Bavaro, M. Martelli, F. Castagnetti, G. Gugliotta, G. Rosti, M. A. Santucci, G. Martinelli, M. Cavo, and S. Soverini. 2019. 'Hyper-Activation of Aurora Kinase a-Polo-like Kinase 1-FOXM1 Axis Promotes Chronic Myeloid Leukemia Resistance to Tyrosine Kinase Inhibitors'. *Journal of Experimental & Clinical Cancer Research* 38(1):216. doi: 10.1186/s13046-019-1197-9.

Müller-Dott, Sophia, Eirini Tsirvouli, Miguel Vazquez, Ricardo O. Ramirez Flores, Pau Badia-i-Mompel, Robin Fallegger, Dénes Türei, Astrid Lægreid, and Julio Saez-Rodriguez. 2023. 'Expanding the Coverage of Regulons from High-Confidence Prior Knowledge for Accurate Estimation of Transcription Factor Activities'. *Nucleic Acids Research* 51(20):10934–49. doi: 10.1093/nar/gkad841.

Nédélec, Yohann, Joaquín Sanz, Golshid Baharian, Zachary A. Szpiech, Alain Pacis, Anne Dumaine, Jean-Christophe Grenier, Andrew Freiman, Aaron J. Sams, Steven Hebert, Ariane Pagé Sabourin, Francesca Luca, Ran Blekhman, Ryan D. Hernandez, Roger Pique-Regi, Jenny Tung, Vania Yotova, and Luis B. Barreiro. 2016. 'Genetic Ancestry and Natural Selection Drive Population Differences in Immune Responses to Pathogens'. *Cell* 167(3):657-669.e21. doi: 10.1016/j.cell.2016.09.025.

Newman, Aaron M., Chloé B. Steen, Chih Long Liu, Andrew J. Gentles, Aadel A. Chaudhuri, Florian Scherer, Michael S. Khodadoust, Mohammad S. Esfahani, Bogdan A. Luca, David Steiner, Maximilian Diehn, and Ash A. Alizadeh. 2019. 'Determining Cell Type Abundance and Expression from Bulk Tissues with Digital Cytometry'. *Nature Biotechnology* 37(7):773–82. doi: 10.1038/s41587-019-0114-2.

Nielsen, Torsten O., Joel S. Parker, Samuel Leung, David Voduc, Mark Ebbert, Tammi Vickery, Sherri R. Davies, Jacqueline Snider, Inge J. Stijleman, Jerry Reed, Maggie C. U. Cheang, Elaine R. Mardis, Charles M. Perou, Philip S. Bernard, and Matthew J. Ellis. 2010. 'A Comparison of PAM50 Intrinsic Subtyping with Immunohistochemistry and Clinical Prognostic Factors in Tamoxifen-Treated Estrogen Receptor-Positive Breast Cancer.' *Clinical Cancer Research : An Official Journal of the American Association for Cancer Research* 16(21):5222–32. doi: 10.1158/1078-0432.CCR-10-1282.

Oshi, Masanori, Hideo Takahashi, Yoshihisa Tokumaru, Li Yan, Omar M. Rashid, Masayuki Nagahashi, Ryusei Matsuyama, Itaru Endo, and Kazuaki Takabe. 2020. 'The E2F Pathway Score as a Predictive Biomarker of Response to Neoadjuvant Therapy in ER+/HER2− Breast Cancer'. *Cells* 9(7):1643. doi: 10.3390/cells9071643.

Paliogiannis, Panagiotis, Alessandro G. Fois, Salvatore Sotgia, Arduino A. Mangoni, Elisabetta Zinellu, Pietro Pirina, Silvia Negri, Ciriaco Carru, and Angelo Zinellu. 2018. 'Neutrophil to Lymphocyte Ratio and Clinical Outcomes in COPD: Recent Evidence and Future Perspectives'. *European Respiratory Review* 27(147):170113. doi: 10.1183/16000617.0113-2017.

Pan, W., and F. Doshi-Velez. 2016. 'A Characterization of the Non-Uniqueness of Nonnegative Matrix Factorizations'.

Prat, Aleix, Joel S. Parker, Olga Karginova, Cheng Fan, Chad Livasy, Jason I. Herschkowitz, Xiaping He, and Charles M. Perou. 2010. 'Phenotypic and Molecular Characterization of the Claudin-Low Intrinsic Subtype of Breast Cancer'. *Breast Cancer Research*. doi: 10.1186/bcr2635.

Prat, Aleix, and Charles M. Perou. 2011. 'Deconstructing the Molecular Portraits of Breast Cancer'. *Molecular Oncology*.

Ramsey, Joseph, Madelyn Glymour, Ruben Sanchez-Romero, and Clark Glymour. 2017. 'A Million Variables and More: The Fast Greedy Equivalence Search Algorithm for Learning High-Dimensional Graphical Causal Models, with an Application to Functional Magnetic Resonance Images'. *International Journal of Data Science and Analytics*. doi: 10.1007/s41060-016-0032-z.

Rooney, Michael S., Sachet A. Shukla, Catherine J. Wu, Gad Getz, and Nir Hacohen. 2015. 'Molecular and Genetic Properties of Tumors Associated with Local Immune Cytolytic Activity'. *Cell* 160(1):48–61. doi: https://doi.org/10.1016/j.cell.2014.12.033.

Rueda, Oscar M., Stephen-John Sammut, Jose A. Seoane, Suet-Feung Chin, Jennifer L. Caswell-Jin, Maurizio Callari, Rajbir Batra, Bernard Pereira, Alejandra Bruna, H. Raza Ali, Elena Provenzano, Bin Liu, Michelle Parisien, Cheryl Gillett, Steven McKinney, Andrew R. Green, Leigh Murphy, Arnie Purushotham, Ian O. Ellis, Paul D.

Pharoah, Cristina Rueda, Samuel Aparicio, Carlos Caldas, and Christina Curtis. 2019. 'Dynamics of Breast-Cancer Relapse Reveal Late-Recurring ER-Positive Genomic Subgroups'. *Nature* 567(7748):399–404. doi: 10.1038/s41586-019-1007-8.

Rye, Inga Hansine, Kanutte Huse, Sarah E. Josefsson, Wanja Kildal, Håvard E. Danielsen, Ellen Schlichting, Øystein Garred, Margit L. Riis, Osbreac, Ole Christian Lingjaerde, June H. Myklebust, and Hege G. Russnes. 2022. 'Breast Cancer Metastasis: Immune Profiling of Lymph Nodes Reveals Exhaustion of Effector T Cells and Immunosuppression.' *Molecular Oncology* 16(1):88–103. doi: 10.1002/1878-0261.13047.

Saal, Lao H., Johan Vallon-Christersson, Jari Häkkinen, Cecilia Hegardt, Dorthe Grabau, Christof Winter, Christian Brueffer, Man-Hung Eric Tang, Christel Reuterswärd, Ralph Schulz, Anna Karlsson, Anna Ehinger, Janne Malina, Jonas Manjer, Martin Malmberg, Christer Larsson, Lisa Rydén, Niklas Loman, and Åke Borg. 2015. 'The Sweden Cancerome Analysis Network - Breast (SCAN-B) Initiative: A Large-Scale Multicenter Infrastructure towards Implementation of Breast Cancer Genomic Analyses in the Clinical Routine'. *Genome Medicine* 7(1):20. doi: 10.1186/s13073-015-0131-9.

Sedgewick, Andrew J., Ivy Shi, Rory M. Donovan, and Panayiotis V. Benos. 2016. 'Learning Mixed Graphical Models with Separate Sparsity Parameters and Stability-Based Model Selection'. *BMC Bioinformatics*. doi: 10.1186/s12859-016-1039-0.

Srivatsan, Sanjay R., José L. McFaline-Figueroa, Vijay Ramani, Lauren Saunders, Junyue Cao, Jonathan Packer, Hannah A. Pliner, Dana L. Jackson, Riza M. Daza, Lena Christiansen, Fan Zhang, Frank Steemers, Jay Shendure, and Cole Trapnell. 2020. 'Massively Multiplex Chemical Transcriptomics at Single-Cell Resolution'. *Science* 367(6473):45–51. doi: 10.1126/science.aax6234.

Steinke, Farrah C., Shuyang Yu, Xinyuan Zhou, Bing He, Wenjing Yang, Bo Zhou, Hiroshi Kawamoto, Jun Zhu, Kai Tan, and Hai-Hui Xue. 2014. 'TCF-1 and LEF-1 Act Upstream of Th-POK to Promote the CD4+ T Cell Fate and Interact with Runx3 to Silence Cd4 in CD8+ T Cells'. *Nature Immunology* 15(7):646–56. doi: 10.1038/ni.2897.

Tenenbaum D, Maintainer B. 2022. 'KEGGREST: Client-Side REST Access to the Kyoto Encyclopedia of Genes and Genomes (KEGG)'. doi: https://doi.org/doi:10.18129/B9.bioc.KEGGREST.

Therneau, Terry M. 2022. 'A Package for Survival Analysis in R'.

Urban, Gudrun, Teresa Golden, Ileana V. Aragon, Lex Cowsert, Scott R. Cooper, Nicholas M. Dean, and Richard E. Honkanen. 2003. 'Identification of a Functional Link for the P53 Tumor Suppressor Protein in Dexamethasone-Induced Growth Suppression'. *Journal of Biological Chemistry* 278(11):9747–53. doi: 10.1074/jbc.M210993200.

Vestbo, J., W. Anderson, H. O. Coxson, C. Crim, F. Dawber, L. Edwards, G. Hagan, K. Knobil, D. A. Lomas, W. MacNee, E. K. Silverman, and R. Tal-Singer. 2008. 'Evaluation of COPD Longitudinally to Identify Predictive Surrogate End-Points (ECLIPSE)'. *European Respiratory Journal* 31(4):869–73. doi: 10.1183/09031936.00111707.

Wang, Zhen, Zhu Yu, Gong-He Wang, Yi-Ming Zhou, Jian-Ping Deng, Yue Feng, Jun-Qiang Chen, and Lei Tian. 2020. 'AURKB Promotes the Metastasis of Gastric Cancer, Possibly by Inducing EMT'. *Cancer Management and Research* 12:6947–58. doi: 10.2147/CMAR.S254250.

van de Wetering, Marc, Wim de Lau, and Hans Clevers. 2002. 'WNT Signaling and Lymphocyte Development'. *Cell* 109(2):S13–19. doi: 10.1016/S0092-8674(02)00709-2.

Wiener, Andy Liaw and Matthew. 2002. 'Classification and Regression by RandomForest'. *R News* 2.

Wu, Sunny Z., Ghamdan Al-Eryani, Daniel Lee Roden, Simon Junankar, Kate Harvey, Alma Andersson, Aatish Thennavan, Chenfei Wang, James R. Torpy, Nenad Bartonicek, Taopeng Wang, Ludvig Larsson, Dominik Kaczorowski, Neil I. Weisenfeld, Cedric R. Uytingco, Jennifer G. Chew, Zachary W. Bent, Chia-Ling Chan, Vikkitharan Gnanasambandapillai, Charles-Antoine Dutertre, Laurence Gluch, Mun N. Hui, Jane Beith, Andrew Parker, Elizabeth Robbins, Davendra Segara, Caroline Cooper, Cindy Mak, Belinda Chan, Sanjay Warrier, Florent Ginhoux, Ewan Millar, Joseph E. Powell, Stephen R. Williams, X. Shirley Liu, Sandra O'Toole, Elgene Lim, Joakim Lundeberg, Charles M. Perou, and Alexander Swarbrick. 2021. 'A Single-Cell and Spatially Resolved Atlas of Human Breast Cancers'. *Nature Genetics* 53(9):1334–47. doi: 10.1038/s41588-021-00911-1.

Xie, Zhuorui, Allison Bailey, Maxim V. Kuleshov, Daniel J. B. Clarke, John E. Evangelista, Sherry L. Jenkins, Alexander Lachmann, Megan L. Wojciechowicz, Eryk Kropiwnicki, Kathleen M. Jagodnik, Minji Jeon, and Avi Ma'ayan. 2021. 'Gene Set Knowledge Discovery with Enrichr'. *Current Protocols* 1(3). doi: 10.1002/cpz1.90.

Xu, Xiufang, Miaofeng Zhang, Faying Xu, and Shaojie Jiang. 2020. 'Wnt Signaling in Breast Cancer: Biological Mechanisms, Challenges and Opportunities'. *Molecular Cancer* 19(1):165. doi: 10.1186/s12943-020-01276-5.

Yotam, Drier, Sheffer Michal, and Domany Eytan. 2013. 'Pathway-Based Personalized Analysis of Cancer'. *Proceedings of the National Academy of Sciences* 110(16):6388–93. doi: 10.1073/pnas.1219651110.

Zappia, Luke, and Alicia Oshlack. 2018. 'Clustering Trees: A Visualization for Evaluating Clusterings at Multiple Resolutions'. *GigaScience* 7(7):giy083. doi: 10.1093/gigascience/giy083.

Zhang, Nu, Ping Wei, Aihua Gong, Wen-Tai Chiu, Hsueh-Te Lee, Howard Colman, He Huang, Jianfei Xue, Mingguang Liu, Yong Wang, Raymond Sawaya, Keping Xie, W. K. Alfred Yung, René H. Medema, Xi He, and Suyun Huang. 2011. 'FoxM1 Promotes β-Catenin Nuclear Localization and Controls Wnt Target-Gene Expression and Glioma Tumorigenesis'. *Cancer Cell* 20(4):427–42. doi: 10.1016/j.ccr.2011.08.016.

Zhou, Mengli, Yushu Ma, Chun-Cheng Chiang, Edwin C. Rock, Kathryn E. Luker, Gary D. Luker, and Yu-Chih Chen. 2023. 'High-Throughput Cellular Heterogeneity Analysis in Cell Migration at the Single-Cell Level.' *Small (Weinheim an Der Bergstrasse, Germany)* 19(6):e2206754. doi: 10.1002/smll.202206754.

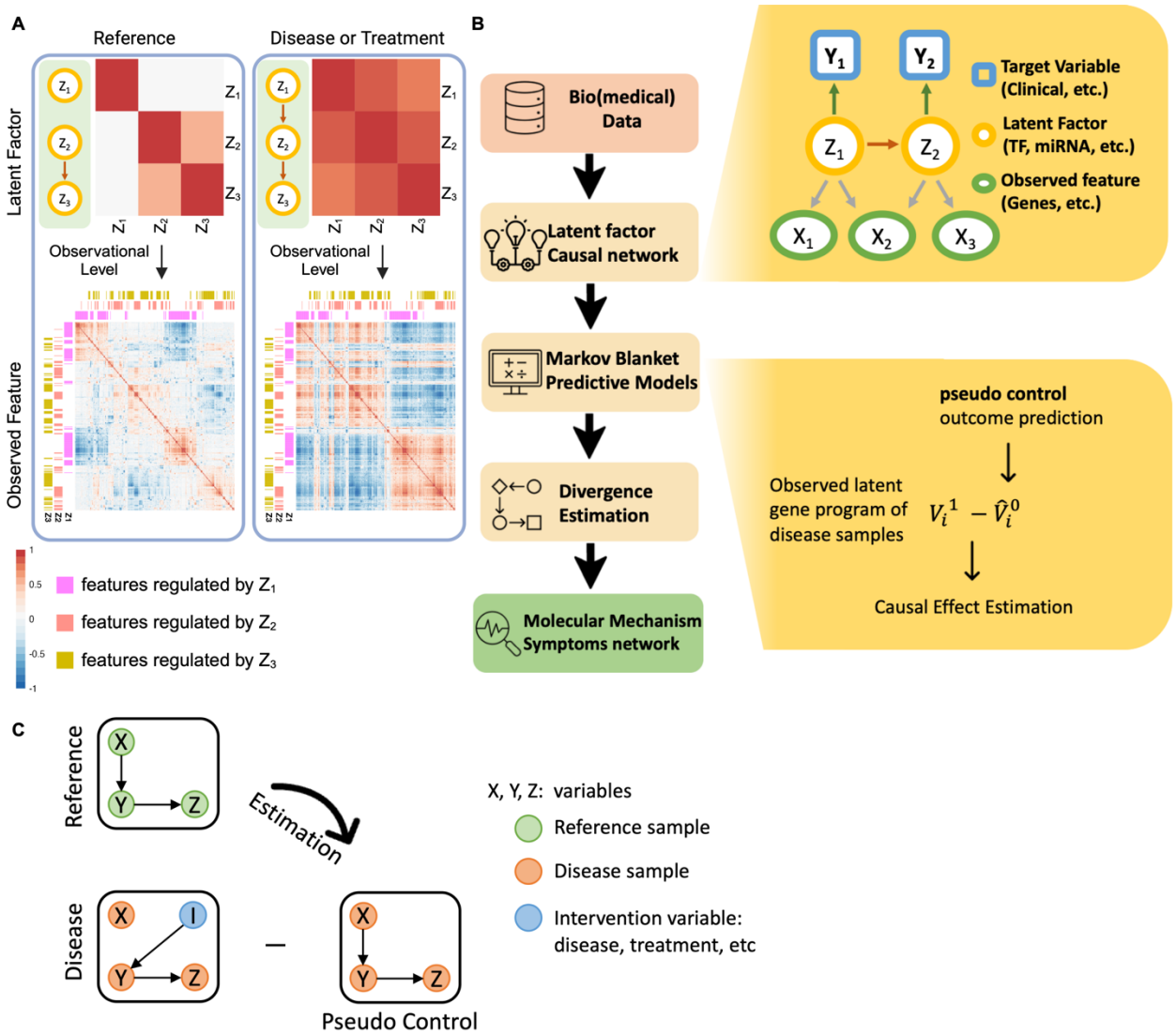

Figure 1. Change in gene program interactions and algorithm overview of LaGrACE.
(A) Three latent factors (LFs) regulate a set of observed features separately in reference condition (left). If interactions among three LFs were changed in disease (or treatment) condition (right), the correlation of regulated features would be altered consequently. (B) LaGrACE is a framework to construct pseudo control leveraging regulatory network of reference sample and quantify heterogeneity of disease group through divergence estimation. (C) Illustration of leveraging pseudo control outcome to calculate divergence score.

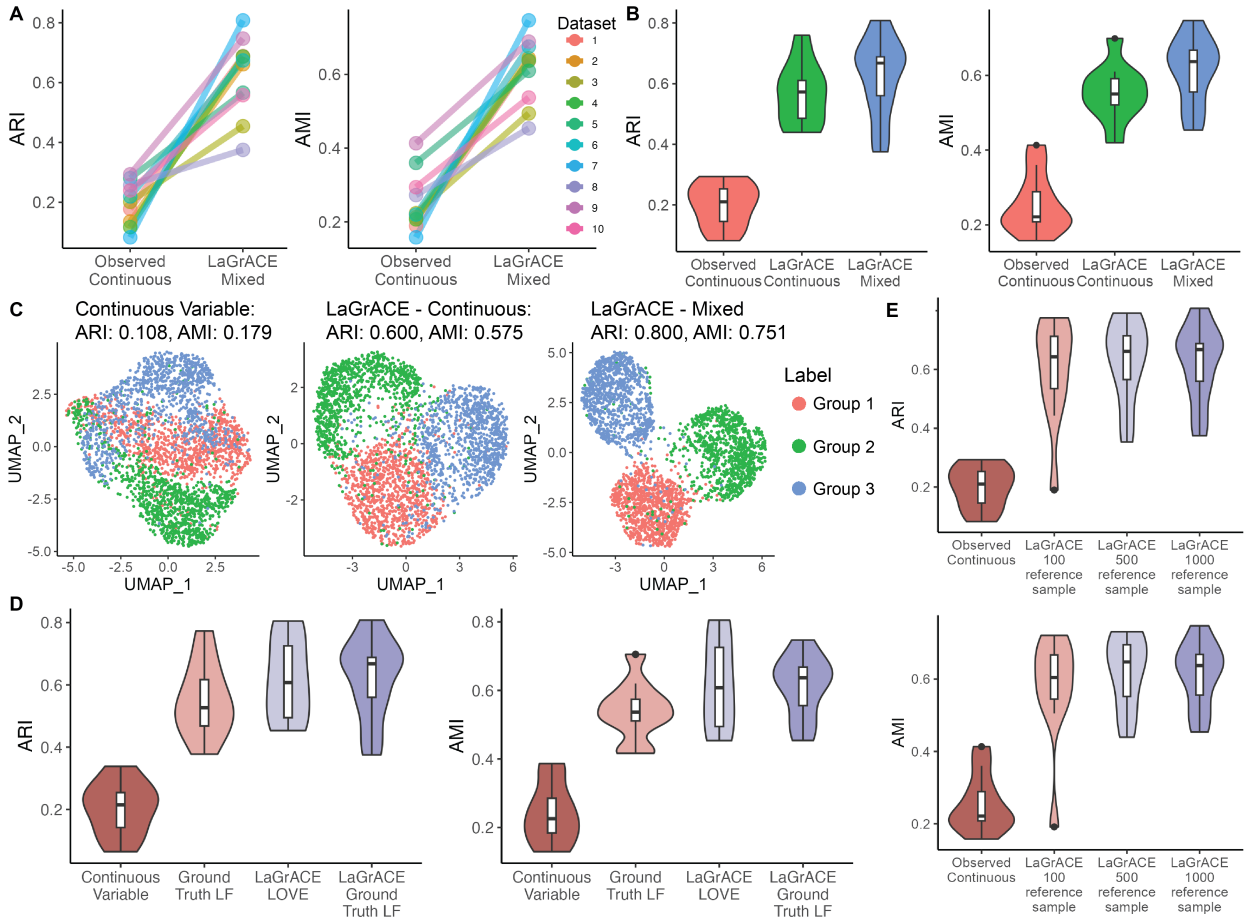

Figure 2. Evaluation of LaGrACE's clustering performance on synthetic dataset with 15% edge change.

(A) Slope chart of ARI and AMI of clustering based on 50 continuous variables (1st point) and LaGrACE feature inferred from both 50 continuous variables and 25 discrete variables (2nd point).

(B) Box plot of ARI and AMI of clustering based on simulated 50 continuous variables (1st point), LaGrACE feature inferred from continuous variables only (2nd point) and LaGrACE feature inferred from both continuous variables and discrete variables (3rd point).

(C) UMAP plot of simulated data colored by clusters based on high-dimensional (2500) continuous variables (ARI: 0.108; AMI:0.179), LaGrACE features inferred from continuous variables only (ARI: 0.600; AMI:0.575) and LaGrACE feature inferred from both continuous variables and discrete variables (ARI: 0.800, AMI: 0.751).

(D) Violin plot of ARI and AMI of clustering based on high-dimensional continuous observed variables, LOVE-recovered latent factors (LOVE LF), ground truth LF, LaGrACE feature with LOVE LF, and LaGrACE feature with ground truth LF.

(E) Violin plot of ARI and AMI of clustering based on 50 continuous variables and LaGrACE features with 100, 500 & 1000 reference samples.

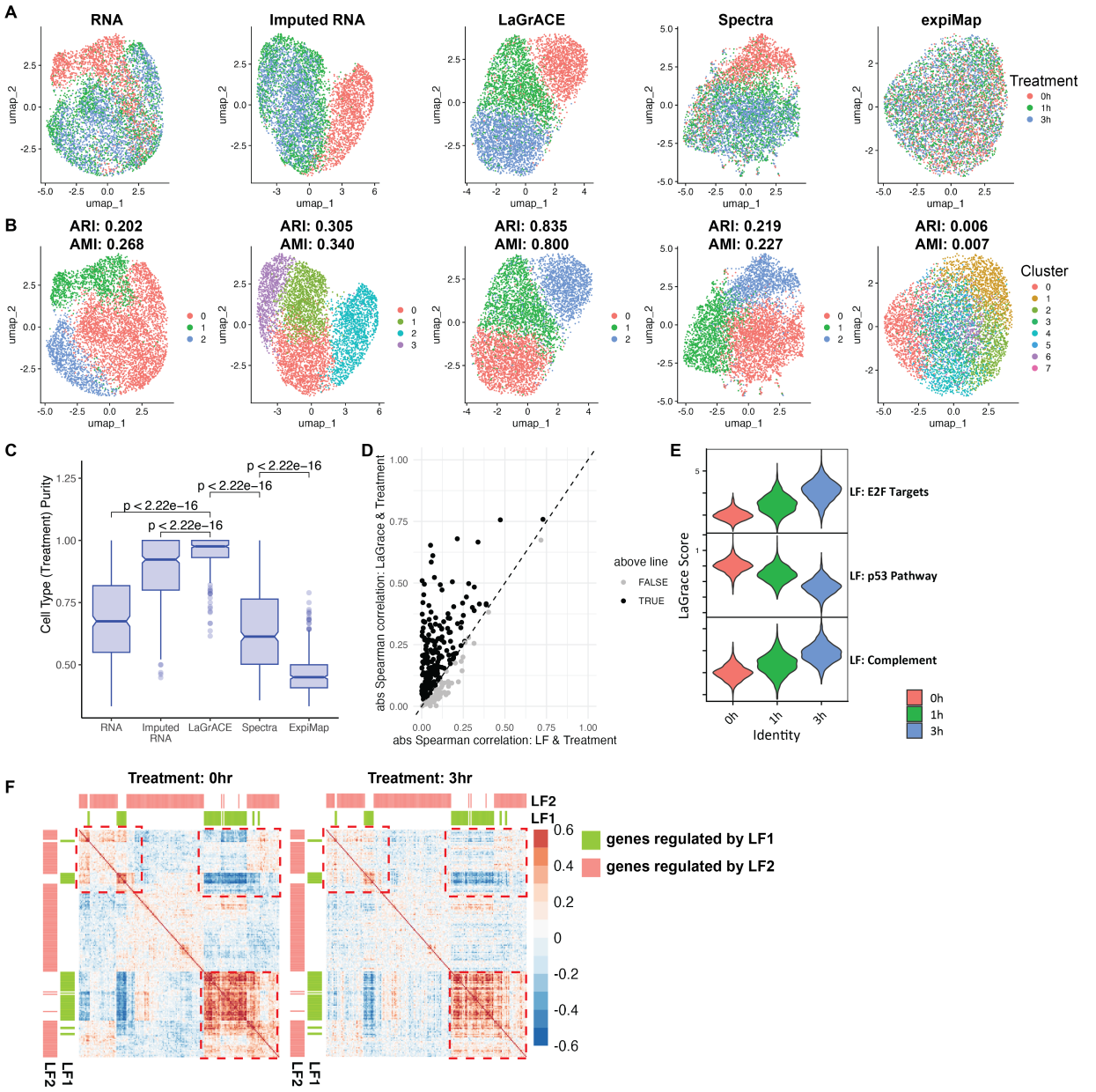

Figure 3. LaGrACE captures treatment effect at single-cell resolution.

A549 lung adenocarcinoma cells were treated with 100 nM dexamethasone for 1 hr or 3 hr or ethanol (control group).

(A) A549 cells visualized and colored by treatment time on UMAP embeddings computed based on unimputed RNA profiles, RNA profiles imputed by SCVI, LaGrACE features and gene set scores inferred by Spectra and expiMap.

(B) UMAP plot of A549 cells colored by clusters based on unimputed RNA (ARI: 0.202, AMI: 0.268), imputed RNA (ARI: 0.305, AMI: 0.340), LaGrACE features (ARI: 0.835, AMI: 0.800), Spectra gene set scores (ARI: 0.219, AMI: 0.227), and expiMap gene set scores (ARI: 0.006, AMI: 0.007).

(C) Boxplot visualized purity score of cell type (treatment time) for single cell Neighborhoods constructed using Milo.

(D) Scatter plot of absolution Spearman rank correlation coefficient between treatment time and latent factors (X axis) / LaGrace scores of latent factors (Y axis)

(E) Violin plot of three LaGrACE features significantly correlated with treatment time (spearman correlation

coefficient of LF: E2F targets, LF: p53 Pathway & LF: Complement: 0.758, -0.756 & 0.666)

(F) Correlation heatmaps of genes regulated by LF1 and its Markov blanket (LF2) at different treatment time points. The region highlighted by dash line box exhibits change in correlation of gene expression regulated by LF1 and its Markov Blanket (LF2) across conditions.

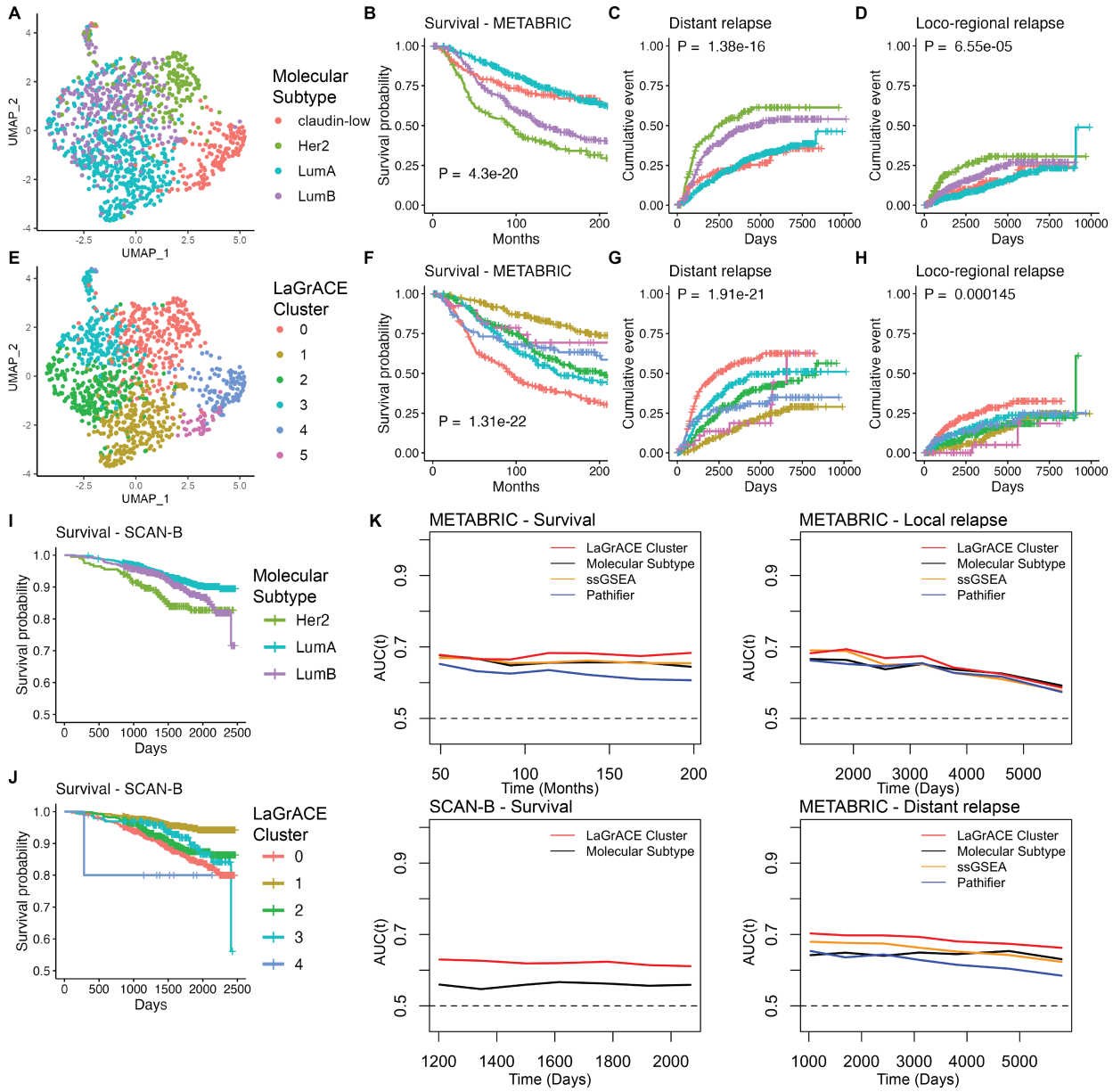

Figure 4. LaGrACE identified six clusters on breast cancer dataset.

(A) UMAP visualization of METABRIC breast cancer samples colored by tumor molecular subtype.
(B) Breast cancer survival curve for molecular subtypes on METABRIC dataset.
(C) Kaplan-Meier estimate of distant relapse on METABRIC dataset by molecular subtype.
(D) Kaplan-Meier estimate of local-regional relapse on METABRIC dataset by molecular subtype.
(E) UMAP visualization of METABRIC breast cancer samples colored by LaGrACE clusters.
(F) Breast cancer survival curve for LaGrACE clusters on METABRIC dataset.
(G) Kaplan-Meier estimate of distant relapse on METABRIC dataset by LaGrACE clusters.
(H) Kaplan-Meier estimate of local-regional on METABRIC dataset by LaGrACE clusters.
(I) Breast cancer survival curve for molecular subtypes on SCAN-B dataset.
(J) Breast cancer survival curve for LaGrACE clusters on SCAN-B dataset.
(K) Time dependent AUC of survival and recurrent event on samples from METABRIC and SCANB dataset.

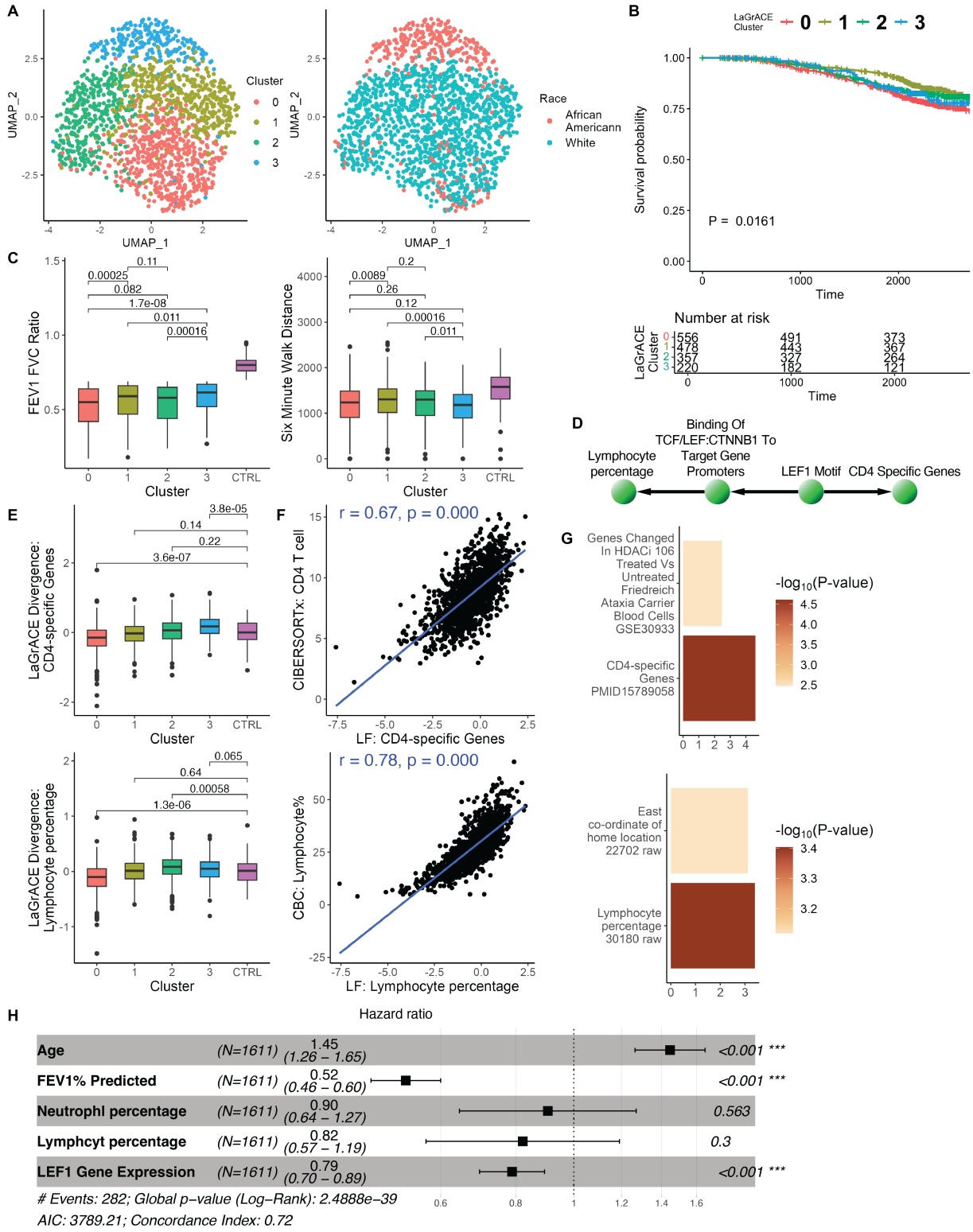

Figure 5. LaGrACE identified novel clusters on COPD dataset

(A) UMAP visualization of COPD patients colored by LaGrACE clusters and race.

(B) Kaplan-Meier survival curve of COPD patients by LaGrACE clusters.

(C) Boxplot of FEV1 FVC ratio and six-minute walk distance.

(D) Subgraph of reference network. Nodes are annotated based on gene set enrichment analysis.
(E) Boxplot of LaGrace score of LF: lymphocyte percentage and LF: CD4-specific genes.
(F) Top, the correlation between LF: lymphocyte percentage and lymphocyte percentage from complete blood count (CBC). Bottom, the correlation between LF: CD4-specific genes and cell type fraction sum of inferred CD4 T cell subpopulations using CIBERSORTx.
(G) Gene set enrichment visualization barplots of LF: lymphocyte percentage (UK Biobank GWAS) and LF: CD4-specific genes (HDSigDB Human).
(H) Forest plot summarizing result of multivariate cox proportional hazards regression model.

Table 1. Molecular subtype composition of LaGrACE breast cancer identified clusters.

	METABRIC				SCANB		
	claudin-low	**Her2**	**LumA**	**LumB**	**Her2**	**LumA**	**LumB**
0	4	140	36	97	193	228	284
1	15	3	216	28	2	741	27
2	0	8	155	79	0	354	153
3	2	10	53	129	2	47	148
4	102	9	0	4	4	2	4
5	41	0	0	0	0	0	0

Table 2. Clinical characteristics of COPD participants vary across clusters. P-values were calculated using a Kruskal-Wallis test for continuous and ordinal variables, and a Chi-squared test for discrete variables, to assess if there are differences in the distribution of variables among clusters.

Variable	Reference Group	Cluster 0	Cluster 1	Cluster 2	Cluster 3	# missing	P.Val
Participants, n	139	556	478	357	220	0	
Age at Phase 2 visit, yr, mean (SD)	64.527 (7.544)	69.182 (8.034)	68.867 (8.375)	69.832 (8.045)	62.075 (7.413)	0	
gender, Female, n (%)	78 (56.12)	224 (40.29)	230 (48.12)	136 (38.1)	110 (50)	0	
race, African American, n (%)	29 (20.86)	60 (10.79)	31 (6.49)	28 (7.84)	217 (98.64)	0	
race, White, n (%)	110 (79.14)	527 (94.78)	447 (93.51)	329 (92.16)	3 (1.36)	0	
Gold Stage 0, n	139	NA	NA	NA	NA	0	
Gold Stage 1, n	NA	120	113	93	49	0	
Gold Stage 2, n	NA	226	224	149	117	0	
Gold Stage 3, n	NA	146	99	83	41	0	
Gold Stage 4, n	NA	64	42	32	13	0	
Complete blood count measurement							
Lymphocyte percentage, mean (SD)	32.266 (9.522)	22.016 (7.14)	28.912 (8.277)	26.764 (8.072)	36.914 (9.046)	5	1.34E-89
Neutrophil percentage, mean (SD)	56.417 (9.616)	66.959 (8.596)	58.895 (8.66)	61.138 (8.856)	51.645 (10.072)	5	1.24E-85
Neutrophil, K/uL, mean (SD)	3.792 (1.504)	5.216 (1.833)	4.404 (1.391)	4.625 (1.523)	3.327 (1.429)	7	8.31E-48
Lymphocytes, K/uL, mean (SD)	2.062 (0.661)	1.636 (0.578)	2.121 (0.774)	1.985 (0.775)	2.256 (0.718)	7	5.88E-38
CIBERSORTx inferred cell type fraction							
Neutrophils percentage, mean (SD)	11.132 (4.05)	14.955 (4.689)	10.968 (3.337)	13.106 (4.079)	9.207 (3.457)	0	6.83E-75
CD4 memory resting T cell percentage, mean (SD)	4.126 (1.16)	3.45 (1.184)	4.116 (1.093)	3.578 (1.165)	4.81 (1.237)	0	5.79E-48
CD4 naive T cell percentage, mean (SD)	4.661 (1.56)	3.637 (1.136)	4.26 (1.331)	4.493 (1.343)	4.626 (1.386)	0	1.37E-29
Plasma cell percentage, mean (SD)	0.172 (0.168)	0.283 (0.213)	0.178 (0.147)	0.248 (0.182)	0.163 (0.143)	0	1.77E-26
Symptoms							
MMRC dyspnea score, mean (SD)	0.626 (1.031)	1.766 (1.459)	1.454 (1.389)	1.529 (1.468)	1.95 (1.469)	0	2.66E-05
CAT score, mean (SD)	7.36 (5.992)	14.959 (8.842)	13.134 (8.526)	13.958 (8.351)	16.086 (9.092)	0	7.58E-05
SGRQ score: Active, mean (SD)	21.225 (23.326)	47.811 (28.956)	41.837 (28.896)	43.972 (29.521)	49.734 (30.07)	0	9.48E-04

distwalked_P2, ft, mean (SD)	1530.109 (408.52)	1174.192 (465.711)	1258.888 (425.263)	1211.714 (439.033)	1132.23 (438.257)	39	1.38E-03
SGRQ score: Total, mean (SD)	12.548 (13.811)	32.415 (21.92)	28.492 (21.018)	30.077 (21.368)	34.1 (22.435)	0	4.14E-03
Spirometry							
total lung capacity, Thirona, liters, mean (SD)	5.203 (1.214)	6.108 (1.41)	6.049 (1.478)	6.251 (1.46)	5.014 (1.231)	153	4.73E-22
FRC/TLC ratio, Thirona, mean (SD)	0.513 (0.084)	0.645 (0.111)	0.625 (0.109)	0.627 (0.107)	0.686 (0.109)	256	5.60E-09
FEV1/FVC ratio, post-BD, mean (SD)	0.798 (0.053)	0.526 (0.129)	0.554 (0.122)	0.541 (0.125)	0.583 (0.098)	1	1.32E-07
DLCO percent predicted, GLI, mean (SD)	92.289 (16.349)	64.716 (22.658)	70.411 (22.755)	66.901 (21.915)	62.546 (19.214)	258	4.79E-05
BODE score, mean (SD)	0.511 (1.03)	2.871 (2.594)	2.316 (2.348)	2.463 (2.405)	2.889 (2.312)	85	9.72E-04
Smoking Status							
Smoked in the past month, Yes, n (%)	0 (0)	160 (28.78)	140 (29.29)	107 (30.06)	140 (63.64)	1	1.48E-21
Smoked over the past 5 years, Yes, n (%)	22 (15.83)	206 (37.18)	195 (40.88)	138 (38.87)	164 (74.55)	5	1.99E-21

BD bronchodilator, *BODE* body mass index, airflow obstruction, dyspnea, and exercise capacity, *CAT* COPD Assessment Test, *DLCO* diffusing capacity for carbon monoxide, *FEV1* forced expiratory volume in 1-s, *FRC* functional residual capacity, *FVC* forced vital capacity, *MMRC* Modified Medical Research Council Dyspnea Scale, *SGRQ* St. George's Respiratory Questionnaire, *TLC* total lung capacity

Table 3. Correlation Analysis between LEF1 gene expression and clinical variables. PCC represents Pearson correlation coefficient; SRCC represents Spearman's rank correlation coefficient.

Correlation Analysis	**Variable**	**PCC (r)**	**P-val**	**SRCC (rho)**	**P-val**
LaGrACE divergence score of LF: Lymphocyte percentage	FEV1 FVC Ratio (n=1749)	**0.12**	**1.34E-06**	**0.10**	**6.05E-05**
	FEV1% Predicted (n=1750)	**0.10**	**1.22E-04**	**0.08**	**1.18E-03**
LaGrACE divergence score of LF: CD4-specific Genes	FEV1 FVC Ratio (n=1749)	**0.20**	**8.68E-16**	**0.19**	**7.08E-14**
	FEV1% Predicted (n=1750)	**0.11**	**1.41E-05**	**0.11**	**1.52E-05**
COPDgene: LEF1 gene expression	Age (n=1750)	**-0.31**	**7.81E-40**	**-0.32**	**3.20E-43**
	FEV1 FVC Ratio (n=1749)	**0.20**	**1.17E-16**	**0.19**	**3.19E-15**
	FEV1% Predicted (n=1750)	**0.18**	**7.10E-14**	**0.16**	**9.53E-12**
ECLIPSE: LEF1 partial region (AF294627)	Age (n=229)	**-0.17**	**8.22E-03**	**-0.24**	**2.72E-04**
	FEV1 FVC Ratio (n=229)	0.04	5.15E-01	0.09	1.96E-01
	FEV1% Predicted (n=229)	0.04	5.58E-01	0.07	2.69E-01
ECLIPSE: LEF1 Full Length (AF288571)	Age (n=229)	**-0.25**	**1.51E-04**	**-0.25**	**1.19E-04**
	FEV1 FVC Ratio (n=229)	**0.25**	**9.95E-05**	**0.27**	**4.66E-05**
	FEV1% Predicted (n=229)	**0.24**	**1.94E-04**	**0.25**	**1.25E-04**
Partial Correlation Analysis	**Variable**	**PCC (r)**	**P-val**	**SRCC (rho)**	**P-val**
COPDgene: LEF1 gene expression condition on AGE	FEV1 FVC Ratio (n=1749)	**0.16**	**4.54E-11**	**0.15**	**1.11E-09**
	FEV1% Predicted (n=1750)	**0.18**	**3.17E-14**	**0.16**	**1.55E-11**
ECLIPSE: LEF1 partial region (AF294627) condition on AGE	FEV1 FVC Ratio (n=229)	0.00	9.70E-01	0.04	5.41E-01
	FEV1% Predicted (n=229)	0.00	9.92E-01	0.04	5.85E-01
ECLIPSE: LEF1 Full Length (AF288571) condition on AGE	FEV1 FVC Ratio (n=229)	**0.20**	**2.11E-03**	**0.23**	**5.33E-04**
	FEV1% Predicted (n=229)	**0.20**	**2.38E-03**	**0.22**	**8.24E-04**