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[1,2,4], Panayiotis V. Benos[1,2,3,*]

[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]University of Florida, Gainesville, Florida, USA, [4]UPMC Hillman Cancer Center, University of Pittsburgh, Pittsburgh, Pennsylvania

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 normal gene program. 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, breast cancer and chronic obstructive pulmonary disease (COPD) datasets, and single-cell RNA sequencing (scRNA-seq) datasets. Our findings demonstrate that LaGrACE is exceptionally robust in identifying biologically meaningful and prognostic subtypes. Additionally, it effectively discerns drug-response signals at a single-cell resolution. 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 disease mechanisms.

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 are capable of working cooperatively through complex mechanisms, some of which are not observed in typical datasets. Such mechanisms include transcription factor (TF) activities, signaling transduction pathway interconnections and epigenetic modifications (Lambert et al., 2018). The regulation of gene programs by those latent factors usually lead to collinearity of the expression of member genes. Differences in these gene programs and their connections could lead to patient phenotypic differences, reflecting their roles in clinical outcomes, response to treatment, or molecular disease subtypes. Although these programs are not directly observable, they provide a more comprehensive picture of the genetic mechanism of disease and facilitate the development of effective therapeutic strategies.

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, 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 gene program information from curated databases. Examples of such methods include Spectra (Kunes et al., 2023), expiMap (Lotfollahi et al., 2023), Pathifier (Yotam et al., 2013) and ssGSEA (Barbie et al., 2009). Nevertheless, these approaches strongly depend on the curated programs, which are derived from multiple biological contexts, and may not correspond to the specific context under study.

In the context of treatment or disease, the regulation of gene programs is often altered, leading to changes in co-expression patterns of transcriptomic profiles (Lee & 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 1 A**). 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 network, 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, interacting with each other, can be modeled using probabilistic graphical models (Friedman, 2004; Kunes et al., 2023). Nonetheless, diseases often present as syndromes encompassing multiple subtypes, and response to a treatment may vary based on a patient's clinical characteristics and medical history. Consequently, a significant challenge arises: how to model alterations in gene programs within a single sample, when this sample can only be observed either in the control or the non-control state?

We propose a novel approach to estimate the changes in gene program interactions in single samples. First, we start by assuming that there is a baseline network of interactions that is present in all control samples. This "pseudo control" network will serve as the reference network. On any given disease (or treatment) sample, we expect that some of the gene programs and their interactions will change. In order to assess which ones, 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 assume that the topology of this 'pseudo control network' for the most part is maintained in samples from the treatment group, although certain part of this network may be altered in the disease (or treatment) samples. Which part is altered 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 control sample could also be observed in a disease state (or after treatment), will gene programs with analogous biological functions exhibit comparable behavior (**Figure 1C**)?

In a case-control study, we can only measure transcriptome 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) represents a treatment label assignment variable for an individual i, where only a part of variables can be directly observed:

$$X_{ij}^{0} = \begin{cases} X_{ij}, W_i = 0 \\ ?, W_i = 1 \end{cases}, X_{ij}^{1} = \begin{cases} ?, W_i = 0 \\ X_{ij}, W_i = 1 \end{cases}$$

To model gene program interactions, we first identify these gene programs as *non-orthogonal* latent factors. For latent factor identification, we use the LOVE algorithm (Bing et al., 2019). Second, we infer a Bayesian network to delineate interactions among gene programs and between gene programs and

clinical variables. With such a reference graph, the value of a gene program X_i can be approximated based on its Markov Blanket $MB(X_i)$ using a regression model. A set of Markov Blanket regression models, which trained on reference samples, are representing reference gene program interactions. The pseudo control outcome of treatment samples, $\hat{X}_i^0$, could then be estimated using those trained models. 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 the pseudo control outcome and the observed value in treatment samples.

$$\text{Divergence Score: } X_i^1 - \hat{X}_i^0$$

These divergence scores can then be used to group samples based on patterns of dysregulation by evaluating alterations within the reference graph 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 graph alongside three perturbed graphs, 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 graph was used to generate 1,000 samples. LaGrACE learned the reference graph from the reference graph samples. There are many ways to calculate the optimum sparsity of the reference graph (Liu et al., 2010). We found that the reference graph sparsity that minimizes the reconstruction error of reference samples (in nested cross-validation setup) also minimizes "disease" sample classification error (**see Methods)**. Subsequently, we calculated divergence scores for each "disease" sample, generated by one of the perturbed graphs. We clustered the samples based on the divergence scores using the Markov blanket approach (**see Methods)**. The resulting clusters demonstrated improved separation and better Adjusted Rand Index (ARI) and Adjusted Mutual Information (AMI) scores compared to unprocessed 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 estimated 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 graph 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 subtypes identified and validated by LaGrACE

Application of LaGrACE to the METABRIC breast cancer dataset uncovered potential new 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). Each SCAN-B sample was assigned to the most proximal of the clusters learned from the METABRIC dataset based on its network perturbation. 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 in both METABRIC and SCAN-B datasets (Supplementary Table S3). Time-dependent AUC analysis also showed LaGrACE's superior discrimination ability in metastasis, survival, and local relapse (**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 & 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 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 group, 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 et al., 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. With both control (reference) and non-control (disease or treatment) samples, our method estimates how gene program regulation differs in the samples in a disease state or those undergoing treatment. Our simulation studies demonstrate LaGrACE's robust performance in subtype identification across varying reference sample sizes, high dimensionality, and collinearity. LaGrACE also outperforms existing single cell gene program methods and effectively discerned drug treatment signals across varying doses and exposure times at a single-cell resolution.

LaGrACE successfully identifies biologically significant and prognostic clusters in breast cancer data (METABRIC). The stability of these identified clusters was affirmed through validation with the SCAN-B cohort, illustrating our method's reliability in patient stratification. Notably, these clusters provide more prognostic insight than molecular subtypes and clusters deduced from *prior knowledge*-based methods. LaGrACE does not depend on prior information, which is an additional advantage. 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.

We also report 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 not fit completely to 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. In the future, we plan to extend LaGrACE to non-Gaussian and non-linear models. Finally, LaGrACE is currently used for identifying subtypes in an unsupervised manner. Nevertheless, this approach can be extended to supervised tasks such as predicting clinical variables of interest.

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. 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. Let X_{ij}^{w} be value of feature i (could be gene program or clinical variable of interest) in a sample j if the value was observed from an individual with a treatment label W = w. For an individual from treatment group, X^1 represents observed X value, but X^0 is unknown. We infer pseudo control outcome $\hat{X}_{i}^{0}$ to approximate X^0. We focus on understanding its impact without the complication of multiple treatments or interference effects.

LaGrACE comprises four main steps, which are summarized below and elaborated in the following sections.

(1) 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) Learn a Bayesian network, based on inferred latent factors and clinical variables of interest, as the reference graph. This graph encapsulates interactions among latent factors from gene expression profiles and clinical variables.
(3) Estimate pseudo control outcome to approximate pre-treatment state on a gene program or a clinical measurement by utilizing its Markov blanket derived from the reference graph.
(4) Calculate difference between pseudo-control outcome and observed value for treatment samples.

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

$$X = AZ + E$$

where $X \in R^p$ denotes p observed variables; $Z \in R^K$ represents K latent factors; $A \in R^{p \times K}$ represents the membership matrix assigning p observed variables to K latent factors; and E denotes independent random noise. 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 and lambda) were selected using Akaike Information Criterion (AIC) method.

Reference Graph 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), over all measured 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) (Andrews et al., 2019):

$$S(\mathcal{g}, \boldsymbol{Z}) = \sum_{j}^{p} s(Z_j, \boldsymbol{Z}_{Pa_j^{\mathcal{g}}})$$

where

$$s\left(Z_j, \boldsymbol{Z}_{Pa_j^{\mathcal{g}}}\right) = \ell\left(\hat{\theta}_{mle}\middle|\boldsymbol{Z}_{\{j\}\cup Pa_j^{\mathcal{g}}}\right) - \ell\left(\hat{\theta}_{mle}\middle|\boldsymbol{Z}_{Pa_j^{\mathcal{g}}}\right) - \frac{PD}{2}\left|Z_j\right|\left|\boldsymbol{Z}_{Pa_j^{\mathcal{g}}}\right|\log(n)$$

And PD is a penalty value (penalty discount) that controls the sparsity of graph. Let $X = (X_1, \dots, X_n)$ be a set of p random variables with n instances. X_j may be either continuous (C) or discrete (D). Z represents a set of variables after performing One-hot Encoding on discrete variables.

$$Z_j = \begin{cases} X_j & if\ j\ \in C \\ \mathbb{I}_1(X_j), \dots, \mathbb{I}_{k-1}(X_j) & if\ j\ \in D \end{cases}$$

Where $\mathbb{I}_k(X_j)$ is an indicator function such that $\mathbb{I}_k(X_j) = 1$ if $X_j = k$, otherwise $\mathbb{I}_k(X_j) = 0$ for $X_j \in D$. The One-hot Encoding transformation embeds the discrete variables into a degenerate distribution. Thus, the transformed data is further treated as jointly Gaussian. $\left(\hat{\theta}_{mle}\middle|\boldsymbol{Z}_{\{j\}\cup Pa_j^{\mathcal{g}}}\right)$ and $\ell\left(\hat{\theta}_{mle}\middle|\boldsymbol{Z}_{Pa_j^{\mathcal{g}}}\right)$ are Gaussian log-likelihood of the transformed data Z given the causal directly acyclic graph (DAG $\mathcal{g}$) and its parameters (θ). The score is decomposable, and the total DG score is the sum of local components (its nodes and their parents).

Pseudo Control 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 graph. 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 graph is estimated, all discrete variables are transformed into binary indicator variables, representing each category of the original discrete variable. Markov blanket based regression model is built for each variable in the reference graph based on reference samples.

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

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

$$\hat{X_i}^0 = \sum_{k \in MB(X_i)} {\beta_{k,i}}^0 {X_k}^1 + {\beta_{0,i}}^0 + \varepsilon$$

Divergence Estimation

Causal reasoning involves comparing the outcome when a treatment is administrated (X^1) against when it is not (X^0). A causal effect on the outcome is inferred if these outcomes differ (X^1≠X^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 gene program 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.

Divergence Score: $X_i^1 - \hat{X_i}^0 = X_i^1 - \left(\sum_{k \in MB(X_i)} {\beta_{k,i}}^0 {X_k}^1 + {\beta_{0,i}}^0\right)$

The divergence score across all variables in the reference graph 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 graph.

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 & 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. This model delineates the joint distribution over datasets with mixed variables. We first simulated a reference graph with a vertex degree of 2 and then created three perturbed graphs through random edge modifications (deletion and addition). Each graph contained 50 continuous and 25 discrete variables, the latter with four categorical levels. Datasets of 1000 samples were simulated for each graph based on the LH model, with continuous variables drawn from $N(\mu, \sigma)$, where $\mu U(-1,1)$ and $\sigma(0.5,1)$. Edge weights, W_{X1X2}, were assigned from a uniform distribution $U(0.5,1.5)$. 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 graph and 3000 from the perturbed graphs.

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 graph and three perturbed graphs. 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

$$X = ZA + E \text{ (X)}$$

where X represents the n×p "gene expression" matrix (*n* and *p* are the number of samples and observed variables, respectively). $Z \in R^{n \times K}$ denotes K latent factors. $A \in R^{p \times K}$ represents the "transcriptional regulatory matrix". $EN(0,1)$denotes independent noise. Each Z factor directly influenced approximately 50 observed features, and A was simulated with weights from a Gaussian distribution. 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 graph and 3000 "treatment" samples from the three perturbed graphs.

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 & 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 et al., 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 & Heller, 2005; Harrell et al., 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 et al., 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.

REFERENCES

Adams, T. S., Schupp, J. C., Poli, S., Ayaub, E. A., Neumark, N., Ahangari, F., Chu, S. G., Raby, B. A., DeIuliis, G., Januszyk, M., Duan, Q., Arnett, H. A., Siddiqui, A., Washko, G. R., Homer, R., Yan, X., Rosas, I. O., & Kaminski, N. (2020). Single-cell RNA-seq reveals ectopic and aberrant lung-resident cell populations in idiopathic pulmonary fibrosis. *Science Advances*, *6*(28). https://doi.org/10.1126/sciadv.aba1983

Andrews, B., Ramsey, J., & Cooper, G. F. (2019). Learning High-dimensional Directed Acyclic Graphs with Mixed Data-types. *Proc Mach Learn Res*, *104*(4–21). http://proceedings.mlr.press/v104/andrews19a/andrews19a.pdf

Badia-i-Mompel, P., Vélez Santiago, J., Braunger, J., Geiss, C., Dimitrov, D., Müller-Dott, S., Taus, P., Dugourd, A., Holland, C. H., Ramirez Flores, R. O., & Saez-Rodriguez, J. (2022). decoupleR: ensemble of computational methods to infer biological activities from omics data. *Bioinformatics Advances*, *2*(1), vbac016. https://doi.org/10.1093/bioadv/vbac016

Barbie, D. A., Tamayo, P., Boehm, J. S., Kim, S. Y., Moody, S. E., Dunn, I. F., Schinzel, A. C., Sandy, P., Meylan, E., Scholl, C., Fröhling, S., Chan, E. M., Sos, M. L., Michel, K., Mermel, C., Silver, S. J., Weir, B. A., Reiling, J. H., Sheng, Q., … Hahn, W. C. (2009). Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature*, *462*(7269), 108–112. https://doi.org/10.1038/nature08460

Bernard, P. S., Parker, J. S., Mullins, M., Cheung, M. C. U., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., Quackenhush, J. F., Stijleman, I. J., Palazzo, J., Matron, J. S., Nobel, A. B., Mardis, E., Nielsen, T. O., Ellis, M. J., & Perou, C. M. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of Clinical Oncology*. https://doi.org/10.1200/JCO.2008.18.1370

Bing, X., Bunea, F., Royer, M., & Das, J. (2019). Latent Model-Based Clustering for Biological Discovery. *IScience*, *14*, 125–135. https://doi.org/10.1016/j.isci.2019.03.018

Cao, J., Cusanovich, D. A., Ramani, V., Aghamirzaie, D., Pliner, H. A., Hill, A. J., Daza, R. M., McFaline-Figueroa, J. L., Packer, J. S., Christiansen, L., Steemers, F. J., Adey, A. C., Trapnell, C., & Shendure, J. (2018). Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science*, *361*(6409), 1380–1385. https://doi.org/10.1126/science.aau0730

Chen, Y., Li, Y., Xue, J., Gong, A., Yu, G., Zhou, A., Lin, K., Zhang, S., Zhang, N., Gottardi, C. J., & Huang, S. (2016). Wnt-induced deubiquitination FoxM1 ensures nucleus β-catenin transactivation. *The EMBO Journal*, *35*(6), 668–684. https://doi.org/10.15252/embj.201592810

Curtis, C., Shah, S. P., Chin, S.-F., Turashvili, G., Rueda, O. M., Dunning, M. J., Speed, D., Lynch, A. G., Samarajiwa, S., Yuan, Y., Gräf, S., Ha, G., Haffari, G., Bashashati, A., Russell, R., McKinney, S., Caldas, C., Aparicio, S., Curtis†, C., … subgroup:, D. analysis. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, *486*(7403), 346–352. https://doi.org/10.1038/nature10983

Dann, E., Henderson, N. C., Teichmann, S. A., Morgan, M. D., & Marioni, J. C. (2022). Differential abundance testing on single-cell data using k-nearest neighbor graphs. *Nature Biotechnology*, *40*(2), 245–253. https://doi.org/10.1038/s41587-021-01033-z

Davidson, S., Efremova, M., Riedel, A., Mahata, B., Pramanik, J., Huuhtanen, J., Kar, G., Vento-Tormo, R., Hagai, T., Chen, X., Haniffa, M. A., Shields, J. D., & Teichmann, S. A. (2020). Single-Cell RNA Sequencing Reveals a Dynamic Stromal Niche That Supports Tumor Growth. *Cell Reports*, *31*(7), 107628. https://doi.org/10.1016/j.celrep.2020.107628

Dimitriadou, F. L. and E. (2021). *mlbench: Machine Learning Benchmark Problems*. https://cran.r-project.org/web/packages/mlbench/

Elyahu, Y., Hekselman, I., Eizenberg-Magar, I., Berner, O., Strominger, I., Schiller, M., Mittal, K., Nemirovsky, A., Eremenko, E., Vital, A., Simonovsky, E., Chalifa-Caspi, V., Friedman, N., Yeger-Lotem, E., & Monsonego, A. (2019). Aging promotes reorganization of the CD4 T cell landscape toward extreme regulatory and effector phenotypes. *Science Advances*, *5*(8). https://doi.org/10.1126/sciadv.aaw8330

Feder, K., Edmaier-Schröger, K., Rawat, V. P. S., Kirsten, N., Metzeler, K., Kraus, J. M., Döhner, K., Döhner, H., Kestler, H. A., Feuring-Buske, M., & Buske, C. (2020). Differences in expression and function of LEF1 isoforms in normal versus leukemic hematopoiesis. *Leukemia*, *34*(4), 1027–1037. https://doi.org/10.1038/s41375-019-0635-1

Freedman, D., Gates, L., Flanders, W., Van Assendelft, O., Barboriak, J., Joesoef, M., & Byers, T. (1997). Black/white differences in leukocyte subpopulations in men. *International Journal of Epidemiology*, *26*(4), 757–764. https://doi.org/10.1093/ije/26.4.757

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

Ghosh, A. J., Saferali, A., Lee, S., Chase, R., Moll, M., Morrow, J., Yun, J., Castaldi, P. J., & Hersh, C. P. (2022). Blood RNA sequencing shows overlapping gene expression across COPD phenotype domains. *Thorax*, *77*(2), 115–122. https://doi.org/10.1136/thoraxjnl-2020-216401

Gönen, M., & Heller, G. (2005). Concordance probability and discriminatory power in proportional hazards regression. *Biometrika*, *92*(4), 965–970. https://doi.org/10.1093/biomet/92.4.965

González-Mariscal, L., Miranda, J., Gallego-Gutiérrez, H., Cano-Cortina, M., & Amaya, E. (2020). Relationship between apical

junction proteins, gene expression and cancer. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, *1862*(9), 183278. https://doi.org/10.1016/j.bbamem.2020.183278

Greenberg, A. K., Hu, J., Basu, S., Hay, J., Reibman, J., Yie, T., Tchou-Wong, K. M., Rom, W. N., & Lee, T. C. (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), 320–328. https://doi.org/10.1165/rcmb.4710

Hänzelmann, S., Castelo, R., & Guinney, J. (2013). GSVA: Gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinformatics*. https://doi.org/10.1186/1471-2105-14-7

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W. M., Zheng, S., Butler, A., Lee, M. J., Wilk, A. J., Darby, C., Zager, M., Hoffman, P., Stoeckius, M., Papalexi, E., Mimitou, E. P., Jain, J., Srivastava, A., Stuart, T., Fleming, L. M., Yeung, B., … Satija, R. (2021). Integrated analysis of multimodal single-cell data. *Cell*, *184*(13), 3573-3587.e29. https://doi.org/https://doi.org/10.1016/j.cell.2021.04.048

Harrell, F. E. J., Lee, K. L., & Mark, D. B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, *15*(4), 361–387. https://doi.org/10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4

Harris, B. D., Crow, M., Fischer, S., & Gillis, J. (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. https://doi.org/10.1016/j.cels.2021.04.010

Harvard Chan Bioinformatics Core. (n.d.). *Single-cell RNA-seq analysis workshop*. https://hbctraining.github.io/scRNA-seq/lessons/elbow_plot_metric.html

Heikenwalder, M., & Lorentzen, A. (2019). The role of polarisation of circulating tumour cells in cancer metastasis. *Cellular and Molecular Life Sciences*, *76*(19), 3765–3781. https://doi.org/10.1007/s00018-019-03169-3

Hill, L. D., Sun, L., Leuschen, M. P., & Zach, T. L. (1993). C3 synthesis by A549 alveolar epithelial cells is increased by interferon-gamma and dexamethasone. *Immunology*, *79*(2), 236–240. http://www.ncbi.nlm.nih.gov/pubmed/8344703

Jia, M., Sayed, K., Kapetanaki, M. G., Dion, W., Rosas, L., Irfan, S., Valenzi, E., Mora, A. L., Lafyatis, R. A., Rojas, M., Zhu, B., & Benos, P. V. (2023). <scp>LEF1</scp> isoforms regulate cellular senescence and aging. *Aging Cell*, *22*(12). https://doi.org/10.1111/acel.14024

Johnson, J., Thijssen, B., McDermott, U., Garnett, M., Wessels, L. F. A., & Bernards, R. (2016). Targeting the RB-E2F pathway in breast cancer. *Oncogene*, *35*(37), 4829–4835. https://doi.org/10.1038/onc.2016.32

Julien Chiquet, Guillem Rigaill, Martina Sundqvist, V. D. (2020). *aricode: Efficient Computations of Standard Clustering Comparison Measures*.

Kamarudin, A. N., Cox, T., & Kolamunnage-Dona, R. (2017). Time-dependent ROC curve analysis in medical research: current methods and applications. *BMC Medical Research Methodology*, *17*(1), 53. https://doi.org/10.1186/s12874-017-0332-6

Kassambara A, Kosinski M, Biecek P, F. S. (2021). *survminer: Drawing Survival Curves using 'ggplot2'*. https://cran.r-project.org/web/packages/survminer/index.html

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

Kunes, R. Z., Walle, T., Land, M., Nawy, T., & Pe'er, D. (2023). Supervised discovery of interpretable gene programs from single-cell data. *Nature Biotechnology*. https://doi.org/10.1038/s41587-023-01940-3

Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T. R., & Weirauch, M. T. (2018). The Human Transcription Factors. *Cell*, *172*(4), 650–665. https://doi.org/10.1016/j.cell.2018.01.029

Lee, T. I., & Young, R. A. (2013). Transcriptional Regulation and Its Misregulation in Disease. *Cell*, *152*(6), 1237–1251. https://doi.org/10.1016/j.cell.2013.02.014

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., & Tamayo, P. (2015). The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Systems*. https://doi.org/10.1016/j.cels.2015.12.004

Liu, H., Roeder, K., & Wasserman, L. (2010). Stability approach to regularization selection (StARS) for high dimensional graphical models. *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010, NIPS 2010*.

Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., & Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. *Nature Methods*, *15*(12), 1053–1058. https://doi.org/10.1038/s41592-018-0229-2

Lotfollahi, M., Rybakov, S., Hrovatin, K., Hediyeh-zadeh, S., Talavera-López, C., Misharin, A. V., & Theis, F. J. (2023). Biologically informed deep learning to query gene programs in single-cell atlases. *Nature Cell Biology*. https://doi.org/10.1038/s41556-022-01072-x

Mancini, M., De Santis, S., Monaldi, C., Bavaro, L., Martelli, M., Castagnetti, F., Gugliotta, G., Rosti, G., Santucci, M. A., Martinelli, G., Cavo, M., & Soverini, S. (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. https://doi.org/10.1186/s13046-019-1197-9

Müller-Dott, S., Tsirvouli, E., Vazquez, M., Ramirez Flores, R. O., Badia-i-Mompel, P., Fallegger, R., Türei, D., Lægreid, A., & Saez-Rodriguez, J. (2023). Expanding the coverage of regulons from high-confidence prior knowledge for accurate estimation

of transcription factor activities. *Nucleic Acids Research*, *51*(20), 10934–10949. https://doi.org/10.1093/nar/gkad841

Nédélec, Y., Sanz, J., Baharian, G., Szpiech, Z. A., Pacis, A., Dumaine, A., Grenier, J.-C., Freiman, A., Sams, A. J., Hebert, S., Pagé Sabourin, A., Luca, F., Blekhman, R., Hernandez, R. D., Pique-Regi, R., Tung, J., Yotova, V., & Barreiro, L. B. (2016). Genetic Ancestry and Natural Selection Drive Population Differences in Immune Responses to Pathogens. *Cell*, *167*(3), 657-669.e21. https://doi.org/10.1016/j.cell.2016.09.025

Newman, A. M., Steen, C. B., Liu, C. L., Gentles, A. J., Chaudhuri, A. A., Scherer, F., Khodadoust, M. S., Esfahani, M. S., Luca, B. A., Steiner, D., Diehn, M., & Alizadeh, A. A. (2019). Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nature Biotechnology*, *37*(7), 773–782. https://doi.org/10.1038/s41587-019-0114-2

Nielsen, T. O., Parker, J. S., Leung, S., Voduc, D., Ebbert, M., Vickery, T., Davies, S. R., Snider, J., Stijleman, I. J., Reed, J., Cheang, M. C. U., Mardis, E. R., Perou, C. M., Bernard, P. S., & Ellis, M. J. (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), 5222–5232. https://doi.org/10.1158/1078-0432.CCR-10-1282

Oshi, M., Takahashi, H., Tokumaru, Y., Yan, L., Rashid, O. M., Nagahashi, M., Matsuyama, R., Endo, I., & Takabe, K. (2020). The E2F Pathway Score as a Predictive Biomarker of Response to Neoadjuvant Therapy in ER+/HER2− Breast Cancer. *Cells*, *9*(7), 1643. https://doi.org/10.3390/cells9071643

Paliogiannis, P., Fois, A. G., Sotgia, S., Mangoni, A. A., Zinellu, E., Pirina, P., Negri, S., Carru, C., & Zinellu, A. (2018). Neutrophil to lymphocyte ratio and clinical outcomes in COPD: recent evidence and future perspectives. *European Respiratory Review*, *27*(147), 170113. https://doi.org/10.1183/16000617.0113-2017

Prat, A., Parker, J. S., Karginova, O., Fan, C., Livasy, C., Herschkowitz, J. I., He, X., & Perou, C. M. (2010). Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast Cancer Research*. https://doi.org/10.1186/bcr2635

Prat, A., & Perou, C. M. (2011). Deconstructing the molecular portraits of breast cancer. In *Molecular Oncology*. https://doi.org/10.1016/j.molonc.2010.11.003

Ramsey, J., Glymour, M., Sanchez-Romero, R., & Glymour, C. (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*. https://doi.org/10.1007/s41060-016-0032-z

Rooney, M. S., Shukla, S. A., Wu, C. J., Getz, G., & Hacohen, N. (2015). Molecular and Genetic Properties of Tumors Associated with Local Immune Cytolytic Activity. *Cell*, *160*(1), 48–61. https://doi.org/https://doi.org/10.1016/j.cell.2014.12.033

Rueda, O. M., Sammut, S.-J., Seoane, J. A., Chin, S.-F., Caswell-Jin, J. L., Callari, M., Batra, R., Pereira, B., Bruna, A., Ali, H. R., Provenzano, E., Liu, B., Parisien, M., Gillett, C., McKinney, S., Green, A. R., Murphy, L., Purushotham, A., Ellis, I. O., … Curtis, C. (2019). Dynamics of breast-cancer relapse reveal late-recurring ER-positive genomic subgroups. *Nature*, *567*(7748), 399–404. https://doi.org/10.1038/s41586-019-1007-8

Rye, I. H., Huse, K., Josefsson, S. E., Kildal, W., Danielsen, H. E., Schlichting, E., Garred, Ø., Riis, M. L., Osbreac, Lingjaerde, O. C., Myklebust, J. H., & Russnes, H. G. (2022). Breast cancer metastasis: immune profiling of lymph nodes reveals exhaustion of effector T cells and immunosuppression. *Molecular Oncology*, *16*(1), 88–103. https://doi.org/10.1002/1878-0261.13047

Saal, L. H., Vallon-Christersson, J., Häkkinen, J., Hegardt, C., Grabau, D., Winter, C., Brueffer, C., Tang, M.-H. E., Reuterswärd, C., Schulz, R., Karlsson, A., Ehinger, A., Malina, J., Manjer, J., Malmberg, M., Larsson, C., Rydén, L., Loman, N., & Borg, Å. (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. https://doi.org/10.1186/s13073-015-0131-9

Sedgewick, A. J., Shi, I., Donovan, R. M., & Benos, P. V. (2016). Learning mixed graphical models with separate sparsity parameters and stability-based model selection. *BMC Bioinformatics*. https://doi.org/10.1186/s12859-016-1039-0

Srivatsan, S. R., McFaline-Figueroa, J. L., Ramani, V., Saunders, L., Cao, J., Packer, J., Pliner, H. A., Jackson, D. L., Daza, R. M., Christiansen, L., Zhang, F., Steemers, F., Shendure, J., & Trapnell, C. (2020). Massively multiplex chemical transcriptomics at single-cell resolution. *Science*, *367*(6473), 45–51. https://doi.org/10.1126/science.aax6234

Steinke, F. C., Yu, S., Zhou, X., He, B., Yang, W., Zhou, B., Kawamoto, H., Zhu, J., Tan, K., & Xue, H.-H. (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), 646–656. https://doi.org/10.1038/ni.2897

Tenenbaum D, M. B. (2022). *KEGGREST: Client-side REST access to the Kyoto Encyclopedia of Genes and Genomes (KEGG)*. https://doi.org/https://doi.org/doi:10.18129/B9.bioc.KEGGREST

Therneau, T. M. (2022). *A Package for Survival Analysis in R*.

Urban, G., Golden, T., Aragon, I. V., Cowsert, L., Cooper, S. R., Dean, N. M., & Honkanen, R. E. (2003). Identification of a Functional Link for the p53 Tumor Suppressor Protein in Dexamethasone-induced Growth Suppression. *Journal of Biological Chemistry*, *278*(11), 9747–9753. https://doi.org/10.1074/jbc.M210993200

van de Wetering, M., de Lau, W., & Clevers, H. (2002). WNT Signaling and Lymphocyte Development. *Cell*, *109*(2), S13–S19. https://doi.org/10.1016/S0092-8674(02)00709-2

Vestbo, J., Anderson, W., Coxson, H. O., Crim, C., Dawber, F., Edwards, L., Hagan, G., Knobil, K., Lomas, D. A., MacNee, W., Silverman, E. K., & Tal-Singer, R. (2008). Evaluation of COPD Longitudinally to Identify Predictive Surrogate End-points (ECLIPSE). *European Respiratory Journal*, *31*(4), 869–873. https://doi.org/10.1183/09031936.00111707

Wang, Z., Yu, Z., Wang, G.-H., Zhou, Y.-M., Deng, J.-P., Feng, Y., Chen, J.-Q., & Tian, L. (2020). AURKB Promotes the Metastasis of Gastric Cancer, Possibly by Inducing EMT. *Cancer Management and Research*, *12*, 6947–6958. https://doi.org/10.2147/CMAR.S254250

Wiener, A. L. and M. (2002). Classification and Regression by randomForest. *R News*, *2*.

Wu, S. Z., Al-Eryani, G., Roden, D. L., Junankar, S., Harvey, K., Andersson, A., Thennavan, A., Wang, C., Torpy, J. R., Bartonicek, N., Wang, T., Larsson, L., Kaczorowski, D., Weisenfeld, N. I., Uytingco, C. R., Chew, J. G., Bent, Z. W., Chan, C.-L., Gnanasambandapillai, V., … Swarbrick, A. (2021). A single-cell and spatially resolved atlas of human breast cancers. *Nature Genetics*, *53*(9), 1334–1347. https://doi.org/10.1038/s41588-021-00911-1

Xie, Z., Bailey, A., Kuleshov, M. V., Clarke, D. J. B., Evangelista, J. E., Jenkins, S. L., Lachmann, A., Wojciechowicz, M. L., Kropiwnicki, E., Jagodnik, K. M., Jeon, M., & Ma'ayan, A. (2021). Gene Set Knowledge Discovery with Enrichr. *Current Protocols*, *1*(3). https://doi.org/10.1002/cpz1.90

Xu, X., Zhang, M., Xu, F., & Jiang, S. (2020). Wnt signaling in breast cancer: biological mechanisms, challenges and opportunities. *Molecular Cancer*, *19*(1), 165. https://doi.org/10.1186/s12943-020-01276-5

Yotam, D., Michal, S., & Eytan, D. (2013). Pathway-based personalized analysis of cancer. *Proceedings of the National Academy of Sciences*, *110*(16), 6388–6393. https://doi.org/10.1073/pnas.1219651110

Zappia, L., & Oshlack, A. (2018). Clustering trees: a visualization for evaluating clusterings at multiple resolutions. *GigaScience*, *7*(7), giy083. https://doi.org/10.1093/gigascience/giy083

Zhang, N., Wei, P., Gong, A., Chiu, W.-T., Lee, H.-T., Colman, H., Huang, H., Xue, J., Liu, M., Wang, Y., Sawaya, R., Xie, K., Yung, W. K. A., Medema, R. H., He, X., & Huang, S. (2011). FoxM1 promotes β-catenin nuclear localization and controls Wnt target-gene expression and glioma tumorigenesis. *Cancer Cell*, *20*(4), 427–442. https://doi.org/10.1016/j.ccr.2011.08.016

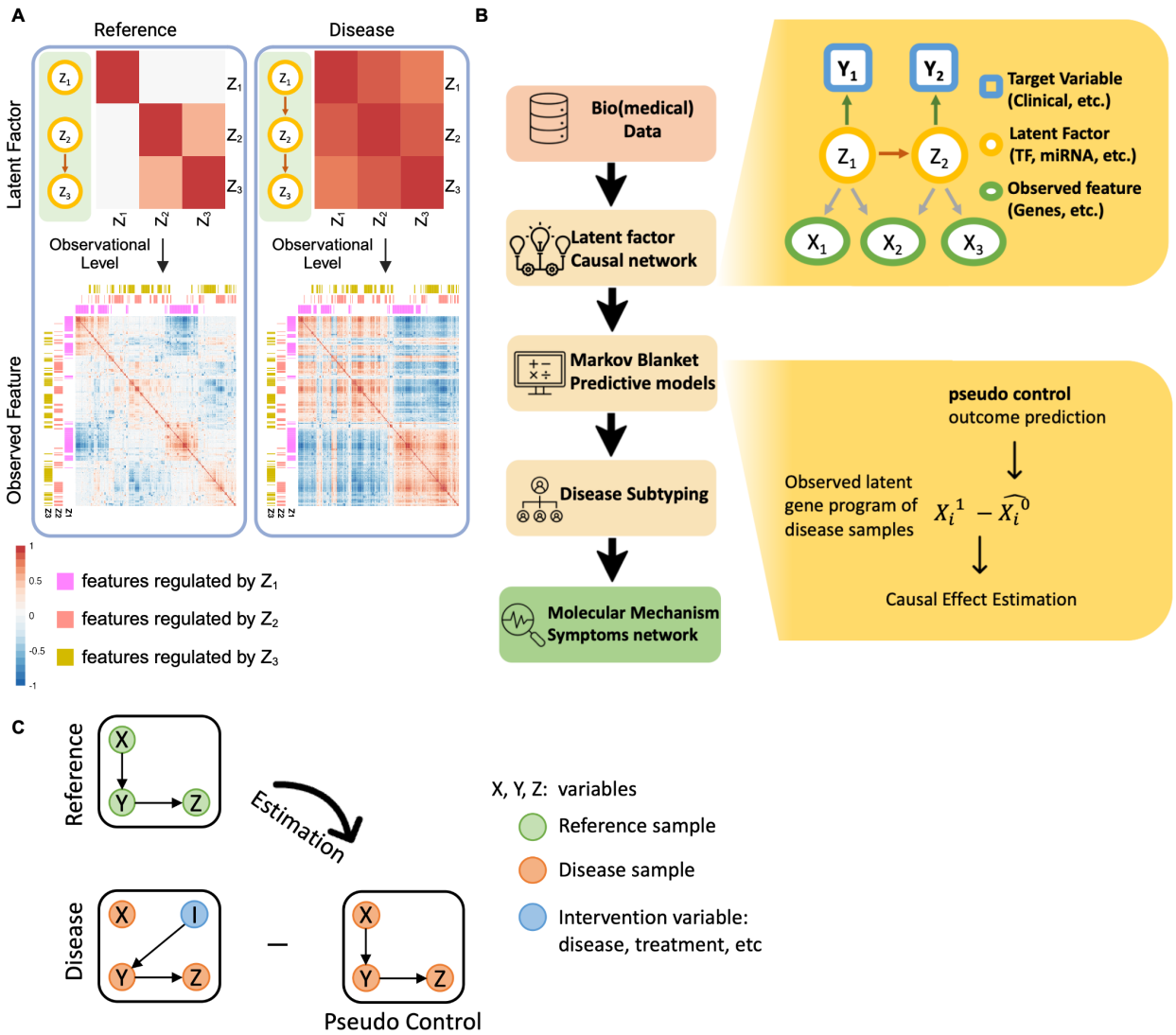

Figure 1. Change in gene program interactions and algorithm overview of LaGrACE.
(A) Three latent factors (LFs) regulate a set of observed features separately in reference condition (left). If interactions among three LFs were changed in disease (or treatment) condition (right), the correlation of regulated 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.

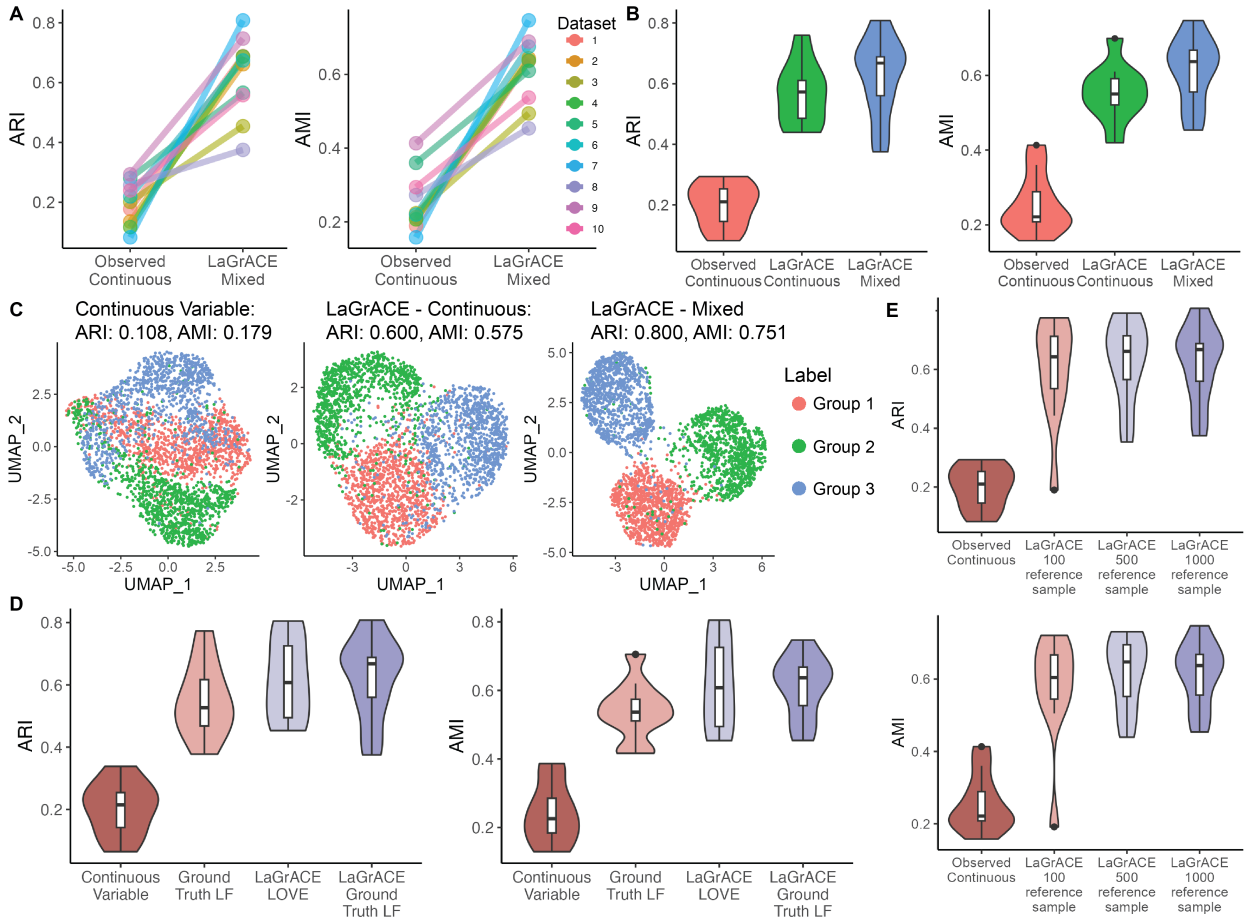

Figure 2. Evaluation of LaGrACE's subtyping performance on synthetic dataset with 15% edge change.

(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.

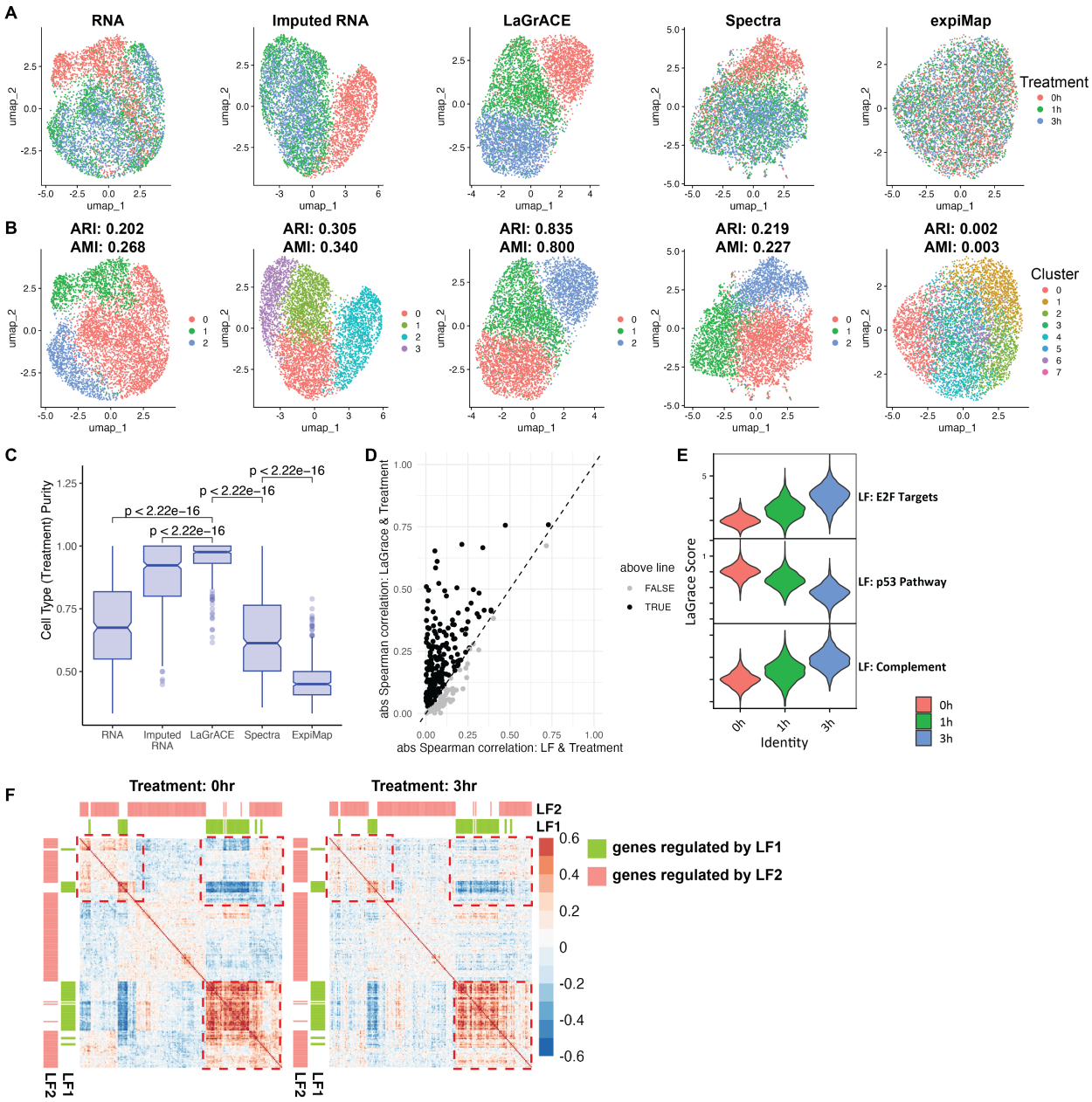

Figure 3. LaGrACE captures treatment effect at single-cell resolution.

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: 0.002, AMI: 0.003).

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

(D) 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)

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

(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.

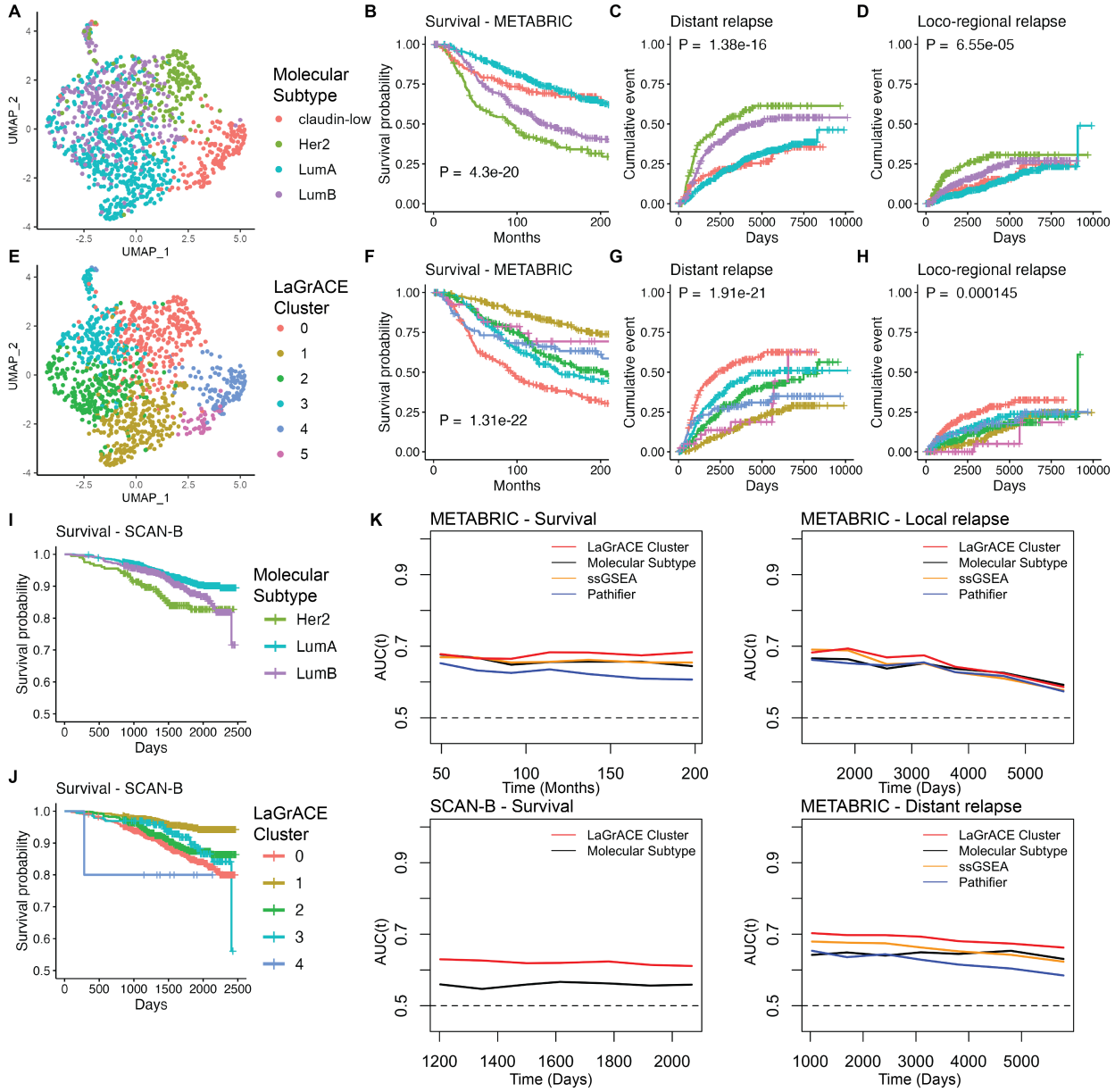

Figure 4. LaGrACE identified six clusters on breast cancer dataset.

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

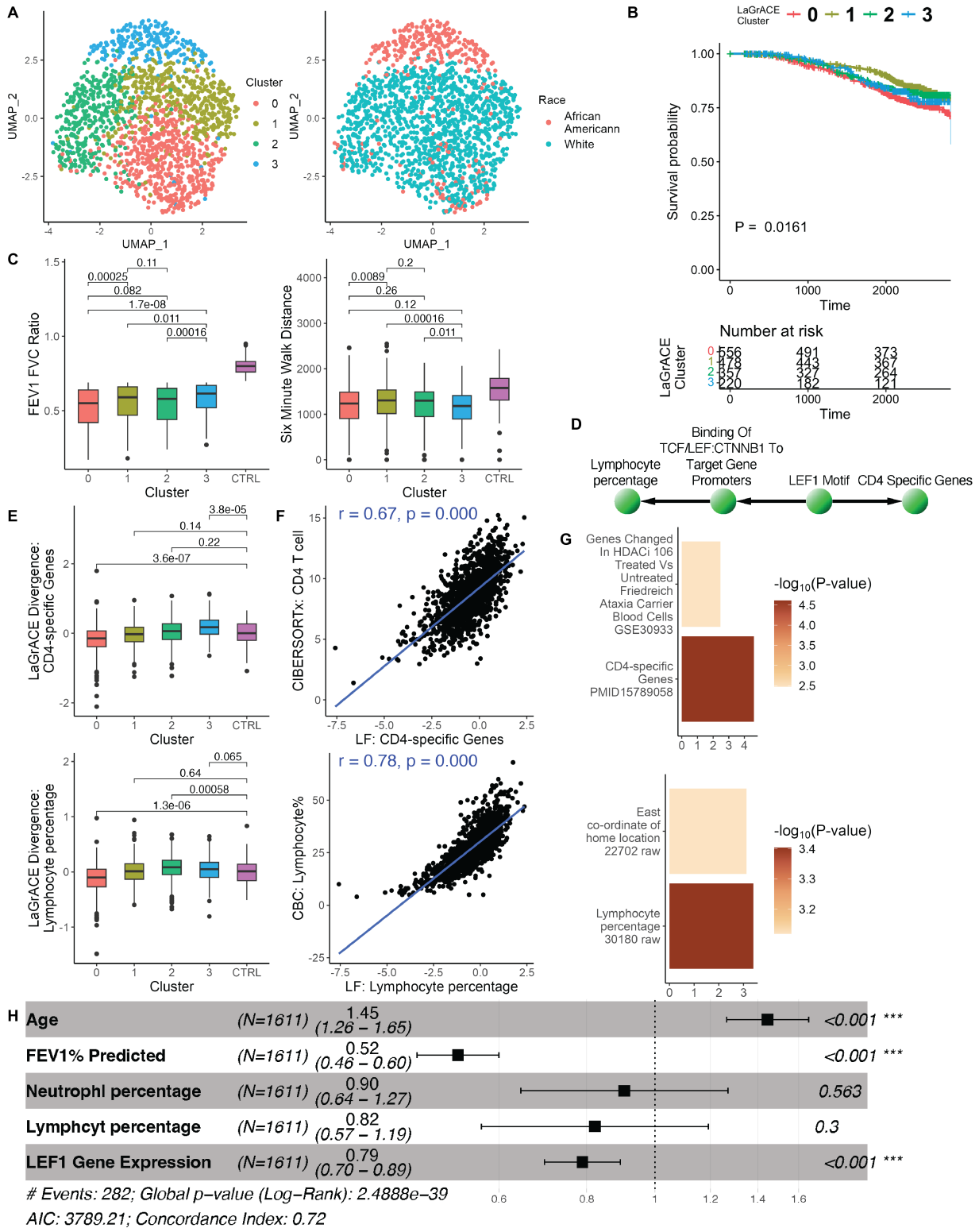

Figure 5. LaGrACE identified novel clusters on COPD dataset

(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 group. 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**