PLOS | ONE

RESEARCH ARTICLE

Using topic modeling via non-negative matrix factorization to identify relationships between genetic variants and disease phenotypes: A case study of Lipoprotein(a) (*LPA*)

Juan Zhao[1], QiPing Feng[2], Patrick Wu[1,3], Jeremy L. Warner[1,4], Joshua C. Denny[1,4], Wei-Qi Wei[1]*

1 Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, United States of America, **2** Division of Clinical Pharmacology, Vanderbilt University Medical Center, Nashville, TN, United States of America, **3** Medical Scientist Training Program, Vanderbilt University School of Medicine, Nashville, TN, United States of America, **4** Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, United States of America

* wei-qi.wei@vumc.org

Abstract

Genome-wide and phenome-wide association studies are commonly used to identify important relationships between genetic variants and phenotypes. Most studies have treated diseases as independent variables and suffered from the burden of multiple adjustment due to the large number of genetic variants and disease phenotypes. In this study, we used topic modeling via non-negative matrix factorization (NMF) for identifying associations between disease phenotypes and genetic variants. Topic modeling is an unsupervised machine learning approach that can be used to learn patterns from electronic health record data. We chose the single nucleotide polymorphism (SNP) rs10455872 in *LPA* as the predictor since it has been shown to be associated with increased risk of hyperlipidemia and cardiovascular diseases (CVD). Using data of 12,759 individuals with electronic health records (EHR) and linked DNA samples at Vanderbilt University Medical Center, we trained a topic model using NMF from 1,853 distinct phenotypes and identified six topics. We tested their associations with rs10455872 in *LPA*. Topics enriched for CVD and hyperlipidemia had positive correlations with rs10455872 *(P* < 0.001), replicating a previous finding. We also identified a negative correlation between *LPA* and a topic enriched for lung cancer (*P* < 0.001) which was not previously identified via phenome-wide scanning. We were able to replicate the top finding in a separate dataset. Our results demonstrate the applicability of topic modeling in exploring the relationship between genetic variants and clinical diseases.

OPEN ACCESS

Citation: Zhao J, Feng Q, Wu P, Warner JL, Denny JC, Wei W-Q (2019) Using topic modeling via non-negative matrix factorization to identify relationships between genetic variants and disease phenotypes: A case study of Lipoprotein(a) (*LPA*). PLoS ONE 14(2): e0212112. https://doi.org/10.1371/journal.pone.0212112

Editor: Zhi Wei, New Jersey Institute of Technology, UNITED STATES

Received: June 13, 2018

Accepted: January 27, 2019

Published: February 13, 2019

Data Availability Statement: The dataset was generated from other previous genetic studies, which has been deposited into dbGap according to the NIH requirement. We have uploaded our scripts on GitHub https://github.com/zhaojuanwendy/pheno_tp/.

Funding: The project was supported NIH grant P50 GM115305, R01 HL133786, R01 GM120523, T32 GM007347 from the National Institute of General Medical Studies for the Vanderbilt Medical-

Using topic modeling via non-negative matrix factorization to identify relationships between genetic variants and disease phenotypes: A case study of Lipoprotein(a) (*LPA*)

Scientist Training Program, 18AMTG34280063 from American Heart Association and T15 LM007450 from the National Library of Medicine for the Vanderbilt Biomedical Informatics Training Program. The dataset used in the analyses described were obtained from Vanderbilt University Medical Centers BioVU which is supported by institutional funding and by the CTSA grant ULTR000445 from NCATS/NIH. Genome-wide genotyping was funded by NIH grants RC2GM092618 from NIGMS/OD and U01HG004603 from NHGRI/NIGMS. The dataset (s) used for the analyses described were obtained from Vanderbilt University Medical Center's BioVU which is supported by institutional funding, the 1S10RR025141-01 instrumentation award, and by the CTSA grant UL1TR000445 from NCATS/NIH. The authors wish to acknowledge the expert technical support of the VANTAGE and VANGARD core facilities, supported in part by the Vanderbilt-Ingram Cancer Center (P30 CA068485) and Vanderbilt Vision Center (P30 EY08126). This study was supported by GM120523, GM109145, HL133786, 5T32GM080178-09, K23AR064768, Rheumatology Research Foundation (K-supplement), American Heart Association (18AMTG34280063, 16SDG27490014 and 15MCPRP25620006), HG008672, R01 LM010685, GM115305 and Vanderbilt Faculty Research Scholar Fund. The dataset used for the analyses described were obtained from Vanderbilt University Medical Center's resources, BioVU and the Synthetic Derivative, which are supported by institutional funding and by the Vanderbilt National Center for Advancing Translational Science grant 2UL1 TR000445-06 from NCATS/NIH. Existing genotypes in BioVU were funded by NIH grants RC2GM092618 from NIGMS/OD and U01HG004603 from NHGRI/NIGMS. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Elucidating associations between genetic variants and human diseases creates new avenues for disease prevention and enables precise identification and treatment of diseases [1,2]. During the past two decades, genetic studies have uncovered thousands of genetic variants that influence risk for disease phenotypes [3], e.g., the discovery of a variant in proprotein convertase subtilisin/kexin type 9 (*PCSK9[4]*) associated with low plasma low-density lipoprotein, which led to a new therapeutic drug class that was approved by the US Food and Drug Administration in 2015. Many of these discoveries come from large-scale association analyses. The two most notable approaches are genome-wide (GWAS) and phenome-wide association studies (PheWAS) [2, 5]. For a given phenotype, GWAS scans hundreds of thousands to millions of single nucleotide polymorphisms (SNPs) across the genome in a hypothesis-free approach. PheWAS, conversely, analyzes thousands of disease phenotypes compared to a single SNP. In a GWAS, the outcome variable is a disease phenotype and predictor variables are SNPs. In a PheWAS, the outcome variable is a SNP and predictor variables are disease phenotypes. Although the output is different, these techniques share many commonalities.

In particular, association analyses test many predictors at one time and assume that each predictor has an independent effect. Nevertheless, diseases often occur as a group of comorbidities, e.g. hyperlipidemia (HLD) and cardiovascular diseases (CVDs). Conventional association analyses may not capture the inter-connections among variables such as phenotypes and thus may not be sensitive to identify important genotype-phenotype relationships. Moreover, association analyses also face the challenge of scaling to an increasing number of phenotypes. Previously, we have described a "networked PheWAS" approach which can address interconnectivity but still requires a degree of supervised interpretation [6].

This study tested the feasibility of topic modeling, an unsupervised machine learning method, for identifying relationships between genetic variants and disease phenotypes. Topic modeling was initially introduced as a text mining technique [7,8]. It extracts latent topics or themes from documents and thus facilitates the understanding of data [9,10]. Each document can have multiple topics. Each topic can be represented by related words (or other inputs). Topic modeling has achieved notable applications in text mining [11,12], social networks [13] and computer vision [14]. Recently, it has been brought into the biomedical research field, primarily focusing on clinical text mining [15–17]. A few studies applied this technique to mining EHRs' events [17–21] and capturing their relationship with the genetic information [15,22,23], e.g. McCoy et al. used latent Dirichlet allocation to find disease clusters and examined their associations with the polygenic risk scores of depression [20].

Our study utilized non-negative matrix factorization (NMF), a topic modeling method to one of the largest biobanks in the U.S. We hypothesized that topic modeling could identify disease clusters among the phenome, which can help replicate known findings and uncover unidentified relationships between genetic variants and disease phenotypes. We used NMF [24,25] to identify latent topics (e.g. disease clusters or relevant comorbidities) from EHR data. We then tested associations between the EHR-derived topics and a lipoprotein(a) (*LPA*) SNP (rs10455872). We chose this *LPA* SNP because previous studies have shown that high levels of the *LPA* protein product, Lp(a), are associated with increased risks of developing HLD and CVD [26]. Specifically, the *LPA* SNP (rs10455872), as a single variant, explains 20–30% of the variation in circulating Lp(a) levels, which makes it an ideal candidate for this study [27]. To demonstrate the benefit of using topic modeling, we compared our result with a traditional PheWAS approach.

Materials and methods

Study cohort

We used data from BioVU, the de-identified DNA biobank at Vanderbilt University Medical Center (VUMC), to conduct this study. BioVU contains DNA samples from >250,000 individuals linked with de-identified EHRs, including diagnostic and procedure codes, clinical notes, laboratory values and medications. We limited our study within a group of selected individuals who received regular, longitudinal care at Vanderbilt to avoid the incomplete data issue. We identified 12,759 adult individuals of European ancestry (Female/Male: 6,018/6,741; mean age: 70.3±12.3) who had both EHRs and genotyped data of rs10455872.

rs10455872 Genotyping

We extracted individual's rs10455872 information from available genotype data. All genotyping was previously conducted using commercially available genome-wide SNP arrays with quality control criteria for variants. Genotype imputation was conducted on the Michigan Imputation server[28] with minimac3[29], using the Haplotype Reference Consortium reference panel, version r1.1[30].

Among the cohort of 12,759 individuals, we observed 85.2% AA, 14.2% AG, 0.6% GG. The minor allele frequency (MAF) of the rs10455872 G allele is 7.7% in our cohort, consistent with the 7% MAF in the European population [31]. We used 0, 1, and 2 to represent the number of *LPA* rs10455872 G alleles that each individual carries.

Disease phenotypes

Following established protocols used in past studies [32], we grouped each individual's International Classification of Disease, 9th edition (ICD-9-CM) codes into disease "phecodes" [33]. There were 1853 phecodes extracted from 12,759 individuals. For each phecode, we labeled individuals with the phecode with a '1', and those having no such phecode with a '0'.

Topic modeling via non-negative matrix factorization (NMF)

We use topic modeling via NMF to extract a set of topics (i.e. clusters of disease) from individuals' phenotypes data (Fig 1). In this study, we used a data matrix X of dimensions n×m to represent the input data, where n denotes the number of an individual (e.g. $n = 12{,}759$), and m denotes the size of the phecodes (e.g. $m = 1853$). The entry of the matrix $X_{ij} \in X$ was a binary value (0 or 1) indicating whether i^{th} individual had the j^{th} phecode.

NMF is based on assumption that a $n\times m$ sparse data matrix X of with n samples and m dimensions, which can often be represented by a small sets of k basis vectors. The linear combination is a $n\times k$ coefficients matrix W, which is a lower-dimensional representation for X.

To be more specific, the input data matrix X can be approximately represented by the product of two non-negative matrices W and H, such that

$$\min_{W\geq 0, H\geq 0} \|X - WH\|_F^2 + \lambda R(W, H), \tag{1}$$

where $\|X - WH\|_F^2$ is a Frobenius norm to measure the error between approximation and the original data, that is, $\|X - WH\|_F^2 = \sum_{i,j} (X - WH)_{ij}^2$. $\lambda R(W, H)$ is a regularization term, where λ is the parameters of the regularization term.

H is a k×m topic–phenotype matrix with k rows and m columns, where k denotes the size of topics and m is the size of phenotypes. It is useful to view each row vector in H as a disease

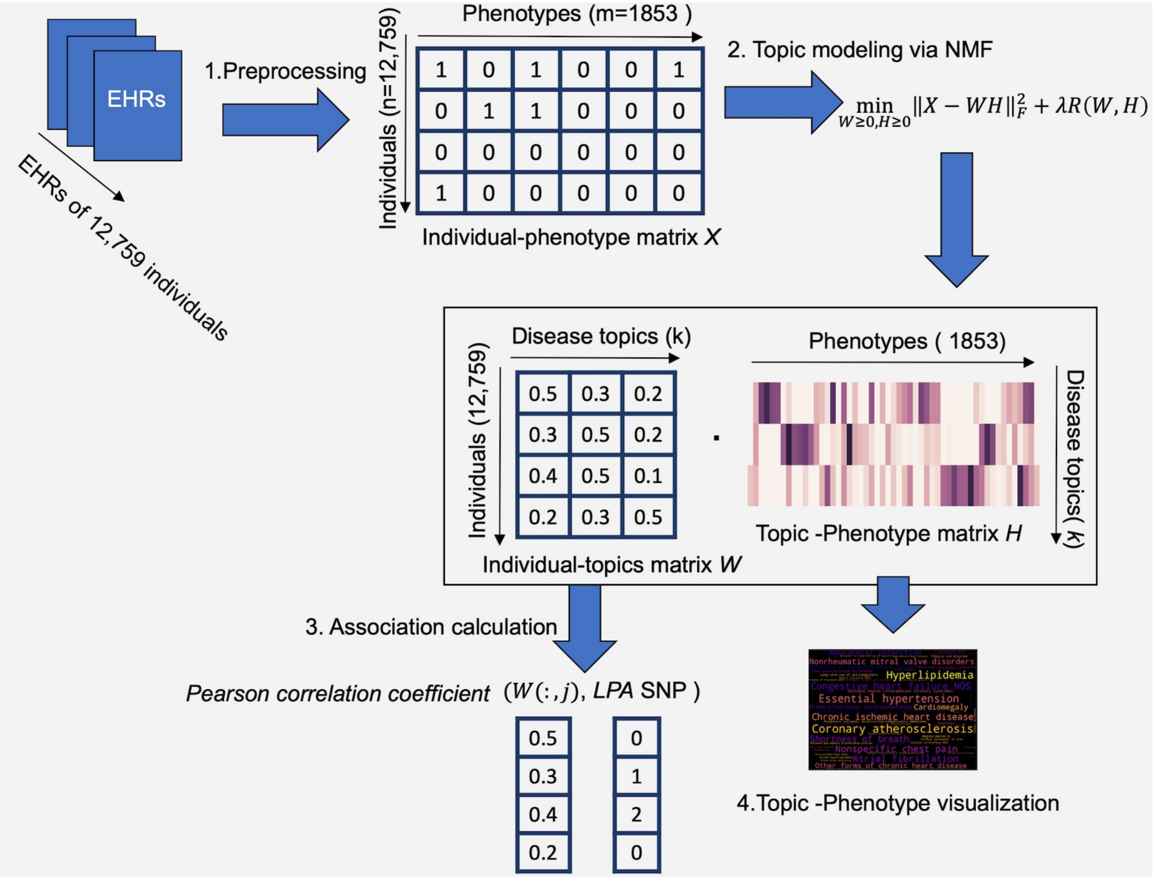
EHRs
EHRs of 12,759 individuals
1.Preprocessing
Phenotypes (m=1853)
Individuals (n=12,759)
1 0 1 0 0 1
0 1 1 0 0 0
0 0 0 0 0 0
1 0 0 0 0 0
Individual-phenotype matrix X
2. Topic modeling via NMF
$\min_{W\geq0,H\geq0} \|X - WH\|_F^2 + \lambda R(W,H)$
Disease topics (k)
Individuals (12,759)
0.5 0.3 0.2
0.3 0.5 0.2
0.4 0.5 0.1
0.2 0.3 0.5
Individual-topics matrix W
Phenotypes (1853)
Disease topics(k)
Topic -Phenotype matrix H
3. Association calculation
Pearson correlation coefficient (W(:, j), LPA SNP)
0.5 0.3 0.4 0.2
0 1 2 0
Myocardial infarction
Nonrheumatic mitral valve disorders
Systolic or combined heart failure
Long-term use of anticoagulants
Hyperlipidemia
Congestive heart failure NOS
Unstable angina (intermediate coronary syndrome)
Essential hypertension
Primary/intrinsic cardiomyopathies
Cardiomegaly
Chronic ischemic heart disease
Coronary atherosclerosis
Shortness of breath
Angina pectoris
Nonspecific chest pain
Atrial fibrillation
Other forms of chronic heart disease
4.Topic -Phenotype visualization

Fig 1. Illustration of topic modeling on EHRs using NMF.

https://doi.org/10.1371/journal.pone.0212112.g001

topic represented by a set of phenotypes, where each cell value defines the phenotype' rank in the topic.

W is an *n*×*k* individual-topic matrix, with *n* individuals and *k* topics. We can view the row vector of *W* as an individual's EHR document with a cell value indicating the individual's relevance to a disease topic. *W* is then used for the association analysis between the topics and rs10455872.

$R(W, H)$ is the regularization term that combines L_1 and L_2 norms, which is defined as:

$$R(W,H) = \gamma(\|W\|_1 + \|H\|_1) + \frac{1}{2}(1-\gamma)\left(\|W\|_F^2 + \|H\|_F^2\right), \tag{2}$$

where γ is the ratio for L_1 penalty. The regularization term is not necessarily included in NMF, but adding the regularization term may help balance the sparsity of the topics, as we assume that each individual may have a small set of diseases.

Topic evaluation

To evaluate topics results, we asked domain experts (authors with clinical background) to view the words describing a topic to determinate if the topic is semantically meaningful. In addition, we used two objective measures, which were also commonly used in topic modeling evaluation—topic dependency and topic coherence (S1 Text) in [34,35]. Topic dependency reflects the overlapping between topics by calculating mean pairwise Jaccard similarity between the topic descriptors (i.e. top-ranked words in a topic). A higher number of similar terms between topics suggest that topics are overlapped (less useful) [34]. Topic coherence reflects if a topic can represent a single theme or similar concepts by measuring the co-concurrency of the topic descriptors in the whole documents. Topics with higher topic coherence are usually easy to interpret.

Since pre-set values of parameters such as γ and λ may impact the results, we tuned the parameters and used agreement scores to measure the stabilities of topics. We calculate the stability of topics using Jaccard similarity and Hungarian method [36]. A higher agreement score indicates a higher stability, i.e., the concepts in a topic remain consistent regardless of the parameter values.

To visualize the topic results, we employed word clouds to present top-ranked phenotypes in a topic. We used 2-dimensional (2D) visualizations to visualize the disease clusters of the cohort by using t-Distributed Stochastic Neighbor Embedding (t-SNE) [37].

Statistical analysis

We tested the association between each extracted topic and *LPA* SNP variation by applying Pearson correlation coefficient (PCC) and logistic regression (LR). PCC measures the strength of a linear association between two variables, and generates a correlation coefficient denoted by $r \in [-1,1]$, showing correlation direction. To compute PCC, we used each topic vector of individual-topic matrix W in NMF as the predictor variable x, where each cell defines the relevance scores on each individual. We used the vector of *LPA* SNP, where each cell represents the number of risk allele, as the variable y. For LR, we used the topic vector as the predictor, and *LPA* SNP as the outcome variable (counts of alleles > 1 is treated as 1), adjusting for age and sex (i.e. 1 represents female; 0 represents male). We reported the coefficient and p-value for each predictor.

To compare the results, we also conducted a conventional PheWAS in our cohort (i.e. 12,759 individuals) to test the association between for *LPA* SNP with 1853 phenotypes. PheWAS is a systematic approach to replicate and discover relationships between targeted genotypes and multiple phenotypes [5].

Validation Cohort

To demonstrate the generalizability of this approach, we also repeated the process using the data from a separate cohort and tested whether or not we can replicate our findings. The data was collected from another project to study stains. All individuals are European ancestry and under statin treatment. The cohort had 3889 adult individuals with more percentage of males and similar age as the study cohort (Male/Female:2,473/1,416 [63.6% vs 52.8%]; mean age: 71.9±11.3 [71.9 vs 70.3]). The study cohort and the validation cohort were mutually exclusive.

Results

We applied a topic modeling algorithm using NMF on the dataset of 12,759 individuals and obtained six topics. Topics are reviewed by a clinician to ensure clinically meaningful, and labeled with a major disease (Fig 2). The topics were coherent and consistent with the comorbidities associated with the phenotypes that is most prevalent in the cohort. For example, topic #2 defined diseases related to CVD (e.g., HLD, hypertension, and chronic ischemic heart disease), topic #3 represented phenotypes relevant to lung cancer and its treatment.

We plot the distribution of the numbers of topics in the cohort in Fig 3. Topic #2 was the most prevalent (33%) topic in the cohort. Topics #1 and #3 were the second and third most

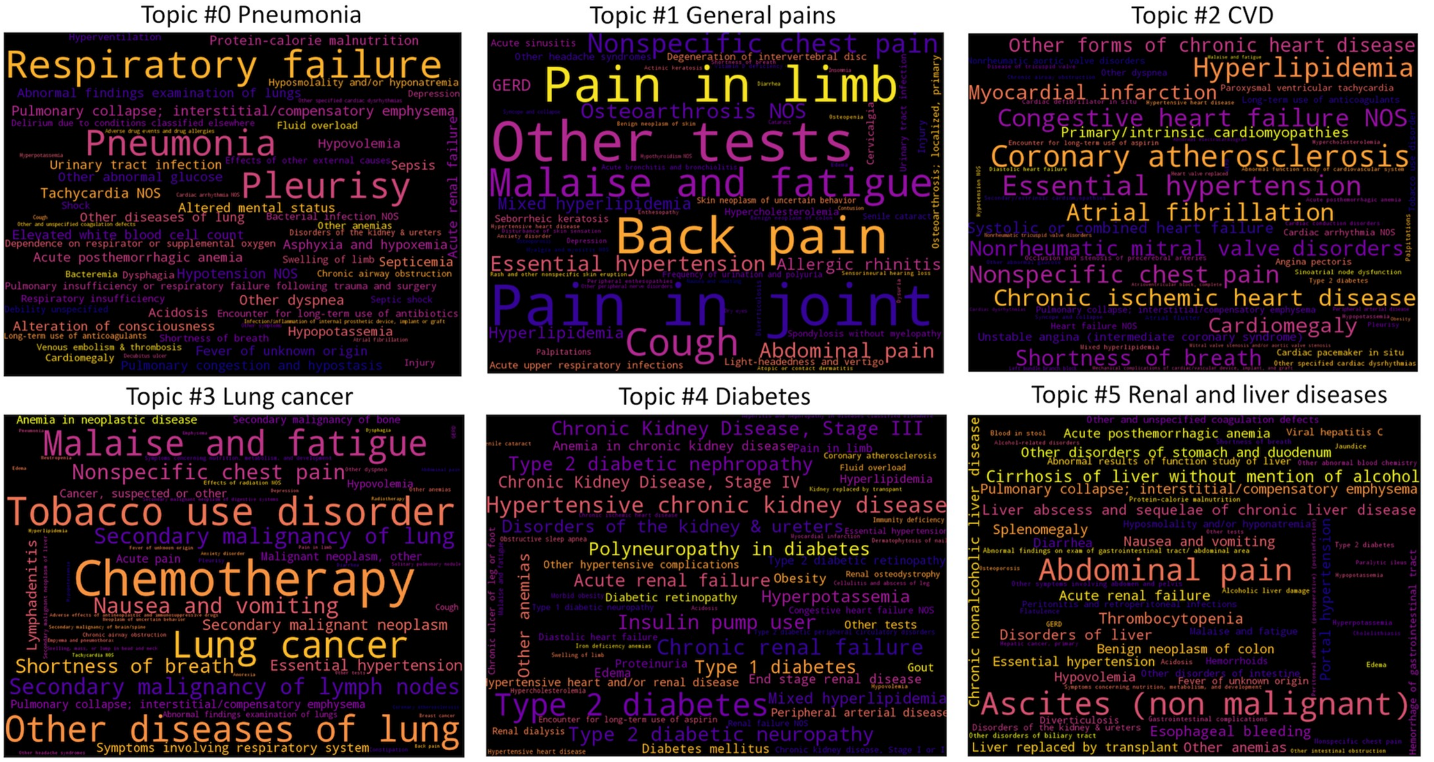

Fig 2. Word clouds for six topics. The size of the words (phecode) in each cloud indicates the weights of the phenotypes on the topic. Phenotypes with larger-sized words have greater influence on the topic compared to phenotypes with smaller-sized words. For each word cloud, we listed the top 60 words.

https://doi.org/10.1371/journal.pone.0212112.g002

prevalent topics in the cohort. We also plotted the scores of individual-phenotypes matrix (*W*) with boxplot in S1 Fig.

We present the visualization of topic modeling results by t-SNE in Fig 4. Each data point represented an individual. We labeled each individual with the assigned topic. We used principal component analysis (PCA) for t-SNE embedding initialization. PCA is a feature reduction method to project high-dimensional data into a lower dimensional space that can explain the most variance of the input data. PCA initialization is more globally stable than random initialization[38]. Consistent with our clinical observation, the topic #2 contains the largest number of individuals in the cohort.

The PCC association test suggests that topic #2 and #3 were significantly associated with rs10455872 (Table 1). Topic #2, a group of lipid and cardiovascular diseases, had a weak but significant positive correlation with rs10455872 ($r = 0.072$, $p = 5.8e\text{-}16$). We also found that topic #3, a group of phenotypes relevant to lung cancer, had a weak but significant negative correlation with rs10455872 ($r = -0.039$, $p = 8.5e\text{-}6$). Although the *r* coefficient is weaker than the topic#2, these correlations are highly statistically significant.

The LR analysis achieved a similar result in the direction and significance for most topics (Table 2). Particularly, topic #2 (CVD) had a positive correlation with rs10455872 (coefficient = 2.789, $p = 3.42E\text{-}13$), and topic #3 (lung cancer) vary inversely with rs10455872 (coefficient = -1.101, $p = 0.009$), which was consistent with PCC test.

PheWAS results on the same data (Fig 5) suggested a significant association between rs10455872 and phenotypes including coronary atherosclerosis, unstable angina, hyperlipidemia and myocardial infarction. Most of these phenotypes were present in topic #2.

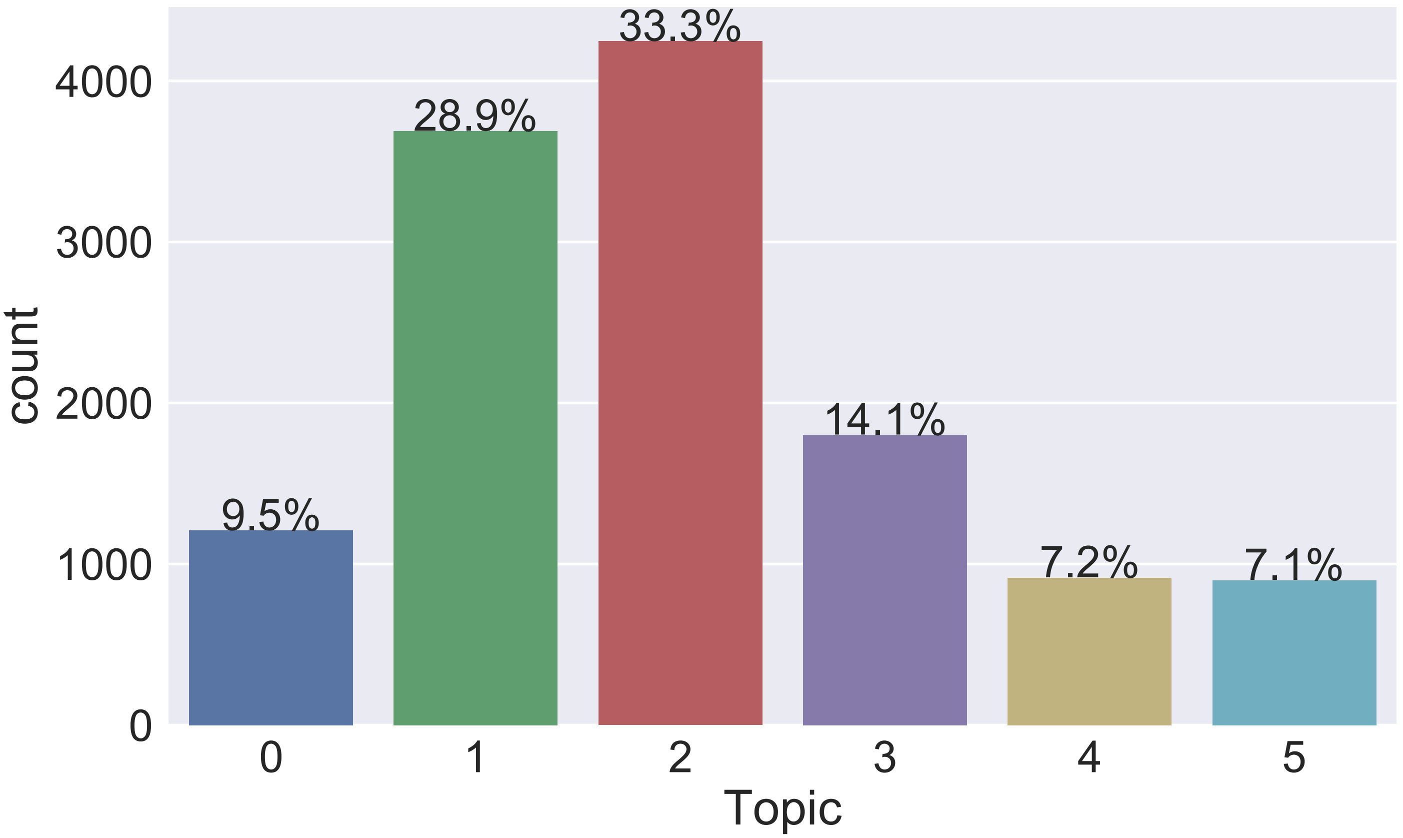

Fig 3. Topic distribution in the cohort. To visualize the prevalence of each topic in the cohort, we assigned an individual to the topic with the maximum score.

https://doi.org/10.1371/journal.pone.0212112.g003

Phenotypes about neoplasms had the second higher associations. However, the association (p-value = 7.209713e-05) did not cross the Bonferroni.

We also validated the approach using data from a separate cohort, without any patient overlap with the study's original cohort. The replicated cohort contains 3889 adults of European ancestry with a larger percentage of males (63.6% vs 52.8%) and similar age (71.9 vs 70.3). The results suggested one positively associated topic containing similar diseases including Coronary atherosclerosis, Essential hypertension, Hyperlipidemia, Nonspecific chest pain, Atrial fibrillation (S6 and S7 Figs, S1 and S2 Tables). We were not able to validate the lung cancer topic, which may due to the smaller data set and limited statistical power.

Discussion

In this paper, we applied topic modeling to explore associations between disease phenotypes and genetic variants. We assumed that some disease phenotypes found simultaneously in EHR have correlated semantic meanings and thus can be learned as topics. We examined the associations between an *LPA* variant (rs10455872) and the six topics derived from EHRs. We observed the expected association between rs10455872 and a topic representing CVD/HLD. We also found a novel association, as of this writing [39], between the *LPA* variant and a lung cancer topic.

The *LPA* gene encodes lipoprotein (a), a major component of the Lp(a) particle. Individuals with elevated Lp(a) levels are more likely to develop CVD compared to those with normal or low Lp(a) level [27,40]. Approximately 70% of Lp(a) variation can be attributed to variants at

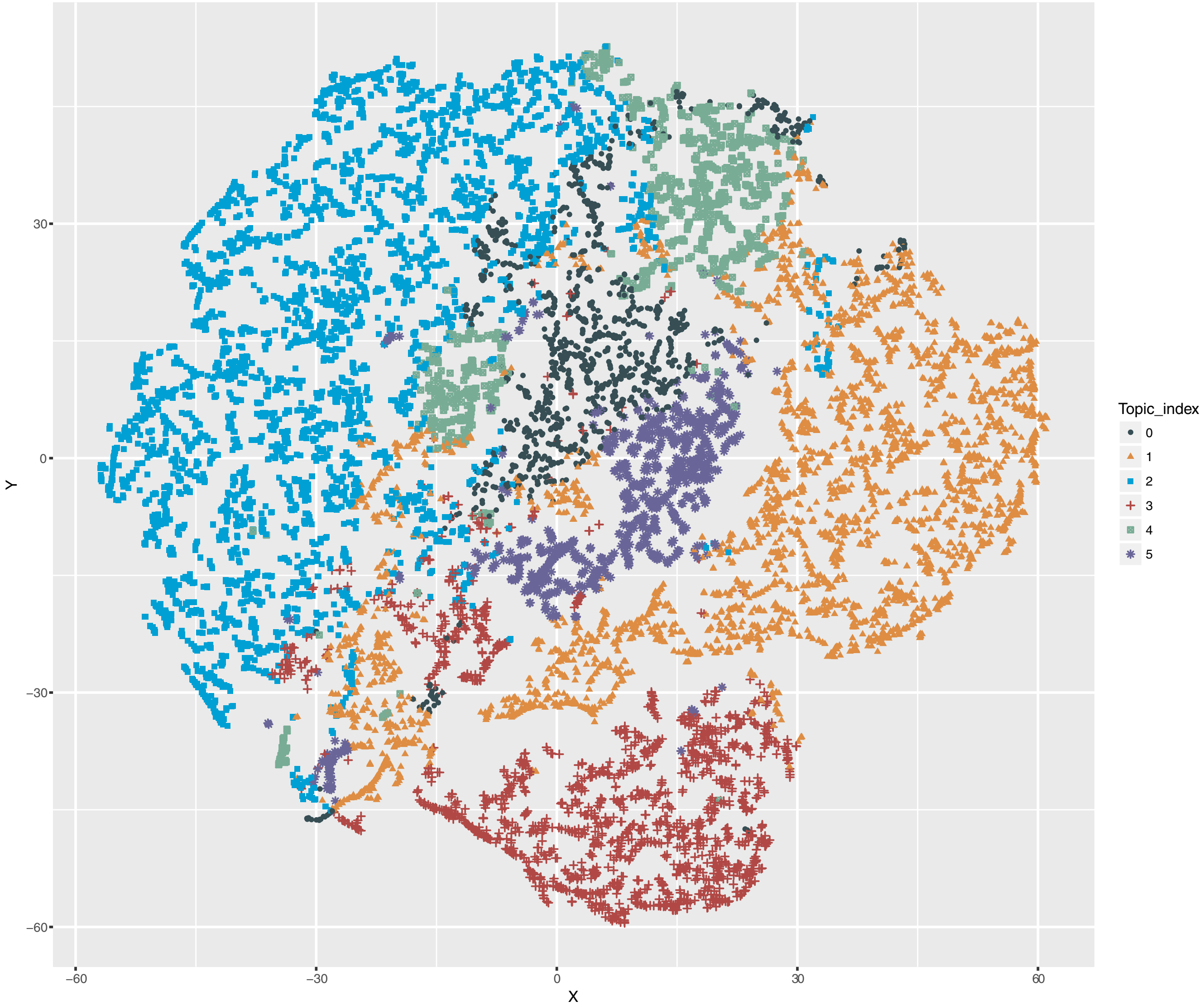

Fig 4. t-SNE plot of visualizing the patient clusters in a projected 2D metric map (The perplexity was set to 30).

https://doi.org/10.1371/journal.pone.0212112.g004

the *LPA* locus [41–43], and rs10455872 alone explains ~25% variation in circulating Lp(a) levels [27]. Further, a previous genetic study suggested that *LPA* variants were strong predictors for CVD risk [27]. In a more recent study of >10,000 patients taking statins, we found that rs10455872 predicted residual CVD risk while on lipid-lowering treatment [44]. This study's finding of a significant association between rs10455872 and the CVD/HLD topic demonstrates the feasibility of topic modeling as a critical tool for uncovering genotype-phenotype relationships.

We also observed a negative correlation between the *LPA* variant and the cancer/lung cancer topic, i.e., possessing this variant is protective. Previous epidemiological studies have reported that individuals with low Lp(a) levels had increased risk of all-cause and cancer-related mortality [45]. Mieno et al. found that hypolipoproteinemia(a) is a risk factor for cancer except for lung cancer. Nevertheless, there are few reports on a relationship between cancer and *LPA* polymorphism or expression levels. Our PheWAS analysis on the same cohort

Table 1. Pearson correlation coefficient testing between LPA variant for each topic.

Topic	Top phenotypes in this topic	*r*	*P*-value
#0	Respiratory failure, Pneumonia, Pleurisy, Pulmonary collapse; interstitial/compensatory emphysema, Hypotension NOS, Tachycardia NOS, Other dyspnea, Hypopotassemia, Sepsis, Septicemia	0.011	0.199
#1	Pain in joint, Other tests, Back pain, Pain in limb, Malaise and fatigue, Cough, Nonspecific chest pain, Essential hypertension, Osteoarthrosis NOS, Abdominal pain	-0.008	0.358
#2	Coronary atherosclerosis, Essential hypertension, Hyperlipidemia, Congestive heart failure NOS, Nonspecific chest pain, Atrial fibrillation, Chronic ischemic heart disease, Shortness of breath, Nonrheumatic mitral valve disorders, Cardiomegaly	0.072	5.8e-16
#3	Chemotherapy, Tobacco use disorder, Lung cancer, Other diseases of lung, Malaise and fatigue, Secondary malignancy of lymph nodes, Secondary malignancy of lung, Nausea and vomiting, Nonspecific chest pain, Shortness of breath	-0.039	8.5e-6
#4	Type 2 diabetes, Hypertensive chronic kidney disease, Chronic renal failure, Insulin pump user, Type 2 diabetic neuropathy, Chronic Kidney Disease, Stage III, Type 2 diabetic nephropathy, Type 1 diabetes, Polyneuropathy in diabetes, Acute renal failure	0.002	0.783
#5	Ascites (nonmalignant), Abdominal pain, Cirrhosis of liver without mention of alcohol, Thrombocytopenia, Liver abscess and sequelae of chronic liver disease, Portal hypertension, Chronic nonalcoholic liver disease, Disorders of liver, Esophageal bleeding, Nausea and vomiting	-0.02	0.021

https://doi.org/10.1371/journal.pone.0212112.t001

identified an association between rs10455872 and secondary malignancy of lymph nodes with borderline p-value (10e-5) but insignificance. To further explore this association between rs10455872 and the cancer/lung cancer topic, we queried gene2pheno (https://imlab.shinyapps.io/gene2pheno_ukb_neale/), which is a publicly available database for testing associations between predicted gene expression levels and phenotypes using data from the UK Biobank. Genetically predicted LPA expression levels were associated with death from T cell lymphomas (p = 6.9e-5, Underlying [primary] cause of death: ICD10: C84.5 Other and unspecified T-cell lymphomas). Given that lung cancer is strongly mediated by environmental exposure and that tobacco use disorder was also part of topic #4, it is possible that the SNP is a marker for propensity to smoking, e.g., similar to what was shown for rs16969968 [46]. Further genetic and epidemiological studies are needed to elucidate the relationship between Lp(a) levels and cancer incidence.

Compared with traditional PheWAS, we identified the most significant associations–between rs10455872 and CVD related diseases, consistent with PheWAS. We also identified a significant association between rs10455872 and cancer/lung cancer, which did not cross the Bonferroni in PheWAS. Besides, our approach automatically identified comorbidities and the interconnections between phenotypes, which cannot be easily identified by PheWAS. We conducted an association test of the comorbidities as a whole instead of each independent

Table 2. Logistic regression analysis between LPA variant for each topic.

Predictor	Coefficient	*P*-value
Age	-0.003	0.079
Sex	0.145	0.005
topic_0	0.542	0.166
topic_1	-0.07	0.820
topic_2	2.789	3.42E-13
topic_3	-1.101	0.009
topic_4	-0.446	0.275
topic_5	-0.695	0.131

https://doi.org/10.1371/journal.pone.0212112.t002

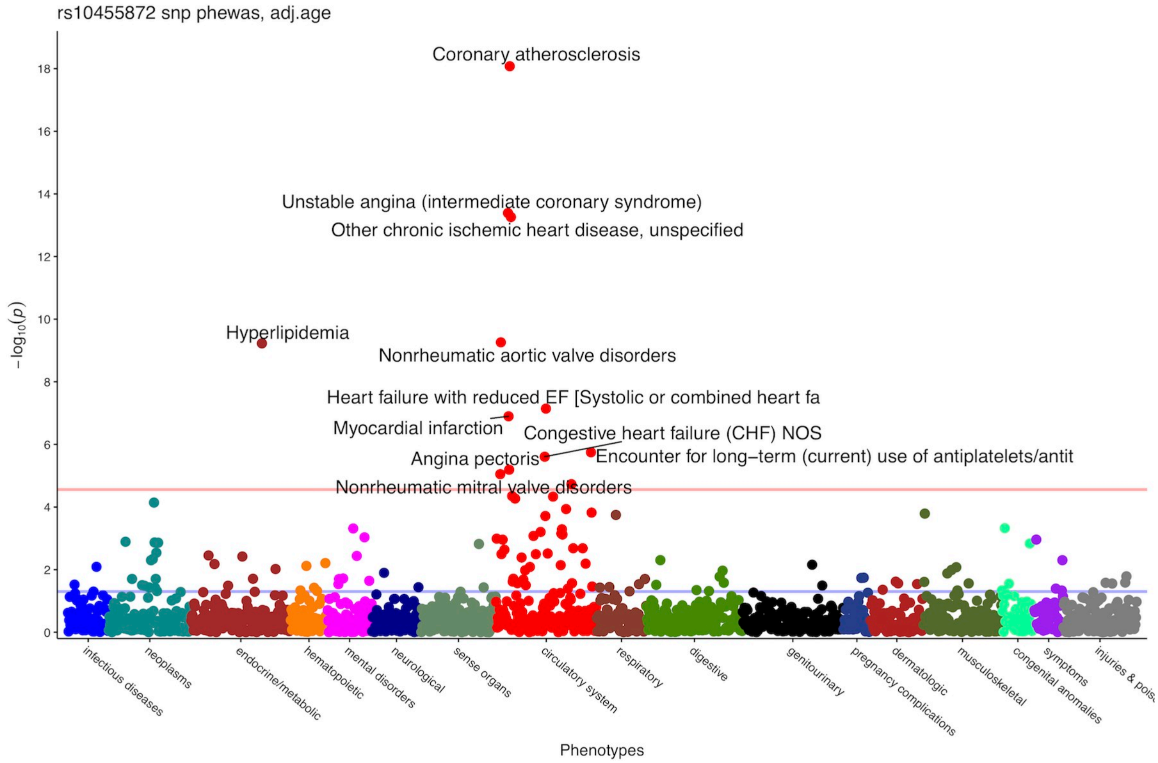

Fig 5. PheWAS results of rs10455872 on 12,759 individuals adjusted by sex and age.

https://doi.org/10.1371/journal.pone.0212112.g005

phenotype. For example, hyperlipidemia often presents with hypertension, as co-occurred in our topic #2. PheWAS can identify the significant association between the genetic variant and hyperlipidemia but cannot automatically identify the cluster of hyperlipidemias and test the relationship between the variant and the cluster.

Compared with existing work on using topic modeling on EHR events [19,20], we explored the choice of number of topics, evaluated the quality of the topics and tested on replicated cohorts.

Choice of number of topics

Topic modeling approaches require pre-specification of the number of topics *k*. In this study, we set *k* = 6, because we focused on capturing the most prevalent diseases such as CVD for quantifying the association. Larger *k* may allow the discover associations between rare diseases and genetic variants, but increases the risk of fracturing common phenotype clusters.

It can be seen that (Fig 4), except for topic #4 (diabetes), the learned topics formed distinct clusters, indicating a good quality of topic modeling. Some of points in topic #4 (diabetes) were close with topic #2 (CVD), which was expected, because type II diabetes is an important risk factor that increases the risk of developing CVD. Compared to the other topics, #1 (Pain), #2 (CVD), and #3 (Lung Cancer) have more concentrated clusters.

For optimal selection of *k* , common approaches includes looking at the error in optimization or having domain experts review the topics to identify which set of topics are most meaningful, and have estimated *k* using singular value decomposition (SVD) to look at the decay of singular values [47–49]. We showed the decay of eigenvalues of the input data using the scree plot (S2 Fig). The top 5–6 components can explain the most variance of the input data.

Therefore, we chose $k = 6$ in this study. Additionally, we also tried $k = [10, 20,30]$. Topics such as CVD and Lung cancer remained associated (S1 Appendix).

We evaluated the topic dependency and average topic coherence by varying k from 5 to 50 in S3 and S4 Figs [34]. The topic dependency drops when k increases from 5 to 15, and then became stable. The average coherences of topics kept dropping as k increases, which means that as the number of topics increases, it becomes difficult to interpret the meaning of the topics.

Stability evaluation

We evaluated the impact of the parameters values (i.e. number of topics k, regularization parameters λ and γ in equations [2] and [3]) on the topic modeling results and the correlation study. In the study, we used $\lambda = 0.2$ and $\gamma = 0.5$ as the default setting. We compared the agreement scores of topics generated on different settings of λ and r with their default values. Results in S5 Fig indicated that λ and γ did not impact the consistency of the topic meaning. We also listed top-ranked phenotypes for each topic and statistical analysis results by using different combinations of λ *and* γ on $k = [10, 20, 30]$ (S1 Appendix). CVD and lung cancer topics were present in all parameter's settings, and their correlations with LPA from PCC and LR analysis were consistent with topics at k = 6 and the default values of λ and r.

Limitations

There are several limitations in this study. First, we tested only one genetic variant in one gene. Rs10455872 explains approximate 25% change in circulating Lp(a) levels according to previous studies; however, it would be interesting to generate a genetic risk score for Lp(a) levels and test its association with disease phenotypes in the future. Second, we used a binary value to indicate if an individual had a diagnosis code. A method accounting for disease severity (e.g., counts of diagnosis codes) could be used in future studies. Finally, the current study was limited to using billing codes to phenotype individuals. We did not include other information, e.g. laboratory test and medications, to assign more accurate phenotypes. This problem can be solved in the future using more sophisticated "deep" phenotyping methods that include more features from EHRs.

Conclusion

In summary, unlike traditional PheWAS that have treated each disease phenotype as a distinct variable, topic modeling via NMF generates more abstract latent factors from disease phenotypes and significantly reduces the number of multiple tests. Our results demonstrate the power of topic modeling in the detection of disease clusters and previously unexplored genotype-phenotype relationships among a large cohort.

Supporting information

S1 Text. Topic evaluation algorithms.
(DOCX)

S1 Appendix. Topic modeling and correlation study results with *topic k = 10, 20,30 and various settings of λ, γ*.
(XLSX)

S1 Fig. Boxplot of individual-phenotypes matrix *W* (after *l2* normalization) to visualize the topic distribution in the cohort.
(EPS)

S2 Fig. Scree plot of topics for $k \in [1, 20]$. This scree plot shows that the eigenvalues start to form a straight line after the sixth principal component. Therefore, the remaining principal components account for a small proportion of the variability and may be less important.
(EPS)

S3 Fig. Topic dependency measured by mean pairwise Jaccard similarity for different *k*.
(EPS)

S4 Fig. Topic coherence for different k.
(EPS)

S5 Fig. Topic stability with tuning parameters $\lambda \in [0, 2]$ and $\gamma = [0.5, 1]$ on $k = [6, 10, 20, 30]$ A: Topic stability for varying $\lambda \in [0, 2]$ on $\gamma = 0.5$; Topic stability for varying $\lambda \in [0, 2]$ on $\gamma = 1$.
(EPS)

S6 Fig. Scree plot of topics for $k \in [1, 20]$ on replicated cohort. This scree plot shows that between the second and third eigenvalues, a straight line starts to form after the third principal component. Therefore, we chose to use 3 topics in the validation study, as the remaining principal components account for a small proportion of variability in the data, and thus may be less important.
(EPS)

S7 Fig. Word clouds for three topics on replicated cohort.
(TIFF)

S1 Table. Pearson correlation coefficient testing between LPA variant for each topic on replicated cohort. * indicates significant association ($p<0.05$).
(DOCX)

S2 Table. Logistic regression result between LPA variant for each topic on replicated cohort. * indicates significant association ($p<0.05$).
(DOCX)

Author Contributions

Conceptualization: Juan Zhao, QiPing Feng, Patrick Wu, Joshua C. Denny, Wei-Qi Wei.

Data curation: QiPing Feng, Joshua C. Denny, Wei-Qi Wei.

Formal analysis: Juan Zhao, QiPing Feng, Patrick Wu, Jeremy L. Warner, Joshua C. Denny, Wei-Qi Wei.

Funding acquisition: QiPing Feng, Joshua C. Denny, Wei-Qi Wei.

Investigation: Juan Zhao, QiPing Feng, Patrick Wu, Wei-Qi Wei.

Methodology: Juan Zhao, QiPing Feng, Patrick Wu, Wei-Qi Wei.

Project administration: Wei-Qi Wei.

Resources: QiPing Feng, Joshua C. Denny, Wei-Qi Wei.

Software: Juan Zhao, QiPing Feng.

Supervision: QiPing Feng, Joshua C. Denny, Wei-Qi Wei.

Validation: Juan Zhao, Patrick Wu, Jeremy L. Warner, Joshua C. Denny, Wei-Qi Wei.

Visualization: Juan Zhao, Patrick Wu.

Writing – original draft: Juan Zhao, QiPing Feng, Wei-Qi Wei.

Writing – review & editing: Juan Zhao, QiPing Feng, Patrick Wu, Jeremy L. Warner, Joshua C. Denny, Wei-Qi Wei.

References

1. Denny JC, Van Driest SL, Wei W-Q, Roden DM. The Influence of Big (Clinical) Data and Genomics on Precision Medicine and Drug Development. Clinical Pharmacology & Therapeutics. 2018; 103: 409–418. https://doi.org/10.1002/cpt.951
2. Manolio TA. Genomewide Association Studies and Assessment of the Risk of Disease. Feero WG, Guttmacher AE, editors. New England Journal of Medicine. 2010; 363: 166–176. https://doi.org/10.1056/NEJMra0905980 PMID: 20647212
3. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 2014; 42: D1001–1006. https://doi.org/10.1093/nar/gkt1229 PMID: 24316577
4. Cohen JC, Boerwinkle E, Mosley TH, Hobbs HH. Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. N Engl J Med. 2006; 354: 1264–1272. https://doi.org/10.1056/NEJMoa054013 PMID: 16554528
5. Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, Brown-Gentry K, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. Bioinformatics. 2010; 26: 1205–1210. https://doi.org/10.1093/bioinformatics/btq126 PMID: 20335276
6. Warner JL, Denny JC, Kreda DA, Alterovitz G. Seeing the forest through the trees: uncovering phenomic complexity through interactive network visualization. J Am Med Inform Assoc. 2015; 22: 324–329. https://doi.org/10.1136/amiajnl-2014-002965 PMID: 25336590
7. Arora S, Ge R, Moitra A. Learning Topic Models–Going Beyond SVD. Proceedings of the 2012 IEEE 53rd Annual Symposium on Foundations of Computer Science. Washington, DC, USA: IEEE Computer Society; 2012. pp. 1–10. https://doi.org/10.1109/FOCS.2012.49
8. Blei DM, Ng AY, Jordan MI. Latent Dirichlet Allocation. Journal of Machine Learning Research. 2003; 3: 993–1022.
9. MacMillan K, Wilson JD. Topic supervised non-negative matrix factorization. arXiv:170605084 [cs, stat]. 2017; Available: http://arxiv.org/abs/1706.05084
10. Blei DM. Probabilistic Topic Models. Commun ACM. 2012; 55: 77–84. https://doi.org/10.1145/2133806.2133826
11. Vosecky J, Jiang D, Leung KW-T, Ng W. Dynamic multi-faceted topic discovery in twitter. Proceedings of the 22nd ACM international conference on Conference on information & knowledge management. New York, NY, USA: ACM; 2013. pp. 879–884. https://doi.org/10.1145/2505515.2505593
12. Zhao WX, Jiang J, Weng J, He J, Lim E-P, Yan H, et al. Comparing Twitter and Traditional Media Using Topic Models. In: Clough P, Foley C, Gurrin C, Jones GJF, Kraaij W, Lee H, et al., editors. Advances in Information Retrieval. Springer Berlin Heidelberg; 2011. pp. 338–349.
13. Cha Y, Cho J. Social-network Analysis Using Topic Models. Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, NY, USA: ACM; 2012. pp. 565–574. https://doi.org/10.1145/2348283.2348360
14. Liu L, Tang L, Dong W, Yao S, Zhou W. An overview of topic modeling and its current applications in bioinformatics. Springerplus. 2016; 5. https://doi.org/10.1186/s40064-016-3252-8 PMID: 27652181
15. Wahabzada M, Mahlein A-K, Bauckhage C, Steiner U, Oerke E-C, Kersting K. Plant Phenotyping using Probabilistic Topic Models: Uncovering the Hyperspectral Language of Plants. Scientific Reports. 2016; 6: 22482. https://doi.org/10.1038/srep22482 PMID: 26957018
16. Backenroth D, He Z, Kiryluk K, Boeva V, Pethukova L, Khurana E, et al. FUN-LDA: A Latent Dirichlet Allocation Model for Predicting Tissue-Specific Functional Effects of Noncoding Variation: Methods and Applications. The American Journal of Human Genetics. 2018; 102: 920–942. https://doi.org/10.1016/j.ajhg.2018.03.026 PMID: 29727691

17. Huang Z, Dong W, Duan H. A probabilistic topic model for clinical risk stratification from electronic health records. Journal of Biomedical Informatics. 2015; 58: 28–36. https://doi.org/10.1016/j.jbi.2015.09.005 PMID: 26370451

18. Chan KR, Lou X, Karaletsos T, Crosbie C, Gardos S, Artz D, et al. An Empirical Analysis of Topic Modeling for Mining Cancer Clinical Notes. 2013 IEEE 13th International Conference on Data Mining Workshops. 2013. pp. 56–63. https://doi.org/10.1109/ICDMW.2013.91

19. McCoy TH, Castro VM, Snapper LA, Hart KL, Perlis RH. Efficient Genome-wide Association in Biobanks Using Topic Modeling Identifies Multiple Novel Disease Loci. Mol Med. 2017; 23: 285–294. https://doi.org/10.2119/molmed.2017.00100 PMID: 28861588

20. McCoy TH, Castro VM, Snapper L, Hart K, Januzzi JL, Huffman JC, et al. Polygenic loading for major depression is associated with specific medical comorbidity. Transl Psychiatry. 2017; 7: e1238. https://doi.org/10.1038/tp.2017.201 PMID: 28926002

21. Limestone: High-throughput candidate phenotype generation via tensor factorization—ScienceDirect [Internet]. [cited 1 Oct 2018]. Available: https://www.sciencedirect.com/science/article/pii/S1532046414001488

22. Pinoli P, Chicco D, Masseroli M. Enhanced probabilistic latent semantic analysis with weighting schemes to predict genomic annotations. 13th IEEE International Conference on BioInformatics and BioEngineering. 2013. pp. 1–4. https://doi.org/10.1109/BIBE.2013.6701702

23. Ye S, Dawson JA, Kendziorski C. Extending Information Retrieval Methods to Personalized Genomic-Based Studies of Disease. Cancer Inform. 2015; 13: 85–95. https://doi.org/10.4137/CIN.S16354 PMID: 25733795

24. Sra S, Dhillon IS. Generalized Nonnegative Matrix Approximations with Bregman Divergences. In: Weiss Y, Schölkopf B, Platt JC, editors. Advances in Neural Information Processing Systems 18. MIT Press; 2006. pp. 283–290. Available: http://papers.nips.cc/paper/2757-generalized-nonnegative-matrix-approximations-with-bregman-divergences.pdf

25. Kim H, Park H. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. Bioinformatics. 2007; 23: 1495–1502. https://doi.org/10.1093/bioinformatics/btm134 PMID: 17483501

26. Nordestgaard BG, Chapman MJ, Ray K, Borén J, Andreotti F, Watts GF, et al. Lipoprotein(a) as a cardiovascular risk factor: current status. Eur Heart J. 2010; 31: 2844–2853. https://doi.org/10.1093/eurheartj/ehq386 PMID: 20965889

27. Clarke R, Peden JF, Hopewell JC, Kyriakou T, Goel A, Heath SC, et al. Genetic Variants Associated with Lp(a) Lipoprotein Level and Coronary Disease. New England Journal of Medicine. 2009; 361: 2518–2528. https://doi.org/10.1056/NEJMoa0902604 PMID: 20032323

28. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. Nature Genetics. 2016; 48: 1284–1287. https://doi.org/10.1038/ng.3656 PMID: 27571263

29. Fuchsberger C, Abecasis GR, Hinds DA. minimac2: faster genotype imputation. Bioinformatics. 2015; 31: 782–784. https://doi.org/10.1093/bioinformatics/btu704 PMID: 25338720

30. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, et al. A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet. 2016; 48: 1279–1283. https://doi.org/10.1038/ng.3643 PMID: 27548312

31. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015; 526: 68–74. https://doi.org/10.1038/nature15393 PMID: 26432245

32. Martin PA, Thorburn MJ, Smith-Read EH. Chromosomal rearrangements in three generations of a Jamaican family. A possible further example of recombinational imbalance. Cytogenetics. 1970; 9: 360–368. PMID: 5501393

33. Wei W-Q, Bastarache LA, Carroll RJ, Marlo JE, Osterman TJ, Gamazon ER, et al. Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record. PLOS ONE. 2017; 12: 1–16. https://doi.org/10.1371/journal.pone.0175508 PMID: 28686612

34. O'Callaghan D, Greene D, Carthy J, Cunningham P. An analysis of the coherence of descriptors in topic modeling. Expert Systems with Applications. 2015; 42: 5645–5657. https://doi.org/10.1016/j.eswa.2015.02.055

35. Stevens K, Kegelmeyer P, Andrzejewski D, Buttler D. Exploring Topic Coherence over Many Models and Many Topics. Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Stroudsburg, PA, USA: Association for Computational Linguistics; 2012. pp. 952–961. Available: http://dl.acm.org/citation.cfm?id=2390948.2391052

36. Greene D, O'Callaghan D, Cunningham P. How Many Topics? Stability Analysis for Topic Models. In: Calders T, Esposito F, Hüllermeier E, Meo R, editors. Machine Learning and Knowledge Discovery in Databases. Springer Berlin Heidelberg; 2014. pp. 498–513.

37. Maaten L van der, Hinton G. Visualizing Data using t-SNE. Journal of Machine Learning Research. 2008; 9: 2579–2605.

38. t-SNE Initialization Options [Internet]. [cited 26 Sep 2018]. Available: https://jlmelville.github.io/smallvis/init.html

39. rs10455872—SNPedia [Internet]. [cited 23 May 2018]. Available: https://www.snpedia.com/index.php/Rs10455872

40. Khera AV, Emdin CA, Drake I, Natarajan P, Bick AG, Cook NR, et al. Genetic Risk, Adherence to a Healthy Lifestyle, and Coronary Disease. New England Journal of Medicine. 2016; 375: 2349–2358. https://doi.org/10.1056/NEJMoa1605086 PMID: 27959714

41. Barlera S, Specchia C, Farrall M, Chiodini BD, Franzosi MG, Rust S, et al. Multiple QTL influence the serum Lp(a) concentration: a genome-wide linkage screen in the PROCARDIS study. Eur J Hum Genet. 2007; 15: 221–227. https://doi.org/10.1038/sj.ejhg.5201732 PMID: 17133260

42. Berglund L, Ramakrishnan R. Lipoprotein(a): an elusive cardiovascular risk factor. Arterioscler Thromb Vasc Biol. 2004; 24: 2219–2226. https://doi.org/10.1161/01.ATV.0000144010.55563.63 PMID: 15345512

43. Sandholzer C, Hallman DM, Saha N, Sigurdsson G, Lackner C, Császár A, et al. Effects of the apolipoprotein(a) size polymorphism on the lipoprotein(a) concentration in 7 ethnic groups. Hum Genet. 1991; 86: 607–614. PMID: 2026424

44. Wei W-Q, Li X, Feng Q, Kubo M, Kullo IJ, Peissig PL, et al. LPA Variants are Associated with Residual Cardiovascular Risk in Patients Receiving Statins. Circulation. 2018; CIRCULATIONAHA.117.031356. https://doi.org/10.1161/CIRCULATIONAHA.117.031356 PMID: 29703846

45. Low Lipoprotein(a) Concentration Is Associated with Cancer and All-Cause Deaths: A Population-Based Cohort Study (The JMS Cohort Study) [Internet]. [cited 14 May 2018]. Available: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0031954

46. Lips EH, Gaborieau V, McKay JD, Chabrier A, Hung RJ, Boffetta P, et al. Association between a 15q25 gene variant, smoking quantity and tobacco-related cancers among 17 000 individuals. Int J Epidemiol. 2010; 39: 563–577. https://doi.org/10.1093/ije/dyp288 PMID: 19776245

47. Bioucas-Dias JM, Nascimento JMP. Estimation of signal subspace on hyperspectral data. Image and Signal Processing for Remote Sensing XI. International Society for Optics and Photonics; 2005. p. 59820L. https://doi.org/10.1117/12.620061

48. Tan VYF, Févotte C. Automatic Relevance Determination in Nonnegative Matrix Factorization with the /spl beta/-Divergence. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2013; 35: 1592–1605. https://doi.org/10.1109/TPAMI.2012.240 PMID: 23681989

49. Kanagal B, Sindhwani V. Rank Selection in Low-rank Matrix Approximations: A Study of Cross-Validation for NMFs. : 5.