Bioinformatics, 37(6), 2021, 785–792
doi: 10.1093/bioinformatics/btaa886
Advance Access Publication Date: 18 October 2020
Original Paper

OXFORD

Genetics and population analysis

Unpaired data empowers association tests

Mingming Gong[1,2,3,†], Peng Liu[1], Frank C. Sciurba[1], Petar Stojanov[2], Dacheng Tao[4], George C. Tseng [1], Kun Zhang[2] and Kayhan Batmanghelich[1,*]

[1]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA 15206, USA, [2]Department of Philosophy, Carnegie Mellon University, Pittsburgh, PA 15213, USA, [3]School of Mathematics and Statistics, The University of Melbourne, Melbourne, VIC 3010, Australia and [4]Australia School of Computer Science, The University of Sydney, Sydney, NSW 2006, Australia

*To whom correspondence should be addressed.

[†]Most of the work was performed as a postdoc at University of Pittsburgh and Carnegie Mellon University.

Associate Editor: Valencia Alfonso

Received on February 25, 2020; revised on September 7, 2020; editorial decision on September 26, 2020; accepted on October 5, 2020

Abstract

Motivation: There is growing interest in the biomedical research community to incorporate retrospective data, available in healthcare systems, to shed light on associations between different biomarkers. Understanding the association between various types of biomedical data, such as genetic, blood biomarkers, imaging, etc. can provide a holistic understanding of human diseases. To formally test a hypothesized association between two types of data in Electronic Health Records (EHRs), one requires a substantial sample size with both data modalities to achieve a reasonable power. Current association test methods only allow using data from individuals who have both data modalities. Hence, researchers cannot take advantage of much larger EHR samples that includes individuals with at least one of the data types, which limits the power of the association test.

Results: We present a new method called the Semi-paired Association Test (SAT) that makes use of both paired and unpaired data. In contrast to classical approaches, incorporating unpaired data allows SAT to produce better control of false discovery and to improve the power of the association test. We study the properties of the new test theoretically and empirically, through a series of simulations and by applying our method on real studies in the context of Chronic Obstructive Pulmonary Disease. We are able to identify an association between the high-dimensional characterization of Computed Tomography chest images and several blood biomarkers as well as the expression of dozens of genes involved in the immune system.

Availability and implementation: Code is available on https://github.com/batmanlab/Semi-paired-Association-Test.

Contact: kayhan@pitt.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Increasingly, data from Electronic Health Records (EHRs) in hospitals are becoming available to clinical researchers. Such massive collections contain various types of data from sources such as high-resolution imaging, genome sequencing and physiological metrics. By studying such a large and diverse data, researchers can provide a holistic view of the underlying mechanisms of human diseases. For example, while a large proportion of human diseases are influenced by genetic variants (Altshuler *et al.*, 2008; Ehret *et al.*, 2011), their mechanisms are not well understood (Visscher *et al.*, 2017; Willer *et al.*, 2013). To understand the mechanism, measuring other variables such as gene expression is required. Unfortunately, it is unlikely that all patients in the EHR have all measurement modalities. For example, due to the high cost of image acquisition and specimen maintenance, hospitals order those only when they are needed. Consequently, only the record of a few patients contains all data modalities, which reduces the power of association tests and increases the chance of false discovery.

Furthermore, a multidimensional phenotype can offer better sensitivity to the clinical and genetic underpinning of human diseases than a one-dimensional scalar phenotype (Csernansky *et al.*, 1998; Ge *et al.*, 2016). For instance, high-dimensional features can be computed to summarize the folding pattern of the brain structure in Magnetic Resonance (MR) imaging (Ge *et al.*, 2016), or the texture and distribution of the lung tissue destruction can be measured and summarized by computed tomography (CT) imaging (Schabdach *et al.*, 2017). Those metrics are highly predictive of the diseases [e.g. Alzheimer's disease (Csernansky *et al.*, 2005) and bipolar disorder (Hwang *et al.*, 2006) for MR, and Chronic Obstructive Pulmonary Disease (COPD) (Schabdach *et al.*, 2017) for CT]. Relating that high-dimensional phenotype to genetic and genomic measurements provides more evidence for understanding the etiology of the disease.

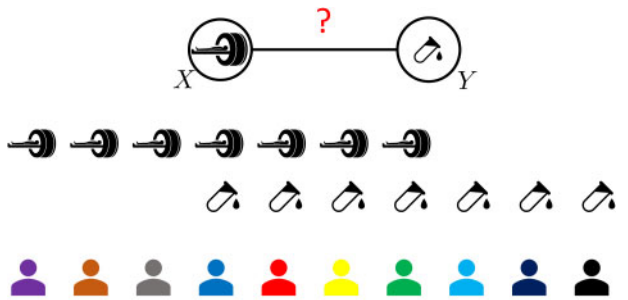

Fig. 1. X and Y represent two modalities. Current approaches only use paired data $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{n}$. Assuming that the total number of samples of X (M) and Y (N) is more than the paired data, we aim to find out how the control of the false discovery and the power of association tests can be improved by the unpaired data $\{\mathbf{x}_i\}_{i=n+1}^{M}$ and $\{\mathbf{y}_i\}_{i=n+1}^{N}$

In this article, we present a new method to formally test the association between two types of potentially high-dimensional data that allows incorporating unpaired samples, i.e. samples with one data modality (see Fig. 1 for a schematic illustration). Our approach provides better control of false-positive and, under some mild assumptions, increases the statistical power of the test. Unpaired data enables us to better estimate the null distribution, which results in more accurate control of the false positive rate. Furthermore, it allows us to leverage the underlying structure of the high-dimensional measurements, which consequently increases the power of the test. The proposed method, the Semi-paired Association Test (SAT), falls in the kernel machine framework (Ge *et al.*, 2015; Gretton *et al.*, 2008a; Liu *et al.*, 2007; Székely *et al.*, 2007; Zhang et al., 2011). More specifically, two variants of our method generalize the Variance Component Score Test (VCST) (Ge *et al.*, 2015; Kwee *et al.*, 2008; Liu *et al.*, 2007) and the Kernel Independence Test (KIT) (Gretton *et al.*, 2008a; Székely *et al.*, 2007; Zhang *et al.*, 2011) such that they can exploit unpaired data. The VCST is commonly used to test for heritability of a phenotype (Ge *et al.*, 2015; Kwee *et al.*, 2008; Liu *et al.*, 2007) and is implemented in popular software such as GCTA (Yang et al., 2011). The KIT is widely used for statistical independence test in various scenarios (Gretton *et al.*, 2008a; Hua and Ghosh, 2015; Wei and Lu, 2017). We provide a connection between those methods. Our proposed test makes unpaired data, previously wasted, available for discovering novel associations in massive uncontrolled datasets, such as EHRs. Unearthing unnoticed associations assists in understanding the underlying mechanism of human diseases.

This article makes two contributions. First, it provides a statistically grounded method for the inclusion of unpaired data. The extensive simulation, as well as theoretical study, supports the hypothesis that the unpaired data is beneficial to control the false discovery and if the conditions are satisfied, can improve the power. Second, we apply our method to two different real studies. In the first experiment, we show that unpaired data can discover a new association between the high-dimensional radiographic measurements of COPD and peripheral blood biomarkers that play a role in the immune system. In this dataset, only a subset of the cohort has blood samples. In the second experiment, we apply our approach to genotype-phenotype data from the General Population Cohort from Uganda (Asiki *et al.*, 2013). In this dataset, all subjects have genotype data but only one-fourth of them have phenotypes. Our method is able to find more heritable phenotypes.

2 Materials and methods

In this section, we first give a brief review of the VCST and the kernel independence test (KIT). We then discuss the connections between them and show that the differences between them lead to different ways to utilize unpaired data. Finally, we detail our SAT method by demonstrating how unpaired data can be incorporated to improve both VCST and KIT.

2.1 Variance component score test (VCST)

We start with the variance component model (a.k.a. the random effect model), which is widely used in statistical genetics for genetic association studies (Ge *et al.*, 2016; Liu *et al.*, 2007; Maity *et al.*, 2012; Yang *et al.*, 2011). We use the same nomenclature where $Y \in \mathbb{R}^p$ is a p-dimensional phenotype and $X \in \mathbb{R}^d$ is genotype. However, our method is general and can be applied elsewhere. Given a paired sample containing n observations $\{\mathbf{y}_i, \mathbf{x}_i\}_{i=1}^{n}$, we consider the following multidimensional variance component model (Ge *et al.*, 2016):

$$y_{ik} = \mu_{ik} + g_k(\mathbf{x}_i) + \epsilon_{ik}, \tag{1}$$

where y_{ik} is the kth element of $\mathbf{y}_i$, g_k is a non-parametric function in a reproducing kernel Hilbert space (RKHS) associated with kernel $k(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle$, μ_{ik} is the offset term and ϵ_{ik} is the error term. Equation (1) can be rewritten in matrix form:

$$\mathbf{Y} = \boldsymbol{\mu} + \mathbf{G} + \epsilon, \tag{2}$$

where $\mathbf{Y} \in \mathbb{R}^{n \times p}$ is the phenotypic matrix of the n observations (subjects) with ith row $\mathbf{y}_i^\top$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p) \otimes 1_n$ is a matrix of offsets (1_n is an $n \times 1$ vector of ones), $\mathbf{G} \in \mathbb{R}^{n \times p}$ is the matrix of the aggregate genetic effects and $\epsilon \in \mathbb{R}^{n \times p}$ is a matrix of residual effects. We have the following distributional assumptions:

$$\mathrm{vec}(\mathbf{G}) \sim \mathcal{N}(0, \boldsymbol{\Sigma}_g \otimes \mathbf{K}), \quad \mathrm{vec}(\epsilon) \sim \mathcal{N}(0, \boldsymbol{\Sigma}_\epsilon \otimes \mathbf{I}_n), \tag{3}$$

where $\mathrm{vec}(\cdot)$ is the matrix vectorization operator that converts a matrix into a vector by stacking its columns, $\otimes$ is the Kronecker product of matrices, $\mathbf{I}_n$ denotes an $n \times n$ identity matrix, $\boldsymbol{\Sigma}_g$ is the genetic covariance matrix, $\boldsymbol{\Sigma}_\epsilon$ is the residual covariance matrix and $\mathbf{K}$ is the kernel matrix with ijth element $[\mathbf{K}]_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$. For example, in the context of statistical genetics, $\mathbf{K}$ denotes identity-by-state (IBS) kernel (Kwee *et al.*, 2008; Schaid, 2010a,b), where $[\mathbf{K}]_{ij}$ represents the relatedness between individual i and j.

To test whether Y and X are associated (whether Y is heritable if X is the genotype), we can test the variance components as $\mathcal{H}_0 : tr(\boldsymbol{\Sigma}_g) = 0$ versus $\mathcal{H}_1 : tr(\boldsymbol{\Sigma}_g) > 0$ using the following score test statistic derived from model (1):

$$\hat{\mathcal{S}}_n(\mathbf{K}, \mathbf{L}) = \frac{1}{n^2} tr(\mathbf{K}\mathbf{H}_n\mathbf{L}\mathbf{H}_n) - \frac{1}{n^3} tr(\mathbf{H}_n\mathbf{L}) tr(\mathbf{H}_n\mathbf{K}), \tag{4}$$

where $tr(\cdot)$ computes the trace of a matrix, $\mathbf{L} = \mathbf{Y}\hat{\boldsymbol{\Sigma}}_Y^{-2}\mathbf{Y}^\top$ and $\mathbf{H}_n = \mathbf{I}_n - \frac{1}{n} 1_n 1_n^\top$ and $\hat{\boldsymbol{\Sigma}}_Y$ is the empirical covariance matrix of Y. The derivation details are provided in Supplementary Section S1 of Supplementary Information. The exact fraction of phenotype variability attributed to genetic variation is defined as heritability. There are various ways to define heritability for a multivariate phenotype (e.g. Ge *et al.*, 2016; Zhou et al., 2013). We adopt the definition by Ge *et al.* (2016) that closely related to the VCST and subsumes the definition of the heritability for the univariate phenotype, which can be calculated as follows (Ge *et al.*, 2016):

$$h^2 = tr\left(\frac{\boldsymbol{\Sigma}_g}{tr(\boldsymbol{\Sigma}_g) + tr(\boldsymbol{\Sigma}_\epsilon)}\right). \tag{5}$$

2.2 Kernel independence test (KIT)

Kernel independence tests are a class of non-parametric methods which are also widely used for genetic association studies (Gretton *et al.*, 2008a; Wei and Lu, 2017). Here we briefly review the Hilbert-Schmidt Independence Criterion (HSIC)-based independence test (Gretton *et al.*, 2008a), which provides a general framework for many association tests (Sejdinovic et al., 2013). Let $\mathcal{F}_y$ be a RKHS associated with the kernel function $l(\mathbf{y}, \mathbf{y}') = \langle \psi(\mathbf{y}), \psi(\mathbf{y}') \rangle$. HSIC tests $\mathcal{H}_0 : \mathbb{P}_{YX} = \mathbb{P}_X\mathbb{P}_Y$ versus $\mathcal{H}_1 : \mathbb{P}_{YX} \neq \mathbb{P}_X\mathbb{P}_Y$ by testing $\mathcal{H}_0 : I = 0$ versus $\mathcal{H}_1 : I > 0$, where I is defined as follows:

Table 1. Comparison of VCST and KIT

	Test statistic (unbiased)	Null distribution (unbiased)	Test statistic (biased)	Null distribution (biased)	Unpaired X	Unpaired Y
VCST	$\hat{S}_n(\mathbf{K},\mathbf{L})$	$\lambda_i\hat{\eta}_j(z_{ij}^2-1)$	$\frac{1}{n^2}tr(\mathbf{K}\mathbf{H}_n\mathbf{L}\mathbf{H}_n)$	$\lambda_i\hat{\eta}_j z_{ij}^2$	✗	✓
KIT	$\hat{I}_n(\mathbf{K},\mathbf{L})$	$\lambda_i\eta_j(z_{ij}^2-1)$	$\frac{1}{n^2}tr(\mathbf{K}\mathbf{H}_n\mathbf{L}\mathbf{H}_n)$	$\lambda_i\eta_j z_{ij}^2$	✓	✓

$$\begin{aligned} I = &\mathbb{E}_{XY}\mathbb{E}_{X'Y'}[k(X,X')l(Y,Y')] \\ &+\mathbb{E}_X\mathbb{E}_{X'}\mathbb{E}_Y\mathbb{E}_{Y'}[k(X,X')l(Y,Y')] \\ &-2\mathbb{E}_{XY}[\mathbb{E}_{X'}[k(X,X')]\mathbb{E}_{Y'}[l(Y,Y')]]. \end{aligned} \tag{6}$$

Given paired data of n subjects, an unbiased estimator of I is the following (Gretton *et al.*, 2008b):

$$\hat{I}_n(\mathbf{K},\mathbf{L}) = \frac{1}{n(n-3)}\left[tr(\tilde{\mathbf{K}}\tilde{\mathbf{L}}) + \frac{1_n^\top\tilde{\mathbf{K}}1_n 1_n^\top\tilde{\mathbf{L}}1_n}{(n-1)(n-2)} - \frac{2 1_n^\top\tilde{\mathbf{K}}\tilde{\mathbf{L}}1_n}{n-2}\right], \tag{7}$$

where $\tilde{\mathbf{K}} = \mathbf{K} - diag(\mathbf{K})$ and similarly for $\tilde{\mathbf{L}}$ and $\mathbf{L}_{ij} = l(\mathbf{y}_i,\mathbf{y}_j)$. To test for statistical independence, one can use characteristic kernels, e.g. the radial basis function $\mathbf{K}_{ij} = \exp\left(\frac{\|\mathbf{x}_i-\mathbf{x}_j\|^2}{\sigma^2}\right)$, such that I can be zero only when X and Y are independent (Sriperumbudur et al., 2011).

2.3 Connections between VCST and KIT

Now we discuss the similarities and differences between VCST and KIT. Table 1 displays the test statistics and null distributions of VCST and KIT.

Test statistic It can be seen from Table 1 that the biased statistics of VCST and KIT are identical to each other, if setting $\psi(\mathbf{y}) = \hat{\mathbf{\Sigma}}_Y^{-1}\mathbf{y}$. The unbiased test statistics of VCST and KIT differ. This is because VCST tests for random effects but assumes that the covariate inducing the random effect (X) and the corresponding kernel matrix ($\mathbf{K}$) are fixed while KIT assumes X is random, leading to different ways to correct for the bias.

Null distribution Let η_j ($\hat{\eta}_j$) be the eigenvalues (empirical) of the covariance of $\phi(X)$ and let λ_i ($\hat{\lambda}_j$) be the eigenvalues (empirical) of the covariance of $\psi(Y)$. As shown in Table 1, the null distributions for VCST and KIT have exactly the same forms, except that VCST uses $\hat{\eta}_j$ while KIT uses η_j. This is also because of their respective fixed or random X assumptions. In practice, because λ_i and η_j are both unknown, we need to replace them with $\hat{\lambda}_i$ and $\hat{\eta}_i$. Therefore, the empirical null distributions of VCST and KIT are identical if only given n paired examples. However, they are inherently different because the null distribution of KIT is derived from asymptotic theory, while the null distribution of VCST is derived from the Gaussian error terms in the variance component model (2). This subtle difference is significant when using unpaired data, which is described as follows.

Unpaired data The main difference between VCST and KIT is that X ($\mathbf{K}$) is considered fixed or random respectively. When given unpaired data, VCST cannot make use of the unpaired data of X due to the fixed X assumption, while KIT can benefit from unpaired data of both X and Y. More specifically, unpaired data can only be used to improve the estimation of λ_i in VCST but they can be used to improve the estimation of both η_j and λ_i in KIT.

2.4 Semi-paired association test

In this section, we present our SAT method that incorporates unpaired data to improve test power. In addition to the n paired data, suppose we also have access to an unpaired sample $\{\mathbf{x}_i\}_{i=n+1}^N$ and an unpaired sample $\{\mathbf{y}_i\}_{i=n+1}^M$. Without loss of generality, we assume $N=M$ and replace M with N for notational simplicity. We will show two ways that unpaired data can improve the association test: (i) better control of type I error by improving the estimation of null distributions and (ii) improved test power by devising a new test statistic under the intrinsic low-dimension assumption of high-dimensional data.

Enhancing type I error control To calculate P-values, we need to estimate the parameters λ_i and η_j in the null distributions from empirical data. Because λ_i and η_j are the eigenvalues of the covariance of $\psi(Y)$ and $\phi(X)$, respectively, the estimation does not require paired Y and X examples. Therefore, we can readily make use of unpaired data to obtain more accurate estimation of λ_i or η_j involved in the null distribution.

For SAT-fx, we add unpaired Y data to estimate the covariance of $\psi(Y)$ and its eigenvalues λ_i from both paired and unpaired data $\{\mathbf{y}_i\}_{i=1}^N$, while η_j should be estimated from only $\{\mathbf{x}_i\}_{i=1}^n$ in the paired sample. For SAT-rx, we can further incorporate unpaired X data and use all the X data $\{\mathbf{x}_i\}_{i=1}^N$ to estimate η_j.

The following theorem shows that (i) the empirical null distribution convergences to the true (asymptotic) distribution and (ii) the variance of the empirical null distribution converges to the variance of the true (asymptotic) null distribution with rate $1/\sqrt{m}$, where m is the sample size of available data for estimating λ_i and η_j.

THEOREM 1 (**Informal**) *Let* $I = \sum_{i=1}^{p}\sum_{j=1}^{q}\lambda_i\eta_j(z_{ij}^2-1)$ *and* $I_m = \sum_{i=1}^{p}\sum_{j=1}^{q}\hat{\lambda}_i\,\hat{\eta}_j(z_{ij}^2-1)$, *where p is the dimension of* $\psi(Y)$ *and q is the dimension of* $\phi(X)$.

1. *As* $m\to\infty$, I_m *converges in distribution to* I.
2. *For all* $\mathbb{P}_{XY}$, $\mathbb{E}(I_m) = \mathbb{E}(I)$ *and* $\mathbb{V}(I_m)$ *converges in probability to* $\mathbb{V}(I)$ *with rate* $1/\sqrt{m}$.

The theorem is developed for SAT-rx and a similar theorem for SAT-fx can be considered as a special case of the above theorem. From the theorem, we can see that if only using paired data, $m=n$; if further using unpaired data, $m=N$. Because $N>n$, incorporating unpaired data to estimate λ_i and η_j leads to lower estimation error and provides more accurate estimation of the null distribution. Hence, our method results in better control of the type I error. This result holds regardless of the dimensionality of X or Y. The proof details of Theorem 1 are given in Supplementary Section S2 of Supplementary Information.

Improving test power Unpaired data contribute to a better estimation of the null distribution, resulting in better control of type I error. It can also improve test power if the high-dimensional data (of at least one modality) live on a lower dimensional space. Such an assumption is mostly the case for real data. For example, previous studies have shown that the space of Magnetic Resonance images of the brain can be modeled by a relatively low-dimensional manifold (Amir *et al.*, 2013; Gerber *et al.*, 2010). A similar assumption has been explored to model the low-dimensional space of gene expression for single-cell expression analysis (Haghverdi *et al.*, 2015; Qiu *et al.*, 2011). Unpaired data help us to estimate the low-dimensional space more accurately. If both X and Y are one-dimensional, our method cannot improve the test power, but the variant without dimension reduction can better control the false discovery rate.

Specifically, if X or Y data (approximately) lie in a low-dimensional space, we show that unpaired data can be used to construct a new test statistic with improved test power. To devise the new test statistics, we first learn the low-dimensional space of X or Y by applying the kernel Principal Component Analysis (PCA) algorithm on both paired and unpaired data. Second, we project the paired data to the learned low-dimensional space and obtain the test statistics of our SAT-fx and SAT-rx by estimating the test statistics of VCST and KIT on the projected data. Due to the use of the kernel trick, calculating the test statistic of SAT-fx and SAT-rx requires

only the kernel matrices $\mathbf{K}_N$ and $\mathbf{L}_M$ which are calculated on all the data, paired and unpaired.

In SAT-fx, because we do not consider X as random as does VCST, we can only incorporate unpaired Y data to learn the low-dimensional structure of Y. In SAT-rx, we further use unpaired data X to learn the low-dimensional space of X. The proposed new test statistics of SAT-fx and SAT-rx have the same form as that of VCST (4) and KIT (7), respectively. We only need to change the kernel matrices in the test statistics. Specifically, the new test statistic for SAT-fx is defined as $\hat{S}_n(\mathbf{K}, \mathbf{L}')$, where

$$\mathbf{L}' = \overline{\mathbf{L}}^\top \mathbf{U} \mathbf{\Lambda}_y^{-1} \mathbf{U}^\top \overline{\mathbf{L}}. \tag{8}$$

In $\mathbf{L}', \overline{\mathbf{L}}$ is the matrix comprised of the first n columns of $\mathbf{L}_N$, $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_{r_Y})$ and $\mathbf{\Lambda}_y = \text{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_{r_Y})$ are the top r_Y eigenvectors and eigenvalues of $\mathbf{L}_N$.

Similarly, the new test statistic of SAT-rx that considers X as random is $\hat{I}_n(\mathbf{K}', \mathbf{L}')$, where

$$\mathbf{K}' = \overline{\mathbf{K}}^\top \mathbf{V} \mathbf{\Lambda}_x^{-1} \mathbf{V}^\top \overline{\mathbf{K}}. \tag{9}$$

In $\mathbf{K}', \overline{\mathbf{K}}$ is the matrix composed of the first n columns of $\mathbf{K}_N$, $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_{r_X})$ and $\mathbf{\Lambda}_x = \text{diag}(\hat{\eta}_1, \ldots, \hat{\eta}_{r_X})$ are the top r_X eigenvectors and eigenvalues of $\mathbf{K}_N$. The asymptotic null distributions of the proposed $\hat{S}_n'$ and $\hat{I}_n'$ have the same forms as the null distributions of $\hat{S}_n$ and $\hat{I}_n$, but using only the top eigenvalues $\{\lambda_i\}_{i=1}^{r_Y}$ and $\{\eta_j\}_{j=1}^{r_X}$, respectively. The derivation details are provided in Supplementary Section S3 of Supplementary Information.

The following theorem shows that the power of the new test statistic of SAT-rx is greater than the classical one that only uses paired data.

THEOREM 2 (**Informal**) Assuming that data from X and Y lie in a low-dimensional manifold, the test power of the proposed SAT-rx is higher than that of the KIT method, which only uses paired data.

SAT-fx follows similar properties as SAT-rx and can be considered as a special case of SAT-rx. The proof details of Theorem 2 are given in Supplementary Section S4 of Supplementary Information. The main steps of our method SAT-rx is summarized in Algorithm 1. The algorithm for SAT-fx is similar to that of SAT-rx, except that SAT-fx does not use the unpaired X data.

Algorithm 1: Semi-paired association test (SAT-rx)

Data: Paired data $\{\mathbf{y}_i, \mathbf{x}_i\}_{i=1}^n$ and unpaired data $\{\mathbf{x}_i\}_{i=n+1}^N$ and $\{\mathbf{y}_i\}_{i=n+1}^N$, the dimension of X after dimension reduction r_X, the dimension of Y after dimension reduction r_Y.

1. According to kernel functions $k(\mathbf{x}_i, \mathbf{x}_j)$ and $l(\mathbf{y}_i, \mathbf{y}_j)$ compute kernel matrices $\mathbf{K}$ and $\mathbf{L}$ on the paired data and $\mathbf{K}_N$ and $\mathbf{L}_N$ on both paired and unpaired data
2. Compute kernel matrices $\mathbf{K}'$ and $\mathbf{L}'$ according to Equation (8) and (9)
3. Compute the test statistic $\hat{I}_n(\mathbf{K}', \mathbf{L}')$ according to Equation (7)
4. Perform eigendecomposition of $\mathbf{K}_N$ and get the top r_X eigenvalues $\hat{\eta}_{j=1}^{r_X}$. Perform eigendecomposition of $\mathbf{L}_N$ and get the top r_Y eigenvalues $\hat{\lambda}_{j=1}^{r_Y}$
5. Sample data points $\{b_i\}_{i=1}^B$ from the asymptotic distribution $\sum_{i=1}^{r_Y}\sum_{j=1}^{r_X} \hat{\lambda}_i \hat{\eta}_j (z_{ij}^2 - 1)$, calculate the ratio R of data points greater than the test statistic $\hat{I}_n(\mathbf{K}', \mathbf{L}')$
6. Return R as the estimated P-value;

3 Simulation study

3.1 Simulation method

To evaluate our method's improvement of type I and type II errors, we mimic the data missingness mechanism by conducting two levels of simulations:

1. We synthesize both modalities X and Y. In this simulation, we evaluate both variants of our method, including SAT-fx and SAT-rx.
2. Following the literature of population genetics in which testing for the heritability of traits is a topic of interest, we use genotype data as X and synthesize Y. We only evaluate SAT-fx because X is fixed.

In simulation (1), to generate X, we first generate N low-dimensional (dim = 10) data points from a Gaussian distribution and then map them to high-dimensional X using a linear transformation plus independent Gaussian noise. To generate Y, we first generate low-dimensional data according to the variance components model (see Equation 1 in the Section 2) and then map them to high-dimensional Y using another linear transformation plus independent Gaussian noise.

In simulation (2), we use genotype data from the COPDGene study (Regan *et al.*, 2011) to simulate low-dimensional phenotype data. The COPDGene study recruited 6751 subjects with Non-Hispanic White (NHW) and 3395 subjects with non-Hispanic African American (AA) Backgrounds. The platform used for data collection is Illumina HumanOmni1 Quad v1. Since the Linkage Disequilibrium (LD) pattern in AA and NHW sub-population are different (Shifman *et al.*, 2003), we focus on the NHW sub-population to avoid introducing confounder to the simulation. We use the standard quality-controlled data that removes SNPs with MAF $< 5\%$ or missing rate $> 1\%$ and SNPs that deviate from Hardy-Weinberg equilibrium. We follow the generative model of a polygenic disease where the polygenic quantitative trait y is modeled with the model 1. The $g_k(\mathbf{x}_i) = \sum_{j \in \mathcal{C}} z_{ij} u_j$ is the genetic caused by $\mathcal{C}$ causal SNPs with effect size u_j and z_{ij} is the genotype of j'th SNP of the i'th individual. We assume that $\mathcal{C}$ is spread out across all SNPs. To control for population structure, we follow the common practice in the human population genetics and use the top six principal components of the relatedness matrix as covariates. To mimic the low-dimensional structure of Y, we map the generated low-dimensional phenotypes to high-dimensional Y using a linear transformation plus independent Gaussian noise.

In all the simulations, we create 1000 simulation replicates to evaluate the type I error rate and test power. Type I error rates and powers are calculated using the percentage of P-values smaller than a given significance level ($\alpha = 0.05$) under null models and alternative models, respectively. We set the heritability $h^2 = 0$ for the evaluation of type I error rates and $h^2 = 0.1$ for the evaluation of power. To show the benefits of incorporating unpaired data, we compare the type I errors of the VCST/KIT as a baseline with two variants of both SAT-fx and SAT-rx: with and without Dimensionality Reduction (DR). VCST and KIT only use n paired data points, while SAT-fx and SAT-rx use n paired data points together with an additional $N - n$ unpaired data points. For evaluation, we have access to the oracle where we can apply VCST and KIT using all N data points as paired, which is the best we can achieve. We set $n = 100$ for simulation (1) and $n = 3000$ for simulation (2).

3.2 Simulation results

Figures 2 and 3 report the type I error rates and power in simulation (1), respectively. Here we only show results for random X. The results for fixed X have similar trends and are available in Supplementary Section S5 of Supplementary Information. The results in Figure 2 demonstrate that the type I error rates of our proposed method approach the predefined significance level (0.05) as we add more unpaired data. In addition, Figure 3 shows that our

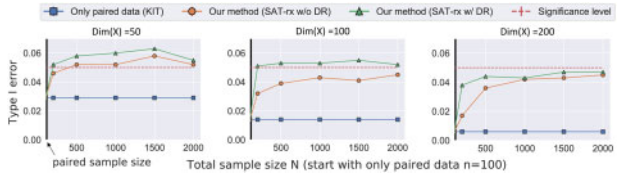

Fig. 2. Evaluation of SAT-rx type I error rate control on the simulated data generated by procedure (1) in the random X setting. The blue line (KIT) is the result of using only paired data; hence it does not change with addition of unpaired data. KIT only uses the $n = 100$ paired data points. Our methods (green and orange) start with n pairs and gradually adds unpaired data to improve type I error control. False-positive rates for both variants of our method SAT-rx are well controlled around the nominal value (DR: Dimension Reduction)

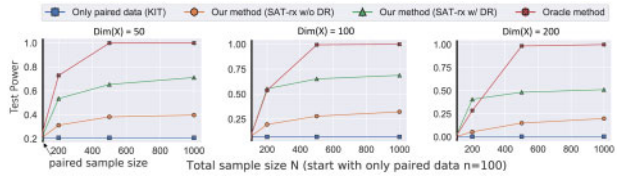

Fig. 3. Evaluation of SAT-rx test power on the simulated data generated by procedure (1) in the random X setting (DR: Dimension Reduction). The results for heritability values $h^2 = 0.1$ and dimensionality $dim(X) = dim(Y) = 50, 100, 200$ are shown. KIT only uses the $n = 100$ paired data points. Our methods start with n pairs and gradually add unpaired data to improve test power

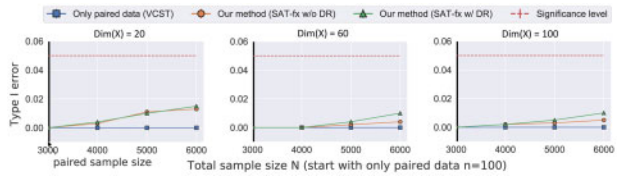

Fig. 4. Evaluation of SAT-fx type I error rate control on the data generated in simulation (2). VCST only uses the $n = 3000$ paired data points. Our method SAT-fx starts with n pairs and gradually adds unpaired data to improve type I error control

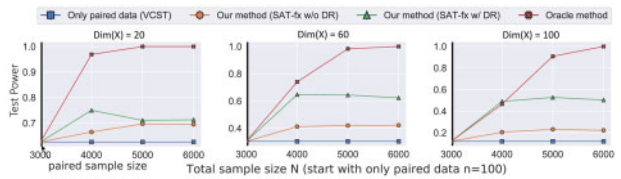

Fig. 5. Evaluation of SAT-fx test power on the data generated in simulation (2). VCST only uses the $n = 3000$ paired data points. Our method SAT-fx starts with n pairs and gradually adds unpaired data to improve test power

method's test power increases when adding unpaired data. Though our method has lower power than the oracle method which has access to all the paired data, it consistently outperforms the baseline KIT method that uses only paired data.

Figures 4 and 5 report the type I error rates and powers of all the methods evaluated in simulation (2). Again, we can see from Figure 4 that the type I error rates of our proposed methods approach the significance level (0.05) as we add more unpaired data. However, because the dimensionality of the genotype is very high, the test is still very conservative even after adding unpaired data. Nevertheless, our method's power exceeds that of VCST and increases as we add unpaired data.

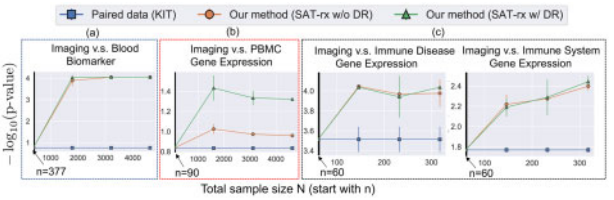

Fig. 6. Experiments on three real imaging and genetics datasets. (**a**) Test an association between multidimensional imaging features and plasma biomarkers. (**b**) Test an association between imaging features and peripheral blood mononuclear cell gene expression data. (**c**) Test an association between imaging features and gene expression of genes in immune system pathway of the disease. In all the experiments, we start with n paired data points and show the behavior of our methods when adding unpaired data, with and without dimensionality reduction (DR)

4 Real data application

4.1 COPD: imaging data and peripheral blood biomarkers

In this experiment, we investigate whether the high-dimensional radiographical measurement from Computed Tomography (CT) imaging is associated with peripheral blood biomarker signature of emphysema. COPD is a highly heterogeneous disease and involves many subprocesses, including emphysema (Vestbo et al., 2013). CT imaging is increasingly used for emphysema diagnosis because it directly characterizes anatomical variation introduced by the disease (Schroeder *et al.*, 2013). Currently, Low Attenuation Area (LAA) is used to quantify the emphysema (Rames and Jones, 2011; Sakai *et al.*, 1994). However, LAA is based on a single intensity threshold value and cannot characterize variation in the texture of the lung parenchyma due to different disease subtypes (Satoh *et al.*, 2001). Over the past year, researchers have proposed various generic and specific local image descriptors that extract higher order statistical features from CT images (Mendoza *et al.*, 2012; Schabdach *et al.*, 2017; Sorensen *et al.*, 2012). However, it is not clear whether such high-dimensional measurements are considered phenotypes, and whether the relationship to the causal biological processes is maintained.

We test the association between one of these multidimensional phenotypes and peripheral blood biomarkers. For the phenotypes, we use the method proposed by Schabdach *et al.* (2017) that computes the similarity between 4629 patients and associates a 100-dimensional vector to each patient (see Supplementary Section S6 for details). For the blood biomarkers, we use the 114 candidate biomarkers in Carolan *et al.* (2014), which were selected based on the pilot work from the BIOSPIR group (O'Neal *et al.*, 2014). A full list of biomarkers analyzed in the COPDGene cohort is available in Supplementary Table S3 of Carolan *et al.* (2014). Because biomarkers are only measured for a subset of subjects, only 377 patients have both the blood biomarker and imaging data. We correct for the effects of covariates including age, sex, BMI (body mass index) and pack-year smoking history. Figure 6a reports the $-\log_{10}(P - \text{value})$ of different methods with respect to size of the unpaired imaging data. The results show that our method takes advantage of unpaired data and detects an association between high-dimensional imaging phenotypes and blood biomarkers that was not detected by the baseline method using only paired data.

4.2 COPD: imaging data and peripheral blood genes

Although smoking is a major risk factor for COPD, not all smokers develop debilitating disease, which suggests that COPD is a systemic disease and other factors might be involved in its development. Bahr *et al.* (2013) identified a set of genes whose expression is associated with two measurements used to diagnose COPD: percent predicted Forced Expiratory Volume in one second (FEV1) and the ratio of FEV1 to forced vital capacity (FEV1/FVC). These genes in Peripheral Blood Mononuclear Cells (PBMC) play a role in the immune system, inflammatory responses and sphingolipid metabolism. Similar to the previous experiment, we investigate whether the

Table 2. *P*-values on Uganda General Population Cohort

	h^2	KIT		SAT-rx (w/o DR)		SAT-rx		Oracle	
		P-value	*P*-value (Bonf)	*P*-value	*P*-value (Bonf)	*P*-value	*P*-value (Bonf)	*P*-value	*P*-value (Bonf)
SBP	0.22	0.293	1.000	0.224	1.000	0.128	0.897	0.010	0.195
DBP	0.29	0.091	0.928	**0.031**	0.537	**7.25e-03**	0.138	<1.00e-05	<1.90e-04
BMI	0.37	0.101	0.907	**0.035**	0.528	**0.011**	0.214	<1.00e-05	<1.90e-04
WHR	0.14	0.249	1.000	0.171	0.901	0.119	0.810	0.033	0.630
Weight	0.43	0.057	0.819	**0.012**	0.235	**1.63e-03**	**0.031**	<1.00e-05	<1.90e-04
Height	0.50	0.031	0.532	3.81e-03	0.072	1.74e-04	**3.31e-03**	<1.00e-05	<1.90e-04
HC	0.37	0.095	0.930	**0.031**	0.503	**0.010**	0.196	<1.00e-05	<1.90e-04
WC	0.31	0.127	0.928	0.057	0.662	**0.022**	0.345	1.20e-05	2.28e-04
ALT	0.37	0.204	0.920	0.172	0.646	0.106	0.617	1.76e-03	0.033
Albumin	0.44	0.117	0.983	**0.046**	0.593	**0.024**	0.395	<1.00e-05	<1.90e-04
ALP	0.12	0.442	1.000	0.419	1.000	0.318	1.000	0.261	1.000
AST	0.25	0.293	1.000	0.322	1.000	0.276	0.875	0.187	1.000
Bilirubin	0.45	0.046	0.629	0.027	0.390	8.43e-03	0.160	<1.00e-05	<1.90e-04
Cholesterol	0.60	0.024	0.448	2.25e-03	**0.043**	1.96e-04	**3.72e-03**	<1.00e-05	<1.90e-04
GGT	0.11	0.307	1.000	0.290	0.801	0.265	0.800	0.039	0.734
HDL	0.51	0.063	0.717	**0.017**	0.326	**4.76e-03**	0.090	<1.00e-05	<1.90e-04
LDL	0.60	0.012	0.222	6.10e-04	**0.012**	2.20e-05	**4.18e-04**	<1.00e-05	<1.90e-04
Triglycerides	0.27	0.242	1.000	0.164	1.000	0.126	0.880	6.76e-04	0.013
HbA1c2	0.56	6.23e-03	0.118	3.66e-04	**6.95e-03**	1.80e-05	**3.42e-04**	<1.00e-05	<1.90e-04
WBC	0.44	6.95e-03	0.139	<1.00e-05	**<2.00e-04**	<1.00e-05	**<2.00e-04**		
RBC	0.39	0.011	0.219	4.40e-05	**8.80e-04**	≪1.00e-05	**<2.00e-04**		
Hemoglobin	0.20	0.041	0.815	1.18e-03	**2.36e-02**	1.40e-05	**2.80e-04**		
HCT	0.22	0.025	0.508	3.36e-04	**6.72e-03**	<1.00e-05	**<2.00e-04**		
MCV	0.57	1.47e-03	0.029	<1.00e-05	<2.00e-04	<1.00e-05	<2.00e-04		
MCH	0.53	2.50e-03	0.050	<1.00e-05	<2.00e-04	<1.00e-05	<2.00e-04		
MCHC	0.72	<1.00e-05	<2.00e-04	<1.00e-05	<2.00e-04	<1.00e-05	<2.00e-04		
RDW	0.33	6.70e-03	0.134	<1.00e-05	**<2.00e-04**	<1.00e-05	**<2.00e-04**		
PLT	0.48	3.00e-03	0.060	<1.00e-05	**<2.00e-04**	<1.00e-05	**<2.00e-04**		
MPV	0.57	1.00e-05	2.00e-04	<1.00e-05	<2.00e-04	<1.00e-05	<2.00e-04		
NEUPr	0.39	0.015	0.304	7.80e-05	**1.56e-03**	<1.00e-05	**<2.00e-04**		
LYMPHPr	0.47	3.30e-03	0.066	<1.00e-05	**<2.00e-04**	<1.00e-05	**<2.00e-04**		
MONOPr	0.48	7.48e-03	0.150	<1.00e-05	**<2.00e-04**	<1.00e-05	**<2.00e-04**		
EOSPr	0.41	1.13e-01	1.000	**0.017**	0.331	**7.08e-04**	**0.014**		
BASOPr	0.47	9.60e-04	0.019	<1.00e-05	<2.00e-04	<1.00e-05	<2.00e-04		
LYMPH	0.52	4.10e-04	0.008	<1.00e-05	<2.00e-04	<1.00e-05	<2.00e-04		
NEU	0.35	0.062	1.000	**3.31e-03**	0.066	**6.00e-05**	**1.20e-03**		
MONO	0.43	0.012	0.236	2.40e-05	**4.80e-04**	<1.00e-05	**<2.00e-04**		
EOS	0.39	0.212	1.000	0.080	1.000	**6.78e-03**	0.136		
BASO	0.50	1.20e-05	2.40e-04	<1.00e-05	<2.00e-04	<1.00e-05	<2.00e-04		

Note: The newly found associations by our method at the significance level 0.05 are marked as bold. Since we mimic the missingness for phenotypes in the top part of the table, we are able to compare our performance with the oracle. In the bottom part of the table, a subset of the subjects has a missing phenotype; hence, the oracle columns are empty.

multidimensional imaging phenotype is associated with the genes identified in (Bahr *et al.*, 2013). We use the same set of imaging phenotypes as done in the previous experiment. In this dataset, 90 subjects have both phenotype and gene expression measurements while more than 4539 subjects only have imaging phenotypes. We use the same covariates as the previous experiment. Figure 6b shows that our method exploits the unpaired data and results in lower *P*-values, suggesting an association between the imaging phenotypes and PBMC gene expression (*P*-value < 0.05) while the *P*-values of the baseline method using only paired data fails to pass the significance level.

4.3 COPD: imaging data and immune system gene expression

In this experiment, we apply our method again in the context of COPD but on a different dataset (Kim *et al.*, 2015). We investigate the hypothesis that anatomical changes manifested on images are related to auto immune pathways. More specifically, we chose the 'immune disease' and 'immune system' gene pathways in the KEGG database (Kanehisa and Goto, 2000). We apply our method to imaging phenotypes and gene expression data containing 319 subjects from several sources (gene expression data from the GEO repository, imaging and clinical information from the Lung Genomics Research Consortium) (Kim *et al.*, 2015). The details of imaging phenotypes are given in Supplementary Table S1. Because only 60 patients have imaging phenotypes, we have a number of unpaired gene expression data. We compare our method with the baseline method that does not use the unpaired gene data and the results are shown in Figure 6c. We can see that our method finds more significant associations as we add more unpaired data.

4.4 Heritable phenotype discovery

In this section, we use the General Population Cohort (GPC), Uganda (Asiki *et al.*, 2013), to establish genotype-phenotype

associations in Genome-Wide Association Studies (GWAS), and show that our method can benefit from unpaired data.

GWAS have discovered many genetic risk variants of common diseases (Ehret *et al.*, 2011; Gratten *et al.*, 2014). Before performing GWAS, one should test the hypothesis that a given phenotype is 'heritable' or not. Given the observation of a phenotype in a population of subjects, so-called narrow sense heritability is defined as an additive genetic portion of the phenotypic variance (Visscher *et al.*, 2006; Yang *et al.*, 2010). A linear mixed model (LMM), which is a form of multivariate regression, is used to estimate the heritability (h^2). Testing for the null hypothesis of $\mathcal{H}_0 : h^2 = 0$ can be done using VCST and the power of the test is affected by the sample size.

We apply our method to study the heritability of a set of phenotypes from the General Population Cohort (GPC), Uganda. More specifically, it contains 37 phenotypes, including anthropometric indices, blood factors, glycemic control, blood pressure, lipid tests and liver function tests (see the complete list of phenotypes in Supplementary Section S8). Initially, 5000 individuals were genotyped using the Illumina HumanOmni 2.5 M BeadChip array, out of which 4778 samples pass the quality control. We follow (Heckerman *et al.*, 2016) exactly for quality control including the Hardy-Weinberg equilibrium (HWE) test, exclusion of Single Nucleotide Polymorphisms (SNPs) with low Minor Allele Frequency (MAF) and computation of the related matrix.

Among all the phenotypes, 18 phenotypes were measured for all the subjects, while the remaining 19 phenotypes were recorded for only 1423 subjects. Thus we conduct two sets of experiments for these two sets of phenotypes. For the 18 phenotypes measured for all individuals, we conduct experiments to mimic the random missingness of phenotypes. We subsample 3000 individuals as unpaired data and allocate the rest as paired data. We compare the *P*-values of the KIT as a baseline with two variants of SAT-rx, with and without dimensionality reduction. In this experiment, we are mimicking the missingness, hence we have access to the oracle, i.e. applying KIT using all data as paired, which is the best we can achieve and which we also compare with our method. For each phenotype, we run the experiments for five times and report the mean of the *P*-values. The standard deviation of the *P*-values is reported in Supplementary Information S9. The upper half of Table 2 reports the *P*-values generated by different methods for all evaluated phenotypes. We can see that the oracle produces much smaller *P*-values in general, while the baseline KIT method can hardly find significant associations. Our SAT-rx method clearly outperforms the KIT method and approaches the performance of the oracle on some phenotypes. Among the 18 phenotypes, our method finds 5 more heritable phenotypes than the baseline method at significance level 0.05.

For the other 19 phenotypes, 1415 individuals have both genotype and phenotype values, and the remaining individuals are considered unpaired (only genotype). For each phenotype, we also run the experiments for five times and report the mean of the *P*-values. The standard deviation of the *P*-values is reported in Supplementary Information S9. We compare the *P*-values of the KIT as a baseline with two variants of SAT-rx, with and without dimensionality reduction. The lower half of Table 2 reports the *P*-values for all methods evaluated on these phenotypes. Among the 19 phenotypes, our method identifies 12 more heritable phenotypes than the baseline method at significance level 0.05.

From the results, we can see that the *P*-values of the 19 unpaired phenotypes are more significant than the 18 paired phenotypes. This suggests that phenotypes with unpaired data may have a stronger correlation with the genotypes. We also provide heritability estimation of all the phenotypes, as shown in Table 2. It can be seen that the heritability values for the 19 unpaired phenotypes are generally larger than those of the paired phenotypes.

5 Discussion

In this article, we have developed SAT, an association test method that can incorporate unpaired data to improve the test power. Unpaired data is unavoidable because existing datasets often require efforts from multiple sites and the data collection process is not necessarily perfectly synchronized. Our method makes better use of current datasets, providing greater test power and thus enabling new discoveries from limited data. First, we have used simulations to show that SAT better controls type I error and has improved power compared to classical methods that only use paired data. Second, in an analysis of 18 phenotypes in the General Population Cohort for which all individuals have paired data, we have found that SAT, using a fraction of the data as paired data and the rest as unpaired data by mimic missingness, produces *P*-values that are closer to the *P*-values generated by the oracle method, which has access to all the paired data, than does the classical KIT method. In the analysis of the other 19 phenotypes in the GPC, for which only 1/4 of the subjects have phenotype data, our SAT method discovers new genetic associations that cannot be discovered by KIT, which ignores the unpaired data. Finally, we applied our method to three real imaging-genetics tasks in which not all subjects have paired data, and the results demonstrate that *P*-values from our method pass the significance level while previous methods fail to find significant associations. All of the results suggest that our method has much better chance to discover new associations when given limited paired data and a reasonable amount of unpaired data. Although we demonstrate the capability of SAT in imaging-genetics tasks, it can be potentially applied to discover relations between many types of multidimensional data when unpaired data exist. It is worth noting that our method relies on the assumption that the data missing mechanism is independent to the association relationship. If this assumption is violated, for example we are only given biased paired data, our method cannot recover the original association.

Financial Support: NIH Award Number 1R01HL141813-01 NSF 1839332 Tripod+X SAP SE.

Conflict of Interest: none declared.

Data availability

The imaging and genetic data for the COPD genomic data are accessible via dbGap (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000179.v1.p1). Further, preprocessed data can be requested from the COPDGene consortium by submitting an ancillary study to SteppL@njhealth.org. Pre-processing of the imaging data follows steps explained in (Schabdach et al. 2017). The GPC genomic data are available at the European Genome-phenome Archive (EGA) under accession number EGAS00001001558.

References

Altshuler,D. *et al.* (2008) Genetic mapping in human disease. *Science*, **322**, 881–888.

Amir,E-a. D. *et al.* (2013) viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nat. Biotechnol.*, **31**, 545–552.

Asiki,G. *et al.* (2013) The general population cohort in rural south-western Uganda: a platform for communicable and non-communicable disease studies. *Int. J. Epidemiol.*, **42**, 129–141.

Bahr,T.M. *et al.* (2013) Peripheral blood mononuclear cell gene expression in chronic obstructive pulmonary disease. *Am. J. Respir. Cell Mol. Biol.*, **49**, 316–323.

Carolan,B.J. *et al.* (2014) The association of plasma biomarkers with computed tomography-assessed emphysema phenotypes. *Respir. Res.*, **15**, 127.

Csernansky,J. *et al.* (2005) Preclinical detection of Alzheimer's disease: hippocampal shape and volume predict dementia onset in the elderly. *Neuroimage*, **25**, 783–792.

Csernansky,J.G. *et al.* (1998) Hippocampal morphometry in schizophrenia by high dimensional brain mapping. *Proc. Natl. Acad. Sci. USA*, **95**, 11406–11411.

Ehret,G.B. *et al.*; CHARGE-HF Consortium. (2011) Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature*, **478**, 103–109.

Ge,T. *et al.* (2015) Massively expedited genome-wide heritability analysis (MEGHA). *Proc. Natl. Acad. Sci. USA*, **112**, 2479–2484.
Ge,T. *et al.* (2016) Multidimensional heritability analysis of neuroanatomical shape. *Nat. Commun.*, **7**, 13291.
Gerber,S. *et al.* (2010) Manifold modeling for brain population analysis. *Med. Image Anal.*, **14**, 643–653.
Gratten,J. *et al.* (2014) Large-scale genomics unveils the genetic architecture of psychiatric disorders. *Nat. Neurosci.*, **17**, 782–790.
Gretton,A. *et al.* (2008a) A kernel statistical test of independence. In: *NIPS08*. MIT Press, Cambridge, MA, pp. 585–592.
Gretton,A. *et al.* (2008b) A kernel statistical test of independence. In: *NIPS 20*. MIT Press, Cambridge, MA, pp. 585–592.
Haghverdi,L. *et al.* (2015) Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics*, **31**, 2989–2998.
Heckerman,D. *et al.* (2016) Linear mixed model for heritability estimation that explicitly addresses environmental variation. *Proc. Natl. Acad. Sci. USA*, **113**, 7377–7382.
Hua,W.-Y. and Ghosh,D. (2015) Equivalence of kernel machine regression and kernel distance covariance for multidimensional phenotype association studies. *Biometrics*, **71**, 812–820.
Hwang,J. *et al.* (2006) Basal ganglia shape alterations in bipolar disorder. *Am. J. Psychiatry*, **163**, 276–285.
Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.
Kim,S. *et al.* (2015) Integrative phenotyping framework (IPF): integrative clustering of multiple omics data identifies novel lung disease subphenotypes. *BMC Genomics*, **16**, 924.
Kwee,L.C. *et al.* (2008) A powerful and flexible multilocus association test for quantitative traits. *Am. J. Hum. Genet.*, **82**, 386–397.
Liu,D. *et al.* (2007) Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. *Biometrics*, **63**, 1079–1088.
Maity,A. *et al.* (2012) Multivariate phenotype association analysis by marker-set kernel machine regression. *Genet. Epidemiol.*, **36**, 686–695.
Mendoza,C.S. *et al.* (2012) Emphysema quantification in a multi-scanner HRCT cohort using local intensity distributions. In: *2012 9th IEEE International Symposium on Biomedical Imaging (ISBI)*, Barcelona, Spain. IEEE (New York), pages 474–477.
O'Neal,W.K. *et al.*; Reporting for SPIROMICS Investigators. (2014) Comparison of serum, EDTA plasma and P100 plasma for luminex-based biomarker multiplex assays in patients with chronic obstructive pulmonary disease in the SPIROMICS study. *J. Transl. Med.*, **12**, 9.
Qiu,P. *et al.* (2011) Extracting a cellular hierarchy from high-dimensional cytometry data with spade. *Nat. Biotechnol.*, **29**, 886–891.
Rames,A. and Jones,P. (2011) TESRA (treatment of emphysema with a selective retinoid agonist) study results. *Am. J. Respir. Crit. Care Med.*, **183**, A6418.
Regan,E.A. *et al.* (2011) Genetic epidemiology of COPD (COPDgene) study design. *COPD J. Chronic Obstr. Pulm. Dis.*, **7**, 32–43.
Sakai,N. *et al.* (1994) An automated method to assess the distribution of low attenuation areas on chest CT scans in chronic pulmonary emphysema patients. *Chest*, **106**, 1319–1325.
Satoh,K. *et al.* (2001) CT assessment of subtypes of pulmonary emphysema in smokers. *Chest*, **120**, 725–729.
Schabdach,J. *et al.* (2017) A likelihood-free approach for characterizing heterogeneous diseases in large-scale studies. In: *International Conference on Information Processing in Medical Imaging*, North Carolina, USA. Springer, New York, pp. 170–183.
Schaid,D.J. (2010a) Genomic similarity and kernel methods I: advancements by building on mathematical and statistical foundations. *Hum. Hered.*, **70**, 109–131.
Schaid,D.J. (2010b) Genomic similarity and kernel methods II: methods for genomic information. *Hum. Hered.*, **70**, 132–140.
Schroeder,J.D. *et al.* (2013) Relationships between airflow obstruction and quantitative CT measurements of emphysema, air trapping, and airways in subjects with and without chronic obstructive pulmonary disease. *Am. J. Roentgenol.*, **201**, W460–W470.
Sejdinovic,D. *et al.* (2013) Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Ann. Stat.*, **41**, 2263–2291.
Shifman,S. *et al.* (2003) Linkage disequilibrium patterns of the human genome across populations. *Hum. Mol. Genet.*, **12**, 771–776.
Sorensen,L. *et al.* (2012) Texture-based analysis of COPD: a data-driven approach. *IEEE Trans. Med. Imaging*, **31**, 70–78.
Sriperumbudur,B. *et al.* (2011) Universality, characteristic kernels and RKHS embedding of measures. *J. Mach. Learn. Res.*, **12**, 2389–2410.
Székely,G.J. *et al.* (2007) Measuring and testing dependence by correlation of distances. *Ann. Stat.*, **35**, 2769–2794.
Vestbo,J. *et al.* (2013) Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: gold executive summary. *Am. J. Respir. Crit. Care Med.*, **187**, 347–365.
Visscher,P.M. *et al.* (2006) Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS Genet.*, **2**, e41.
Visscher,P.M. *et al.* (2017) 10 years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.*, **101**, 5–22.
Wei,C. and Lu,Q. (2017) A generalized association test based on U statistics. *Bioinformatics*, **33**, 1963–1971.
Willer,C.J. *et al.* (2013) Discovery and refinement of loci associated with lipid levels. *Nat. Genet.*, **45**, 1274.
Yang,J. *et al.* (2010) Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.*, **42**, 565–569.
Yang,J. *et al.* (2011) GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.*, **88**, 76–82.
Zhang,K. *et al.* (2011) Kernel-based conditional independence test and application in causal discovery. In: *27th Conference on Uncertainty in Artificial Intelligence (UAI 2011)*. AUAI Press, pp. 804–813.
Zhou,J.J. *et al.* (2013) Heritability of chronic obstructive pulmonary disease and related phenotypes in smokers. *Am. J. Respir. Crit. Care Med.*, **188**, 941–947.