Bioinformatics, 2018, 1–9
doi: 10.1093/bioinformatics/bty804
Advance Access Publication Date: 15 September 2018
Original Paper

Data and text mining

Integrating hypertension phenotype and genotype with hybrid non-negative matrix factorization

Yuan Luo[1,*], Chengsheng Mao[1], Yiben Yang[1], Fei Wang[2], Faraz S. Ahmad[1], Donna Arnett[3], Marguerite R. Irvin[4] and Sanjiv J. Shah[1]

[1]Department of Preventive Medicine, Feinberg School of Medicine, Northwestern University, Chicago, IL 60611, USA, [2]Department of Healthcare Policy & Research, Weill Cornell Medicine, Cornell University New York, NY 10065, USA, [3]Department of Epidemiology, College of Public Health, University of Kentucky, Lexington, KY 40536, USA and [4]Department of Epidemiology, University of Alabama at Birmingham, Birmingham, AL 35294, USA

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on May 8, 2018; revised on August 20, 2018; editorial decision on September 11, 2018; accepted on September 13, 2018

Abstract

Motivation: Hypertension is a heterogeneous syndrome in need of improved subtyping using phenotypic and genetic measurements with the goal of identifying subtypes of patients who share similar pathophysiologic mechanisms and may respond more uniformly to targeted treatments. Existing machine learning approaches often face challenges in integrating phenotype and genotype information and presenting to clinicians an interpretable model. We aim to provide informed patient stratification based on phenotype and genotype features.

Results: In this article, we present a hybrid non-negative matrix factorization (HNMF) method to integrate phenotype and genotype information for patient stratification. HNMF simultaneously approximates the phenotypic and genetic feature matrices using different appropriate loss functions, and generates patient subtypes, phenotypic groups and genetic groups. Unlike previous methods, HNMF approximates phenotypic matrix under Frobenius loss, and genetic matrix under Kullback-Leibler (KL) loss. We propose an alternating projected gradient method to solve the approximation problem. Simulation shows HNMF converges fast and accurately to the true factor matrices. On a real-world clinical dataset, we used the patient factor matrix as features and examined the association of these features with indices of cardiac mechanics. We compared HNMF with six different models using phenotype or genotype features alone, with or without NMF, or using joint NMF with only one type of loss We also compared HNMF with 3 recently published methods for integrative clustering analysis, including iClusterBayes, Bayesian joint analysis and JIVE. HNMF significantly outperforms all comparison models. HNMF also reveals intuitive phenotype–genotype interactions that characterize cardiac abnormalities.

Availability and implementation: Our code will be made publicly available on github upon publication.

Contact: yuan.luo@northwestern.edu

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

1 Introduction

Precision medicine aims to utilize information from multiple modalities—including phenotypic and genetic measurements—to develop an individualized and comprehensive view of a patient's pathophysiologic progression, to identify unique disease subtypes, and to administer personalized therapies (Kohane, 2015). Existing efforts are often based on only a selected set of biomarkers. The rapid growth of phenotypic and genetic data for many common diseases poses technical challenges for subtyping them, due to the high dimensionality of data, diversity of data types, uncertainty and missing data. However, the rapid growth of multiple data modalities, when linked to the right patients, may provide a prismatic view of the underlying pathophysiology of these diseases and offers a basis for meaningful subtyping of these patients.

Hypertension is an example of a complex, heterogeneous clinical syndrome characterized by elevated blood pressure. Although typically considered a single disease, primary hypertension (i.e., essential hypertension) is in fact a heterogeneous group of subtypes with varying etiologies and pathophysiology. This common form of hypertension is highly prevalent and is polygenic in nature. However, genetic studies of hypertension have focused primarily on analyzing single variants at a time and then ranking them in terms of significance, as has been done in several genome-wide association studies [see Poulter *et al.* (2015) for a review]. However, it is more likely that genetic variants interact with each other to increase susceptibility to disease. Furthermore, genetic variants interact with phenotypic risk factors to further promote the development of diseases such as hypertension. With the growing availability of high throughput genotyping and phenotyping data (such as through NIH/NHLBI TOPMed program), the need for integrating both data modalities is becoming increasingly pressing. Thus, it is critical to develop a methodology to combine phenotypic and genetic data when clustering patients for the identification of novel subtypes of the disease. Such work could help identify novel molecular and pathophysiological pathways of disease and also may identify subgroups of patients who are more homogeneous in their response to specific therapies.

Major contributions of this paper are: (i) Aiming to provide informed patient stratification, we propose Hybrid Non-negative Matrix Factorization (HNMF) that approximates phenotype and genotype matrices using different appropriate loss functions, instead of single loss function in previous joint NMF methods. (ii) We use simulation to show HNMF converges fast and accurately to true factor matrices, and we use a real-world clinical dataset to show HNMF-generated patient factor matrix is more effective in predicting indices of cardiac mechanics compared to multiple non-NMF, NMF and joint NMF based methods. (iii) We show that HNMF-generated group matrices lead to insights on phenotype–genotype interactions that characterize cardiac abnormalities.

From the clinical perspective, there have been only a few previous studies that have examined the clustering of hypertensive patients. Katz *et al.* applied model-based clustering to a cohort of 1273 hypertensive individuals, using only phenotypic data as features (Katz *et al.*, 2017). Study participants were clustered into two distinct groups that differed markedly in clinical characteristics, cardiac structure/function, and indices of cardiac mechanics. Guo *et al.* (2017) used K-means clustering of phenotypic data (clinical and blood pressure characteristics) and found four groups of interest. However, neither of these studies utilized genetic data, which could have provided an additional important dimension to the clustering of hypertension, particularly when combined with phenotypic data.

From the method perspective, non-negative matrix factorization (NMF) refers to the set of problems on approximating a non-negative matrix as the product of several non-negative matrices. The problem has become popular since Lee and Seung's Nature paper (Lee and Seung, 1999), where they form a nonnegative matrix by concatenating the set of pixel intensity vectors stretched from human facial images. After factorizing such matrix into the product of two matrices, they found that one matrix can be interpreted as the set of image basis with part based representation of human faces, and the other matrix is the coefficients if we were to reconstruct the face image from those bases. Because of the non-negativity constraints, NMF is not a convex problem and they developed a multiplicative update algorithm to obtain a stationary solution, with provable convergence of the algorithm (Lee and Seung, 2001).

Since then researchers have been working on NMF from various aspects. Ding *et al.* (2005) showed that there is some equivalence between NMF and Kmeans/spectral clustering and claim NMF can be used for data clustering. Ding *et al.* (2006) further developed a t-NMF approach that can perform co-clustering on both matrix columns and rows. They also discussed the various NMF variants (Ding *et al.*, 2010). Sra and Dhillon (2006) extended NMF to the case when the matrix approximation loss is measured by Bregman divergence, which is a much more general loss with both Frobenius norm and KL divergence, which are discussed in (Lee and Seung, 2001)) as its special cases. On the solution procedure aspect, multiplicative updates have been recognized for its slow convergence and poor quality. Lin (Lin, 2007) proposed a projected gradient approach for NMF. Kim and Park (2011) also proposed an active set type of method called principal block pivoting to solve the NMF problem.

NMF is a highly effective unsupervised method to cluster similar patients (Hofree *et al.*, 2013; Luo *et al.*, 2016b) and sample cell lines (Müller *et al.*, 2008), and to identify subtypes of diseases (Collisson *et al.*, 2011). Conventional NMF can only model either phenotypic measurements (e.g. using Frobenius loss) or genetic variants (e.g., using KL loss) but not both. Recent studies have investigated methods for joint matrix factorization, serving the purpose of meta-analysis (Wang *et al.*, 2015), multi-view clustering (Liu *et al.*, 2013) or imposing multiple characterization of documents (Kim *et al.*, 2015). However, these methods focus on using Frobenius loss to measure approximation accuracy of multiple matrices, and cannot readily integrate phenotypic measurements and genetic variants where approximating the two matrices admit different types of loss functions. Gunasekar *et al.* (2016) proposed collective matrix factorization based on the Bregman divergence framework to integrate multi-source EHR phenotyping data, implemented KL-divergence as matrix approximation loss and experimented on discrete diagnosis and medications data.

In theory, both KL divergence and Frobenius loss are special cases of Bregman divergence, but care needs to be taken when materializing the theoretical framework to the concrete case of hybrid genotypic and continuous phenotypic data. Challenges include how to derive useful genetic variant information from terabytes of whole exome sequencing data, how to filter deleterious variants, how to properly implement HNMF with missing continuous data, etc. Our paper is one such concrete materialization to integrate phenotypic and genotypic information for patient subtyping. Addressing both the clinical and methodological challenges, we propose the model of HNMF which models the approximations of phenotypic and genetic matrices under Frobenius loss and KL loss respectively. We develop an alternating project gradient descent method for optimizing the HNMF objective, and demonstrate its fast convergence and

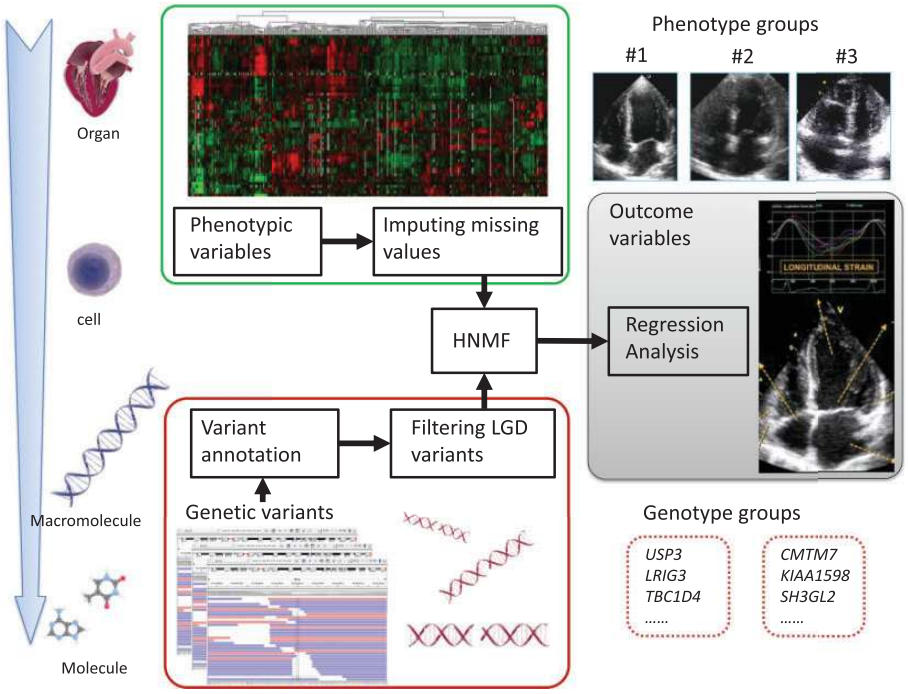

Fig. 1. Study workflow. HNMF stands for hybrid non-negative matrix factorization. LGD stands for likely gene disruptive

effectiveness in integrating both the phenotypic and genetic data using both simulated and real-world studies.

2 Materials and methods

We develop a hybrid matrix factorization method to integrate both phenotypic and genetic features. The model applies non-negative matrix factorization to discover groups of phenotypic variables and genetic variants simultaneously that collectively and interactively characterize the groups of the patients. The approximation error is measured using Frobenius loss for the phenotypic matrix, and KL loss for the genetic matrix; hence we name our algorithm the HNMF. We have made the following assumptions throughout the paper:

1. The phenotype matrix and the genotype matrix share a common set of subtypes;
2. The relationship between the variables and the groups are linear and can be modeled with matrix multiplication.

2.1 Workflow of the study

We first outline the workflow of the study in Figure 1. This study considers two types of patient data: phenotypic measurements and genetic variants.

We first impute missing values in the phenotypic variables. For genetic variants, we first annotate the variants and then keep those variants that are likely gene disruptive (LGD). The pre-processed phenotypic measurements and genetic variants are then used as input to our HNMF algorithm. The patient factor matrix is then used as the feature matrix to perform regression analysis to predict main cardiac mechanistic outcomes. We next explain each step in detail.

2.2 Cohort construction and data collection

Our cohort comes from the Hypertension Genetic Epidemiology Network (HyperGEN) study. HyperGEN, part of the NIH Family Blood Pressure Program, is a cross-sectional study consisting of individuals with hypertension, their siblings and offspring, and a random sample of normotensives, all recruited from 4 cities in the United States (Williams *et al.*, 2000). We focus on the African American participants (660 total), for whom we have both the phenotypic data (e.g., vitals) and whole exome sequencing (WES) data. We used two measurements from the echocardiograms that are main reflectors of systolic (longitudinal strain) and diastolic (septal e' velocity) cardiac mechanics as outcome variables (Table 1) (Mitter *et al.*, 2017; Mor-Avi *et al.*, 2011). As opposed to conventional cardiac function measures such as ejection fraction, indices of cardiac mechanics obtained by speckle-tracking echocardiography are more sensitive measures of intrinsic cardiomyocyte function (Shah *et al.*, 2014). Furthermore, indices of cardiac mechanics are thought to be subclinical measures of myocardial dysfunction that occur during the transition from risk factors (e.g., hypertension, obesity, diabetes, renal disease) to overt heart failure (Selvaraj *et al.*, 2016). WES identifies the variants found in the coding region of genes (exons). In order to accurately and consistently call variants from across all datasets, we adopt the GATK framework (DePristo *et al.*, 2011) for a standardized processing of WES data.

Table 1. Outcome variables reflecting cardiac mechanics

Outcome	Description
Septal e' velocity	Left ventricular early diastolic relaxation velocity, measured at the septal mitral annulus in the apical 4-chamber view. Lower values reflect slower left ventricular relaxation and worse diastolic function.
Longitudinal strain	Left ventricular longitudinal strain measured in the apical 4-chamber view, a marker of subendocardial longitudinal systolic function. Lower absolute values reflect worse systolic function.

2.3 Imputation on phenotypic variables

Biomedical, epidemiological and clinical data often contain missing values for test results, some due to issues during data acquisition and archiving, but others due to the fact that clinicians do not order certain tests based on patient-specific diagnostic and treatment course. The missing percentage of the phenotypic variables considered in our study ranges from 0% to 37%. We had our cardiologist colleagues pick rather inclusively 129 phenotypic variables (see Supplementary Material) that can characterize the hypertension risk and cardiac physiology of the patients. We are rather tolerant on missing rate in order to retain as many variables as possible. Nevertheless, only 13/129 (10%) of the phenotypic variables had missingness > 10%. Six of these variables with missingness > 10% (including those with missingness > 30%) were phenotypes related to mitral inflow, which characterize diastolic function. Given the importance of diastolic dysfunction (i.e., abnormal cardiac relaxation and/or reduced cardiac compliance) in the setting of hypertension, we chose to retain these variables because of their clinical importance. We use the Multivariate Imputation by Chained Equations (MICE) algorithm to perform the imputation. This approach assumes a conditional model for each variable to be imputed, with the other variables as possible predictors (van Buuren and Groothuis-Oudshoorn, 2011). The term chained equation comes from the adoption of a Gibbs sampler, which is an iterative Markov Chain Monte Carlo (MCMC) algorithm. Previous studies e.g., Luo *et al.* (2016a)) showed that even at the presence of high missing rate (over 50%), MICE imputation may still render clinically useful information to predict patient outcome due to redundant information in phenotypic variables.

2.4 Annotation-based variant filtration and LGD variant detection

We next used the ANNOVAR toolkit (Wang *et al.*, 2010) to comprehensively annotate called variants with a wide array of information, including their hosting gene ;[using several gene models such as RefSeq, UCSC Known Gene, Gencode (Harrow *et al.*, 2012)]; the variant function; its predicted pathogenicity according to PolyPhen2 (Adzhubei *et al.*, 2013), SIFT (Kumar *et al.*, 2009), CADD (Kircher *et al.*, 2014), and other meta predictors; its minor allele frequency among the 1000 Genomes populations and ExAC (Lek *et al.*, 2016); and its phenotype associations according to ClinVar, and HGMD (Stenson *et al.*, 2012).

To address issues of reference mis-annotation, we resort to the recently released Exome Aggregation Consortium (ExAC) exome dataset (Lek *et al.*, 2016), which aims to aggregate exome sequencing data from a wide range of large-scale sequencing projects including the cohorts of Myocardial Infarction Genetics Consortium, Swedish Schizophrenia & Bipolar Studies and The Cancer Genome Atlas (TCGA). We filter out those variants whose allele frequencies are observed to be over 90% among the 60, 706 individuals aggregated by ExAC. We also apply a similar 90% filtering threshold on the alternate allele frequency in our cohort. We further focus on likely gene disruptive (LGD) variants, which include frame-shift insertion, frame-shift deletions, nonsense variants and splice site alterations. We have 6430 gene features for our cohort, 660 subjects. We follow the common practice and exclude the genes that have very rare variants (<10 subjects) or very frequent variants (> 50% of the subjects), resulting in 1481 genes. We then follow the common approach of gene prioritization (Moreau and Tranchevent, 2012) and further select the genes that show significant difference between the two hypertension groups (patient taking 1 vs. multiple antihypertensive medications) by two-tailed binomial exact tests (Howell, 2012). The gene selection is based on the entire patient cohort but uses a categorical label that is different from the final continuous outcomes of cardiac mechanics indexes. Eventually, 349 (110) genes (Supplementary Material) are selected for our cohort with p-value of binomial test less than 0.1 (0.01). Each entry of our genetic matrix specifies how many variants a patient has on that gene.

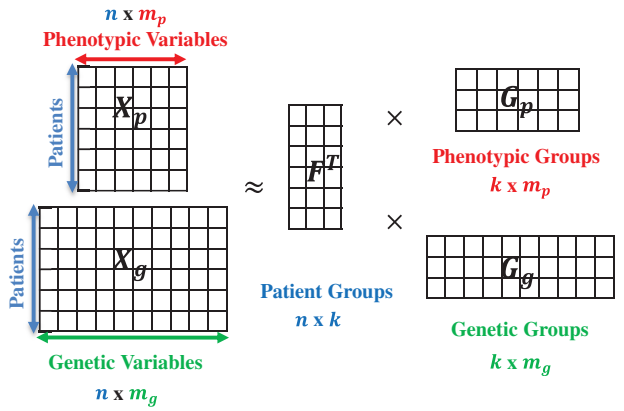

Fig. 2. Hybrid non-negative matrix factorization model. In the figure, X_p is the patient-by-phenotype-measurement matrix. X_g is the patient-by-genetic-variant matrix. F is the patient factor matrix specifying patient groups. G_p is the phenotype factor matrix specifying groups of phenotypic measurements. G_g is the genetic factor matrix specifying groups of genetic measurements

Table 2. Notations

Variable	Description
n	Number of patients
m_p	Number of phenotypes
m_g	Number of genotypes
k	Number of patient groups
$X_p \in R^{n\times m_p}$	Patient by phenotype matrix, continuous value
$X_g \in R^{n\times m_g}$	Patient by genotype matrix, count value
$F \in R^{k\times n}$	Patient group assignment matrix
$G_p \in R^{k\times m_p}$	Phenotype group assignment matrix
$G_g \in R^{k\times m_g}$	Genotype group assignment matrix

2.5 Hybrid NMF

We propose the hybrid NMF (HNMF) model that integrates both phenotypic and genetic measurements of patients. The phenotypic measurements we consider are continuous values, hence we use Gaussian distribution to model the approximation error. The genetic measurements are counts of the genetic variants that happen to a particular gene, thus we use Poisson distribution to model the variant count. A schematic view of our HNMF model is shown in Figure 2.

Our goal is to maximize the joint likelihood of the two approximations. Let the variables be defined as in Table 2, we establish the following constrained optimization problem

$$\max\ \lambda \log\ P(X_g|F,G_g) + \log P(X_p|F,G_p)$$

$$st.\ F \geq 0, G_p \geq 0, G_g \geq 0 \qquad (1)$$

where λ indicates the trade-off between the phenotypic approximation and genetic approximation ($\lambda = 1$ for our experiment), and the log likelihood functions are defined as follows.

$$\log P(X_g|F,G_g) = -\frac{1}{2\delta^2}\sum_{ij}\left(X_{p_{ij}} - \sum_u F_{ui}G_{p_{uj}}\right)^2 + C_1 \quad (2)$$

$$\log P(X_g|F,G_g) = \sum_{ij}\left(X_{g_{ij}}\log\left(\sum_u F_{ui}G_{g_{uj}}\right) - \sum_u F_{ui}G_{g_{uj}}\right) + C_2 \quad (3)$$

We require F, G_g and G_p to be nonnegative in order to achieve better interpretability. Since the values of the entries in both X_p and X_g are nonnegative, with the nonnegativity constraints on F, G_g and G_p we are essentially assuming an additive reconstruction of X_p and X_g from the product of those factors under a hybrid loss. According to the seminal paper from Lee and Seung in Nature (Lee and Seung, 1999), such additive reconstruction can result in better interpretation of F, G_g and G_p.

By minimizing the negative log likelihood, we arrive at the following objective function.

$$\min\mathcal{L}(F,G_p,G_g) = \sum_{ij}\left[\frac{\lambda}{2}\left(X_{p_{ij}} - \hat{X}_{p_{ij}}\right)^2 + \hat{X}_{g_{ij}} - X_{g_{ij}}\log\left(\hat{X}_{g_{ij}}\right)\right]$$

$$st.\ F \geq 0, G_p \geq 0, G_g \geq 0 \quad (4)$$

where $\hat{X}_{p_{ij}} = \sum_u F_{ui}G_{p_{uj}}$ and $\hat{X}_{g_{ij}} = \sum_u F_{ui}G_{g_{uj}}$ Writing $\mathcal{L}$ in the matrix form, we have

$$\mathcal{L}(F,G_p,G_g) = \sum_{ij}\left[\frac{\lambda}{2}||X_p - \hat{X}_p||_F^2 + \hat{X}_g - X_g\log\left(\hat{X}_g\right)\right] \quad (5)$$

where $\hat{X}_p = F^T G_p$ and $\hat{X}_g = F^T G_g$. We can use the following alternating projected gradient descent procedure to solve the objective and establish the stopping criteria that the partial gradients should be small enough or all factor matrix updates cannot produce a feasible direction along which the objective function decreases (let $\mathcal{P}_+(\cdot)$ denote the non-negative projector):

$$F^{t+1} = \mathcal{P}_+\left[F^t - \eta_F^t\nabla_F\mathcal{L}\left(F,G_p^t,G_g^t\right)|_{F=F^t}\right] \quad (6)$$

$$G_p^{t+1} = \mathcal{P}_+\left[G_p^t - \eta_{G_p}^t\nabla_{G_p}\mathcal{L}\left(F^{t+1},G_p,G_g^t\right)|_{G_p=G_p^t}\right] \quad (7)$$

$$G_g^{t+1} = \mathcal{P}_+\left[G_g^t - \eta_{G_g}^t\nabla_{G_g}\mathcal{L}\left(F^{t+1},G_p^{t+1},G_g\right)|_{G_g=G_g^t}\right] \quad (8)$$

These equations take turns in optimizing each factor matrix while keeping the other two fixed. We next present the partial gradients with respect to each of the three factor matrices. For phenotype group matrix G_p, we have

$$\nabla_{G_p}\mathcal{L}(F,G_p,G_g) = \lambda\left(FF^T G_p - FX_p\right) \quad (9)$$

Let $\hat{X}_g = F^T G_g$, and $\tilde{X}_{g_{ij}} = X_{g_{ij}}/\hat{X}_{g_{ij}}$, for genotype group matrix G_g, we have

$$\nabla_{G_g}\mathcal{L}(F,G_p,G_g) = F(E_G - \tilde{X}_g) \quad (10)$$

where $E_G \in R^{n\times m_g}$ is an all-one matrix. For the patient group matrix F, we have

$$\nabla_F\mathcal{L}(F,G_P,G_g) = \lambda\left(-G_pX_p^T + G_pG_p^TF\right) + G_g\left(E_F - \sim X_g^T\right) \quad (11)$$

With those gradients, we can adopt an alternating projected gradient descent procedure to solve the hybrid matrix factorization problem. This is an iterative procedure, at each iteration, the algorithm optimizes the objective with one specific group of variables with all other variables fixed. The optimization procedure used at each iteration will be projected gradient descent. In order to determine the step size at each gradient descent step, we use the Armijo rule as a sub-procedure which looks for the largest η (step size) that satisfies the following sufficient decrease condition. Let Θ, Θ^{new} denote the parameters (e.g., F, G_g and G_p) before and after each iteration respectively, and $\delta \in (0,\ 1)$ be a predefined number. General sufficient decrease condition can be written as

$$\ell 5(\Theta^{new}) - \mathcal{L}(\Theta) \leq \delta tr\left(\nabla_\Theta\mathcal{L}(\Theta)(\Theta^{new} - \Theta)^T\right) \quad (12)$$

If $\mathcal{L}$ is a quadratic form of Θ, we have a special fast-to-check sufficient decrease condition as Formula (13) (Lin, 2007). The algorithm for projected gradient descent with Armijo rule can be outlined as Algorithm 1. Note that ρ in the algorithm is a step size controlling parameter that is set to the common choice of 0.1 (Lin, 2007).

$$(1-\delta)tr\left(\nabla_\Theta\mathcal{L}(\Theta)(\Theta^{new} - \Theta)^T\right) + \frac{1}{2}tr\left((\Theta^{new} - \Theta)\nabla_\Theta^2\mathcal{L}(\Theta)(\Theta^{new} - \Theta)^T\right) \leq 0 \quad (13)$$

2.6 Feature group discovery using HNMF

In HNMF, the row vectors in the phenotype factor matrix G_p and in the genetic factor matrix G_g specify the grouping of phenotypic measurements and genetic variants respectively. Such groupings can be viewed as mixtures of phenotypic (or genetic) features, as they allow sharing of these features among different groups as specified by its fractional weights across groups. The motivation is to identify paired phenotypic group and genetic group that together characterize pathophysiologic underpinnings. The approximated phenotypic matrix can be viewed as rank-one sum of outer-product of patient group (e.g., $[F^T]_{\cdot j}$, jth column of the patient group matrix) and phenotypic group (e.g., $[G_p]_{j\cdot}$, jth row of the phenotypic group matrix). Similar argument holds for genetic group matrix. Thus the patient group (e.g., $[F^T]_{\cdot j}$) bridges the corresponding phenotypic group (e.g., $[G_p]_{j\cdot}$) and genetic group (e.g., $[G_g]_{j\cdot}$). We used the patient group matrix F^T as the instance-feature matrix in Ridge regression and used the numeric values of the cardiac mechanic variables as outcomes (listed in Table 1), and identify a column with maximum coefficient (e.g., $[F^T]_{\cdot j}$). We selected the corresponding phenotypic and genetic groups (e.g., $[G_p]_{j\cdot}$ and $[G_g]_{j\cdot}$), which are paired through the shared patient group (e.g., $[F^T]_{\cdot j}$) and provide interpretation

Algorithm 1 Projected gradient descent with Armijo rule

1: Initialize Θ. Set $\eta = 1$
2: **for** $i = 1$ to k **do**
3: **if** η satisfies Eq. (13) (or (12) if quadratic) **then**
4: Repeatedly increase η as $\eta \leftarrow \eta/\rho$ until either η does not satisfy Eq. (13)) (or (12)) if quadratic) or $\Theta(\eta/\rho) = \Theta$
5: **else**
6: Repeatedly decrease η as $\eta \leftarrow \rho\eta$ until η satisfy Eq. (13) (or (12) if quadratic)
7: **end if**
8: Set $\Theta^{new} = \max(0, \Theta - \eta\nabla_\Theta\mathcal{L}(\Theta))$
9: end for

advantage. Using the trained regression model, we rank the patient groups by their regression coefficients and focus on the top patient groups (and associated phenotypic and genetic groups) that are associated with large effect size.

2.7 Evaluating the groups discovered by HNMF

Because there is no innate way (except for simulation) to determine whether the groupings of phenotypic measurements and genetic variants discovered by HNMF are good or poor, we evaluate their utility as features, abstracted from the base data, in a prediction model. We assume that good features will improve prediction and will give us some insights into which phenotypic and genetic patterns are indicative of patient cardiac mechanic abnormality. We use the phenotypic and genetic data for participants from the hypertension genetic epidemiology network (HyperGEN) study. We take a subset of the African American patients who are hypertensive, and for whom we have both phenotypic and genetic data available at large scale. We predict the numeric values of the cardiac mechanic variables as outcomes (listed in Table 1). For each outcome variable, we randomly split these patients into a 7: 3 train and held-out test dataset, and repeat the random initializations of HNMF and other NMF based comparison models 50 times in order to improve the statistical robustness of the results. We did not require that all the individuals from the same family to be included in either the training or the test set, but not both. This is out of the consideration that we want to minimize the potential bias from family variant patterns during model training. However, we did perform additional experiments requiring all the individuals from the same family to be included in either the training or the test set, but not both, which yielded similar numerical results, please refer to the Supplementary Material for more details.

To evaluate the effectiveness of HNMF in abstracting raw data into more predictive features, we use the patient factor matrix F to train a Ridge regression model. We chose Ridge regression over alternatives such as support vector regression or random forest regression for its capability to generate deterministic weights for individual features. We match the groups in the phenotypic factor matrix and genetic factor matrix according to their row indices, and link them to the corresponding row in the patient factor matrix F. Linear regression then provides a convenient way to directly assess phenotypic and genetic group contribution.

3 Results

In this section, we first evaluate the algorithmic performance using a simulated dataset where the actual factor matrices are known. Then, we evaluate the hybrid matrix factorization performance using the HyperGEN dataset.

3.1 Simulation

We first analyze simulated data where the underlying factor matrices are known. Specifically, we consider a 20×10 X_p matrix and a 20×100 X_g matrix with the true number of factors being 3. That is, they are generated by a 3×20 F matrix, a 3×10 G_p matrix, and a 3×100 G_g matrix. We first sample the F, G_p, and G_g matrices. We then generate the X_p matrix by adding an error term ϵ_p on top of $F^T G_p$ where ϵ_p adopts standard normal distribution. Next we generate the X_g matrix by sampling according to Poisson distribution with the parameter set to $F^T G_g$.

In order to evaluate the similarity between the factorized matrix and its true counterpart, we use the following similarity score:

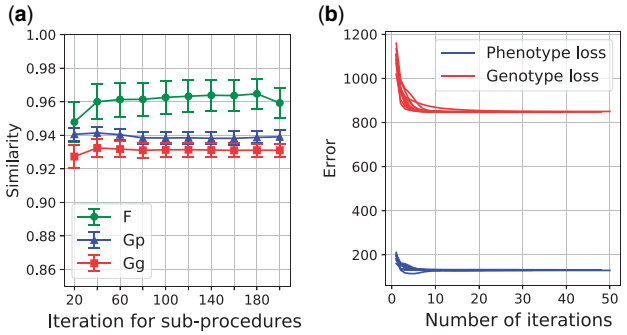

Fig. 3. Simulation results on a hybrid matrix factorization problem with rank 3. Ten random initializations are performed. (**a**) Similarity scores with error bars as a function of number of iterations for sub-procedures (**b**) decreasing trend of loss functions for phenotype (Frobenius loss) and genotype matrix (KL loss) approximations during HNMF, shown are 10 runs. Error bars indicate standard deviations

$$similarity(A, B) = \frac{tr(A^T B)}{\sqrt{tr(A^T A)}\sqrt{tr(B^T B)}} \tag{14}$$

where $tr(\cdot)$ is trace and $tr(A^T B)$ can be considered as matrix inner product. This similarity score is essentially the cosine similarity, which quantifies the closeness between the computed solution and the actual factor matrix and provides a single number between 0 and 1 (Chi and Kolda, 2012). In order to test the sensitivity of estimates to the initialization, we performed random initialization 10 times. The simulation results are shown in Figure 3a where the similarity score is plotted as a function of maximum number of iterations for sub-procedures (optimizing F, G_p, G_g one at a time while fixing the other two, using the Armijo rule), which represents the closeness to the sub-problem optima. Figure 3 shows that as we have extra sub-procedure iterations, the similarity scores first rise slightly and then plateau quickly. We can also see that the similarity between the true factor matrices and those recovered by HNMF quickly reaches to an accurate level (>0.9). Figure 3b shows the convergence speed of the proposed alternating projected gradient descent method with the number of iterations for sub-procedures set to 100. We can see that both loss functions (Frobenius loss for phenotype matrix and KL loss for genotype matrix) quickly decrease within a few iterations. In fact, for our simulation, the stopping criteria is usually met in less than 50 iterations. Regarding the sensitivity of estimates, Figure 3a shows that the variation across runs with different initializations is relatively low; Figure 3b shows that although the loss function curves may differ in the first few iterations across different initializations, they usually converge to the same levels quickly.

3.2 Application on cardiac mechanics

We then evaluate HNMF on its effectiveness of abstracting raw data into more predictive features. Using the 2 indices of cardiac mechanics listed in Table 1 as the outcome and the patient factor matrix F as the predictors, we train a Ridge regression model. We evaluate the root-mean-square error (RMSE) of our model on the held-out test set, and compare it against two baselines: (b1) Using only genetic variants as regression features; (b2) Using only phenotypic measurements as regression features. We also established five groups of comparison models as follows: (c1) Using only the genetic groups as regression features by applying NMF on the genetic variant matrix only; (c2) In disease with polygenic risk factors, each variant may contribute a small portion of risk, thus we added the total count of risky variants as additional feature to the genetic groups (Liu *et al.*,

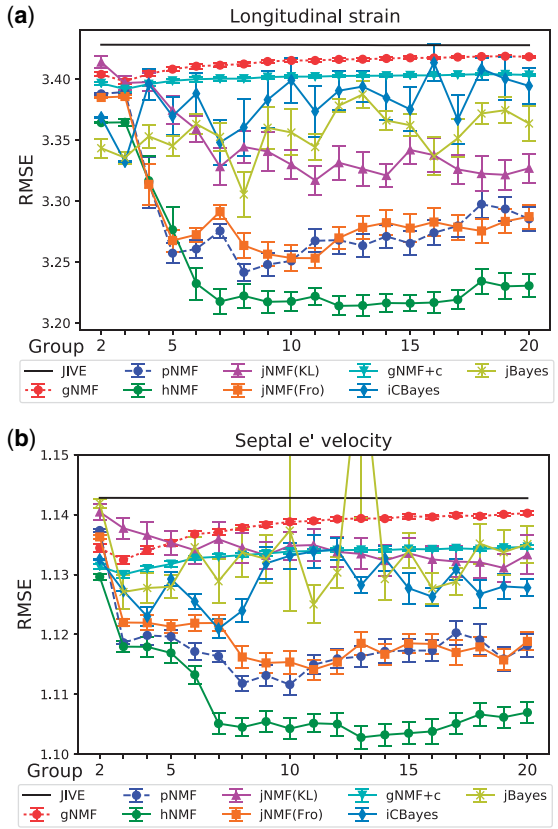

Fig. 4. RMSE with 95% confidence interval for HNMF and comparison methods. gNMF – using genotype factor matrix as features; pNMF –phenotype factor matrix as features; hNMF –hybrid factor matrix as features; jNMF(KL) – joint matrix factorization using KL loss; jNMF(Fro) – joint matrix factorization using Frobenius loss; gNMF+c – genotype factor matrix and the total count of risky variants. Other recently published methods include: iCBayes: Bayesian latent variable model for integrative clustering analysis (Mo *et al.*, 2018); jBayes - Bayesian joint analysis (Ray *et al.*, 2014); JIVE – Joint and individual variation explained (Lock *et al.*, 2013)

2014); (c2) Using only the phenotypic groups as regression features by applying NMF on the phenotypic measurement matrix only; (c3) Using joint matrix factorization but use KL loss for both matrices; (c4) Using joint NMF but use Frobenius loss for both matrices; (c5) Using other recently published methods include: iCBayes—Bayesian latent variable model for integrative clustering analysis (Mo *et al.*, 2018); jBayes—Bayesian joint analysis (Ray *et al.*, 2014); JIVE – Joint and individual variation explained (Lock *et al.*, 2013). For the two suggested Bayesian sampling methods, we used the optimal setting described in their respective papers regarding sampling iterations (burn-in iterations and max iterations, e.g., 3000 and 4000 respectively for jBayes).

We follow Ho *et al.* (2014) on the evaluation procedure in that we vary the group number k from the smallest 2 to where the evaluation metric plateaus and show that across the spectrum HNMF outperforms multiple separate and joint NMF comparison models. The baseline RMSE performances are: 1.25 and 3.88 for geno-baseline on septal e' velocity and longitudinal strain respectively, 1.20 and 3.55 for pheno-baseline respectively. The RMSE performance results of HNMF and comparison models are shown in Figure 4. Comparing all the factorization models and non-factorization models, we can see that using factor matrices as features results in significant improvement (smaller RMSE) over using phenotypic measurements and genetic variants directly as features. Phenotype-only factor matrices often show better regression accuracy than genotype-only factor matrices, likely due to the fact that genetic raw matrix is much sparser than the phenotypic raw matrix. The HNMF factor matrix for regression also significantly outperforms all comparison models including genotype-only or phenotype-only factor matrix for regression, as well as the two joint NMF model results using either KL loss or Frobenius loss for both matrices. This suggests that HNMF can effectively integrate the phenotype and genotype features to predict cardiac mechanics outcomes. HNMF also outperformed recently published methods including iCBayes, jBayes, and JIVE regarding both cardiac mechanics indexes. Note that JIVE is a deterministic model (hence no confidence intervals in the figure) whose performance varies little with the rank of the matrix corresponding to joint variation (hence appearing as a flat line in the figure). The joint Bayesian methods occasionally may have large variations possibly due to the fact that our study has a moderate number of subjects with both phenotype and genetic data. Bayesian sampling based methods likely prefer more subjects to achieve stable estimation while HNMF is more stable as it directly optimizes the objective function. We also noted that jBayes occasionally produced large RMSEs (e.g., $k = 13$ for septal s' velocity), when the corresponding matrices contain large negative entries. This likely suggests overfitting; on the contrary, HNMF produced matrices with entries that have controlled magnitude due to non-negative constraints, and likely reduced overfitting.

3.3 Sensitivity analysis

When performing annotation-based genetic variant filtration, we select the genes that show significant difference in number of LGD variants between the two hypertension groups (patient taking 1 vs. multiple anti-hypertensive medications) by two-tailed binomial exact tests. Using a *P*-value threshold of being less than 0.01 produces 110 genes for our cohort. This is a relatively stringent threshold and in this section we perform sensitivity analysis by varying the *P*-value threshold and including 0.05 and 0.1. With these *P*-value thresholds, we include considerably more genes into consideration: 239 genes for 0.05 as threshold and 349 genes for 0.1 as threshold. The genotype baseline RMSEs are 4.87 (4.63) for longitudinal strain and 1.56 (1.50) for septal e' velocity under *P*-value threshold 0.1 (0.05). Supplementary Figure S1 (Supplementary materials) shows the results of the sensitivity analysis in comparison with Figure 4. Comparing these figures, it is easy to see that under all p-value thresholds, HNMF consistently outperforms all baselines and NMF comparison models including pheno- and geno- separate NMF models and joint NMF models with KL or Frobenius losses. On the other hand, as one tightens the *P*-value threshold, the plateau region becomes wider, suggesting that the regression performance is less sensitive as the group number varies in the plateau region. Thus in the following phenotype and genotype group analysis, we chose *P*-value threshold of 0.01. Another reason is that with a stricter *P*-value threshold, we are more confident that selected genes are likely implicated in the pathogenesis of abnormal cardiac mechanics. We also note that with large enough patient cohort size, techniques such as cross-validations can be used to accurately determine the optimal group number. The larger the patient cohort size, the more effective cross-validation is, under more relaxed filtering criteria that result in more genes to consider.

4 Discussion

Using the method in the feature group discovery section, we identified the top phenotypic and genetic groups that are associated with

Table 3. Top phenotypic and genetic groups (and their representative components) associated with lower values of septal e′ velocity and absolute longitudinal strain (worse cardiac mechanics)

	Top phenotype group	Top gene group
Septal e′velocity	Sodium	GPRC6A
	Calcium	MSMP
	Albumin	NPR2
	Left ventricular ejection fraction	IDI2
	Relative wall thickness	TPM2
Longitudinal strain	Sodium	COX6B2
	Calcium	PAX5
	Albumin	BMP4
	Waist/hip ratio	TPM2
	Sitting heart rate	CLDN5

Note: Paired phenotypic group and genetic group are linked by patient group.

worse cardiac mechanics. Due to space limitation, we only show the top phenotypic and genetics groups associated with lower values of septal e' velocity and longitudinal strain, as listed in Table 3. The phenotypic groups can help us identify variables that are correlated with abnormal cardiac mechanics. The associated genetic group consists of genes that potentially mediate the corresponding multivariable phenotypic abnormality. They collectively indicate problematic multi-factor genotype and phenotype interaction and attribute such interaction to a specific patient group (in F), thus can more comprehensively and precisely characterize and stratify these patients in an evidence-driven fashion.

More specifically, the echocardiographic septal e' velocity is one of several variables used during the assessment of diastolic dysfunction. In general lower septal e' values are reflective of a higher degree of diastolic dysfunction, which is associated with the development of heart failure and/or adverse cardiovascular outcomes (Mitter *et al.*, 2017). In septal-e' phenotype group, preserved (higher) left ventricular ejection fraction is often present in patients with diastolic dysfunction, other variables are associated with the development of diastolic dysfunction, including abnormal sodium, calcium, and albumin levels, and abnormal left ventricular wall thickness during diastole. In the septal-e' gene group, TPM2 shows strong susceptibility to variants that lead to cardiomyopathies and IDI2 to chronic kidney disease (comorbidity and risk factor for cardiovascular disease). NPR2 is linked to cardiac conduction. GPRC6A is responsible for calcium sensing that affects L-type calcium channel and is critical to cardiac cell function (Mackenzie *et al.*, 2005). MSMP is involved in resting heart rate modulation. For longitudinal strain, lower values suggest worse longitudinal systolic function of the subendocardium (inner layer of the heart), thus worse cardiac mechanics (Shah *et al.*, 2014). In longitudinal strain phenotype group, besides abnormal sodium, calcium and albumin levels, both higher waist/hip ratio and faster sitting heart rate have a known association with the development of heart failure (Bui *et al.*, 2011). In the longitudinal strain gene group, COX6B2 is in the cardiac muscle contraction pathway, CLDN5 is expressed in heart muscle, other genes also show strong susceptibility to variants that lead to cardiomyopathies (TPM2), other cardiovascular diseases (BMP4), and obesity as comorbidity (PAX5).

This study is subject to potential limitations. First, we only consider the genetic variants that are in coding regions. Genetic variations in coding regions are thought to be the most clinically significant because they often result in a change in the amino acid sequence of a protein. Thus, variations in coding regions of genes typically are associated with more clinical sequelae than variants in non-coding regions. However, variants in non-coding regions could have clinical implications through gene regulation or epigenetic modifications etc. The lack of non-coding variants is a limitation in our study. Applying HNMF on both coding and non-coding variants will be more computationally intensive. Thus in future work, we will develop a more computationally efficient algorithm, and obtain Whole Genome Sequencing (WGS) data to systematically capture potential regulatory variants. Regarding the identified subgroups, we only assessed and discussed their consistency to known knowledge. In the future, we also plan to provide more evidence, and in particular, biological validation to confirm potential novel discoveries. The second limitation concerns the gene feature selection using the entire patient cohort. This is out of consideration that genetic features are sparse and we only have a moderate sized patient cohort. In addition, we use a categorical label that is different from the final continuous outcomes of cardiac mechanics indexes to reduce the impact on generalizability evaluation. Despite our best efforts, we acknowledge that the impact on generalizability cannot be fully eliminated, and we plan to sequence more subjects from external sites to more strictly evaluate the generalizability of our algorithm and how applicable the selected genes would be to future cohorts. The third limitation concerns the fact that some individuals from the same family may be split into the training set while others in the test set. We did so in order to minimize the potential bias from family variant patterns during model training. This may result in an overly optimistic view of the generalizability. However, as neither HNMF nor all the comparison methods explore the family structure, we expect that their relative performances are similar and models' ranks will hold in general. We also performed additional experiments by assigning all individuals from the split families to the training set, therefore guaranteeing family-preserving training-testing split. As shown in Supplementary Figure S2, these experiments yielded similar numerical results and confirmed our expectation, please refer to the Supplementary Material for more detail.

To sum, we proposed a novel HNMF algorithm that integrates both phenotypic measurements and genetic variants as features in order to subtype patients. HNMF models the approximation error for the phenotypic matrix using Gaussian distribution, and models the variant count for the genetic matrix using Poisson distribution. The objective function is the negative log-likelihood of the data given parameters. We developed an alternating projected gradient descent method to solve the approximation problem. Using the simulated dataset, we demonstrated that HNMF has fast convergence and high accuracy when approximating the true factor matrices. Using the real-world HyperGEN dataset, we demonstrated the effectiveness of HNMF in integrating both the phenotypic and genetic features to derive informative patient subgroupings. We used the patient factor matrix as features to predict the cardiac mechanics outcome variables. We compared HNMF with six different models using phenotype or genotype features directly, using NMF on these features separately, and using joint matrix factorization but with only one type of loss function. HNMF significantly outperforms all comparison models. Analyzing the identified phenotype and genotype groups reveals intuitive phenotype-genotype interactions that characterize cardiac abnormality. For future study, we plan to extend HNMF to consider prior medical knowledge (e.g., known phenotypic and genotypic characteristics associated with heart failure) in guiding the generation of the factor matrices for better patient stratification. We also plan to extend HNMF to a tri-factorization model that allows for different group numbers in patient, genotype and phenotype factor matrices, in order to benefit

HNMF with more flexibility to handle heterogeneous and distinct modality of data sources. We plan to model the genetic matrix approximation using zero-inflated Poisson distribution, as genetic matrix is sparse. We also plan to relax LGD criteria to include more genetic variants and obtain Whole Genome Sequencing data to systematically capture potential regulatory variants.

Funding

This work has been supported in part by the supported by the following grants: NIH 1R21LM012618-01, NIH R01 HL107577, AHA #15CVGPSD27260148, NSF IIS-1716432, NSF IIS-1750326 and ONR N00014-18-1-2585.

Conflict of Interest: none declared.

References

Adzhubei,I. *et al*. (2013) Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet*.,**76**, 7–20.

Bui,A.L. *et al*. (2011) Epidemiology and risk profile of heart failure. *Nat. Rev. Cardiol*., **8**, 30–41.

Chi,E.C., and Kolda,T.G. (2012) On tensors, sparsity, and nonnegative factorizations. *SIAM J. Matrix Analysis Appl*., **33**, 1272–1299.

Collisson,E.A. *et al*. (2011) Subtypes of pancreatic ductal adenocarcinoma and their differing responses to therapy. *Nat. Med*., **17**, 500–503.

DePristo,M.A. *et al*. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet*., **43**, 491.

Ding,C. *et al*. (2005). *On the equivalence of nonnegative matrix factorization and spectral clustering*. Paper presented at the Proceedings of the 2005 SIAM International Conference on Data Mining, Newport Beach, CA, USA, April 21-23, 2005.

Ding,C. *et al*. (2006). *Orthogonal nonnegative matrix t-factorizations for clustering*. Paper presented at the Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, Philadelphia, PA, USA, August 20-23, 2006.

Ding,C.H. *et al*. (2010) Convex and semi-nonnegative matrix factorizations. *IEEE Trans. Pattern Anal. Machine Intell*., **32**, 45–55.

Gunasekar,S. *et al*. (2016) Phenotyping using structured collective matrix factorization of multi–source ehr data. *arXiv Preprint arXiv*, *1609.04466*.

Guo,T. *et al*. (2017) Integrative variants, haplotypes and diplotypes of the CAPN3 and FRMD5 genes and several environmental exposures associate with serum lipid variables. *Sci. Rep*., **7**, 45119.

Harrow,J. *et al*. (2012) GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res*., **22**, 1760–1774.

Ho,J.C. *et al*. (2014). *Marble: high-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization*. Paper presented at the Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, New York, NY, USA, August 24–27, 2014.

Hofree,M. *et al*. (2013) Network-based stratification of tumor mutations. *Nat. Meth*., **10**, 1108–1115.

Howell,D.C. (2012). *Statistical Methods for Psychology*: Cengage Learning.

Katz,D.H. *et al*. (2017) Phenomapping for the identification of hypertensive patients with the myocardial substrate for heart failure with preserved ejection fraction. *J. Cardiovasc. Transl. Res*., **10**, 275.

Kim,H. *et al*. (2015). *Simultaneous discovery of common and discriminative topics via joint nonnegative matrix factorization*. Paper presented at the Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015.

Kim,J., and Park,H. (2011) Fast nonnegative matrix factorization: an active-set-like method and comparisons. *SIAM J. Sci. Comput*., **33**, 3261–3281.

Kircher,M. *et al*. (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet*., **46**, 310.

Kohane,I.S. (2015) Ten things we have to do to achieve precision medicine. *Science*, **349**, 37–38.

Kumar,P. *et al*. (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc*., **4**, 1073–1082.

Lee,D.D., and Seung,H.S. (1999) Learning the parts of objects by non-negative matrix factorization. *Nature*, **401**, 788–791.

Lee,D.D., and Seung,H.S. (2001). *Algorithms for non-negative matrix factorization*. Paper presented at the Advances in Neural Information Processing Systems, Denver, CO, USA.

Lek,M. *et al*. (2016) Analysis of protein-coding genetic variation in 60, 706 humans. *BioRxiv*, 030338.

Lin,C.-J. (2007) Projected gradient methods for nonnegative matrix factorization. *Neural Computation*, **19**, 2756–2779.

Liu,J. *et al*. (2014) New genetic variants improve personalized breast cancer diagnosis. *AMIA Summits on Translational Science Proceedings*, *2014*, 83.

Liu,J. *et al*. (2013). *Multi-view clustering via joint nonnegative matrix factorization*. Paper presented at the Proceedings of the 2013 SIAM International Conference on Data Mining, Austin, Texas, USA, May 2-4, 2013.

Lock,E.F. *et al*. (2013) Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *Ann. Appl. Stat*., **7**, 523.

Luo,Y. *et al*. (2016a) Using machine learning to predict laboratory test results. *Am. J. Clin. Pathol*., **145**, 778–788.

Luo,Y. *et al*. (2016b). *Predicting ICU Mortality Risk by Grouping Temporal Trends from a Multivariate Panel of Physiologic Measurements*. Paper presented at the Proceedings of the 30th AAAI Conference on Artificial Intelligence, Phoenix, Arizona, USA

Mackenzie,P.I. *et al*. (2005) Nomenclature update for the mammalian UDP glycosyltransferase (UGT) gene superfamily. *Pharmacogenet. Genomics*, **15**, 677–685.

Mitter,S.S. *et al*. (2017) A test in context E/A and E/e ' to assess diastolic dysfunction and LV filling pressure. *J. Am. Coll. Cardiol*., **69**, 1451–1464.

Mo,Q. *et al*. (2018) A fully Bayesian latent variable model for integrative clustering analysis of multi-type omics data. *Biostatistics*, **19**, 71–86.

Mor-Avi,V. *et al*. (2011) Current and evolving echocardiographic techniques for the quantitative evaluation of cardiac mechanics: aSE/EAE consensus statement on methodology and indications: endorsed by the Japanese. *Soc. Echocardio. J. Am. Soc. Echocardiography*, **24**, 277–313.

Moreau,Y., and Tranchevent,L.-C. (2012) Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nat. Rev. Genet*., **13**, 523.

Müller,F.-J. *et al*. (2008) Regulatory networks define phenotypic classes of human stem cell lines. *Nature*, **455**, 401–405.

Poulter,N.R. *et al*. (2015) Hypertension. *Lancet*, **386**, 801–812.

Ray,P. *et al*. (2014) Bayesian joint analysis of heterogeneous genomics data. *Bioinformatics*, **30**, 1370–1376.

Selvaraj,S. *et al*. (2016) Association of central adiposity with adverse cardiac mechanics findings from the hypertension genetic epidemiology network study. *Circ. Cardiovasc. Imaging*, **9**, e004396.

Shah,S.J. *et al*. (2014) Ultrastructural and cellular basis for the development of abnormal myocardial mechanics during the transition from hypertension to heart failure. *Am. J. Physiol. Heart Circ. Physiol*., **306**, H88–H100.

Sra,S., and Dhillon,I.S. (2006). *Generalized nonnegative matrix approximations with Bregman divergences*. Paper presented at the Advances in neural information processing systems, Vancouver, British Columbia, Canada.

Stenson,P.D. *et al*. (2012) The human gene mutation database (HGMD) and its exploitation in the fields of personalized genomics and molecular evolution. *Curr Protoc Bioinformatics, Chapter 1*, **13**, Unit1 doi: 10.1002/0471250953.bi0113s39

van Buuren,S., and Groothuis-Oudshoorn,K. (2011) mice: multivariate Imputation by Chained Equations in R. *J. Stat. Software*, **45**, 1–67.

Wang,H.-Q. *et al*. (2015) j NMFMA: a joint non-negative matrix factorization meta-analysis of transcriptomics data. *Bioinformatics*, **31**, 572–580.

Wang,K. *et al*. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*., **38**, e164.

Williams,R.R. *et al*. (2000) NHLBI Family Blood Pressure Program: methodology and recruitment in the HyperGEN network. *Ann. Epidemiol*., **10**, 389–400.