MolecularMedicine

Efficient Genome-wide Association in Biobanks Using Topic Modeling Identifies Multiple Novel Disease Loci

Thomas H McCoy, Jr,[1] *Victor M Castro*,[1,2] *Leslie A Snapper*,[1] *Kamber L Hart*,[1] *and Roy H Perlis*[1]

[1]Center for Quantitative Health, Division of Clinical Research and Center for Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts, United States of America; and [2]Partners Research Information Systems and Computing, Partners HealthCare System, Boston, Massachusetts, United States of America

Biobanks and national registries represent a powerful tool for genomic discovery, but rely on diagnostic codes that can be unreliable and fail to capture relationships between related diagnoses. We developed an efficient means of conducting genome-wide association studies using combinations of diagnostic codes from electronic health records for 10,845 participants in a biobanking program at two large academic medical centers. Specifically, we applied latent Dirichilet allocation to fit 50 disease topics based on diagnostic codes, then conducted a genome-wide common-variant association for each topic. In sensitivity analysis, these results were contrasted with those obtained from traditional single-diagnosis phenome-wide association analysis, as well as those in which only a subset of diagnostic codes were included per topic. In meta-analysis across three biobank cohorts, we identified 23 disease-associated loci with $p < 1e\text{-}15$, including previously associated autoimmune disease loci. In all cases, observed significant associations were of greater magnitude than single phenome-wide diagnostic codes, and incorporation of less strongly loading diagnostic codes enhanced association. This strategy provides a more efficient means of identifying phenome-wide associations in biobanks with coded clinical data.

Online address: http://www.molmed.org
doi: 10.2119/molmed.2017.00100

INTRODUCTION

In the search for common genetic variations associated with medical disorders, the traditional analytic approach examines single disorders in case-control cohorts ascertained for a specific disorder. With the availability of large-scale biobanks with broad ascertainment, multiple approaches to phenome-wide association – ie, looking across a range of clinical phenotypes to detect genetic association – have been proposed (1). However, relying on individual disorders represented in diagnostic codes may not efficiently capture the underlying architecture of genetic risk. First, the ability of claims codes to accurately capture a given diagnosis varies widely, even when diagnosis-specific classifiers are applied to augment single codes (2,3). As such, approaches that focus on individual diagnostic codes are limited by inaccurate, missing or heterogeneous diagnoses; eg, where individuals with cystic fibrosis might be represented by male infertility, diabetes and chronic rhinosinusitis even in the absence of a diagnostic code for cystic fibrosis (4). Second, under conditions of pleiotropy, where a single variant contributes to risk for multiple disorders, as in some autoimmune and neuropsychiatric disorders, standard phenome-wide approaches do not make efficient use of the correlation structure between diagnoses. Finally, single-code approaches do not capture disease subtypes with different genetic architecture, where these subtypes may be reflected in different patterns of comorbidity, as a recent investigation of diabetes mellitus suggests (5–8).

Here, we describe a method for addressing the problem of mapping genetic space to high-dimensional phenotype space that leverages comorbidity and diagnostic uncertainty to allow efficient genome-wide or single-locus association. This approach facilitates association by capturing diagnostic co-occurrence patterns to reduce dimensionality, thereby decreasing the number of hypotheses being tested, while increasing power by including individuals who may have different manifestations of the same underlying pathology. Specifically, we apply latent Dirichilet allocation (LDA), a means of identifying commonly co-occurring features, to derive a set of 50 disease topics. Then we test those topics for association with common genetic variation and compare this approach to

Address correspondence to *Roy H Perlis, Massachusetts General Hospital Simches Research Building, Sixth Floor, Boston, MA 02114 USA. Phone: 617-726-7426; E-mail: rperlis@partners.org*

Submitted June 8, 2017; Accepted for Publication August 24, 2017;
Published Online (www.molmed.org) August 31, 2017.

Feinstein Institute
for Medical Research
Northwell Health™

standard methods using single International Classification of Diseases, Ninth Revision (ICD-9)/phenome-wide association studies (PheWAS) codes (9).

MATERIALS AND METHODS

Cohort Derivation and Genotyping

We drew on three cohorts of patients seen in the Brigham and Women's Hospital network and the Massachusetts General Hospital network, representing the first 15,064 individuals genotyped as part of the Partners HealthCare Biobank initiative (10). These individuals provided informed consent for their electronic health records (EHRs) to be examined in investigations approved by the Partners Institutional Review Board, and provided blood samples for DNA extraction.

DNA was extracted from buffy coat and genotyped using the Illumina Expanded Multi-Ethnic Genotyping Array (MEGA or MEGA-EX) platforms, with common variant arrays incorporating content from the 1000 Genomes Project Phase 3. Single nucleotide polymorphism (SNP) coordinates were remapped based on the TopGenomicSeq provided from Illumina (MEGA_Consortium_v2_15070954_A2.csv); all rsIDs correspond to build 142 of dbSNP. To determine the forward strand of the SNP, we aligned both SNP sequences (alleles A and B) to hg19 using BLAT with default parameters set by the University of California, Santa Cruz Genome Browser (11).

Quality Control and Imputation

Genotyping was done using three versions of the Illumina Multi-Ethnic Global (MEG) array (MEGA n = 4927, MEGA EX n = 5353, MEG n = 4784; mappable variants available for each were 1,411,334, 1,710,339 and 1,747,639, respectively). Each cohort was cleaned, imputed and analyzed separately to avoid batch effects. For each batch, we included subjects with genotyping call rates exceeding 99%; no related individuals based on identity by descent were included (12). From these individuals, any genotyped SNP with a call rate of at least 95%, minor allele frequency of 0.01 or greater and Hardy-Weinberg equilibrium p value $< 1 \times 10^{-6}$ was included. We then imputed using the Michigan Imputation Server implementing Minimac3 (13–15). Imputation used all population subsets from the 1000 Genomes Project Phase 3 v5 as reference panel; haplotype phasing was performed using SHAPEIT (16).

Ancestry

For each cohort, we used principal components analysis of linkage-disequilibrium-pruned genotyped SNPs to characterize population structure, based on EIGENSTRAT, as implemented in PLINK v1.9, and plotted these components with superimposition of HapMap samples to confirm locations of northern European individuals (17–19). Limiting the analysis to these individuals yielded 3,728 + 3,402 + 3,715 = 10,845 analyzable participants.

Topic Identification

For both cohorts, ICD-9 diagnosis codes extracted from each individual's medical record were grouped into 1,667 PheWAS codes corresponding to clinically meaningful disease categories, with the total number of codes within each PheWAS category preserved (20). Initial model fitting was performed using cohort 1 only, with cohort 2 preserved as an out-of-sample test set. To fit the topic model cohort, the PheWAS code count by subject matrix was frequency-controlled to eliminate PheWAS codes that occurred in < 1% or > 99% of subjects. After frequency control, 508 distinct PheWAS codes were used for the initial unsupervised learning step of model fit.

The PheWAS code count by subject matrix for cohort 1 was used to train an LDA model with 50 topics (9). LDA is a form of unsupervised machine learning typically found in natural language processing (NLP). As topic modeling is drawn from the NLP literature, this preprocessing can be conceptualized as treating each subject's medical record as a document composed of ICD-9 codes that are lemmatized to PheWAS codes and thereafter analyzed as a term-count document matrix. LDA postulates that the words of a document are a mixture of underlying topics, and documents are composed of each of these topics to varying degrees. The resulting trained LDA model is a distribution of all PheWAS codes over each topic. This distribution can be used to score each collection of PheWAS codes for membership in each of the topics. In the case of illness in a biobank, we use LDA to model biology as a collection of topics or underlying generator processes of observable, but potentially overlapping and incompletely penetrant, pathological states. These states are captured as PheWAS "words." Having trained the topic model on cohort 1, this model was then used to score each subject in cohort 1 for membership in each of the topics (in-sample) and each subject in cohort 2 for membership in each of the topics (out-of-sample). To perform the PheWAS LDA, we used the Gensim implementation of the LDA algorithm (21,22).

There is no widely accepted method for naming topics, since by definition all PheWAS words arise from all topics at some probability, albeit a vanishingly small probability in many cases. To aid in interpretability, in our discussion of results we name topics subjectively in terms of the preponderance of codes represented toward the top of the list, as interpreted by the two physician authors (THM, RHP); we refer to them below as "topic-name-plus," as a reminder to the reader that the topic contains more than just a single diagnosis and may contain apparently unrelated terms.

Analysis

Single-locus associations in each cohort were examined individually, and then combined in inverse variance–weighted fixed-effects meta-analyses. In these analyses, only bi-allelic SNPs with minor allele frequencies of at least 1% were retained. Tests for association used linear regression assuming an additive allelic effect, treating each topic as a quantitative trait and adjusting for the first 10 principal components *a priori* (analyses incorporating 5 or 20 components did not yield meaningfully different results). Association results are presented in terms of independent loci after pruning, using the clump command in PLINK 1.9, with a 250kb window and $r^2 = 0.2$. We present uncorrected p values, but elected to focus on p values less than 1e-15 and loci with at least two associated SNPs (23).

To facilitate comparisons across topics and methods, reported p values are not adjusted for linkage disequilibrium scores. Adjustment for lambda-1000 or linkage disequilibrium score regression intercept did not meaningfully change relative results; lambda values range from 0.990 to 1.017 λ across topics (24).

Secondary analyses examined alternate topic-based phenotypes in which either the most strongly loading diagnostic codes (ie, those with loading > 0.05) or least strongly loading diagnostic codes (ie, those with loading < 0.01) for a given topic were omitted, as a means of understanding the relative contributions of these sets of codes. These analyses utilized the same approach as for the primary analysis of topics. For comparison, we also examined association with the presence or absence of the single most strongly loading diagnostic code in each topic, using logistic regression.

All supplementary materials are available online at www.molmed.org.

RESULTS

After exclusions for genotyping quality control, relatedness and ancestry, cohorts 1, 2 and 3 included 2,141/3,728 (57.4%), 1,690/3,402 (49.7%) and 2,089/3,715 (56.23%) female participants, respectively. Mean ages were 57.9 (standard deviation [SD] 16.2), 62.4 (SD 16.0) and 59.4 (SD 16.5).

After imputation, a total of 7,781,941 SNPs with minor allele frequency (MAF) of 0.01 or greater were analyzed for each of the 50 topics and meta-analyzed. After genome-wide association analysis for each of the 50 topics, a total of 56 loci spanning 24 topics included at least one SNP with $p < $ 1e-11; 39 of these loci across 22 topics included at least one additional associated SNP at $p < 0.01$. Table 1 reports the physical position, annotation and association for the most strongly associated SNPs for topics with at least one $p < $ 5e-15 and at least two associated SNPs in a given locus, while Supplementary Table S1 reports the 10 most associated independent SNPs for all 50 topics (for effects by cohort, see Supplementary Table S2). The strongest associations (all $p < $ 1e-15) were observed for pulmonary disease/cystic fibrosis-plus, anemia and fracture-plus, rheumatoid arthritis-plus, pregnancy complications-plus, uterine neoplasm-plus, viral-plus, neoplasm-plus, adrenal and electrolyte disorders-plus, and pituitary and adrenal disorder-plus. Figures 1 and 2 show Manhattan and locus plots for pulmonary disease/cystic fibrosis-plus and neoplasm-plus; for plots for the remainder of these, see Supplementary Materials. Diagnostic codes loading most strongly for each of these topics are listed in Table 2; for the codes loading on all 50 topics, see Supplementary Table S3.

We also examined (Table 3) the effect of three alternate phenotypic definitions: examining the topic "tail" only (ie, diagnostic codes with weights < 0.05, the "tail" of the list) or the topic "head" only (ie, diagnostic codes with weights > 0.01, the "head" of the list) and including only the single top-weighted diagnostic code (ie, a standard single diagnosis association). This last comparison allows direct contrast with nominal associations returned by traditional PheWAS, recognizing that here only 50 phenotypes are examined rather than 500 or more.

DISCUSSION

We applied a topic-modeling approach to identify 50 groups of diagnostic codes in biobank-associated EHR data and then used genome-wide data to examine common-variant associations for each topic. With this novel approach, we identified multiple known loci for autoimmune and pulmonary disease, as well as multiple apparently novel disease loci for pregnancy complications, viral susceptibility, anemia/fracture risk and uterine cancer not previously associated at a genome-wide threshold with disease (based on searching the National Human Genome Research Institute–European Bioinformatics Institute Catalog of published genome-wide association studies) (25). We compared our results to those arising from a standard single-diagnostic-code PheWAS; this approach would not have yielded association at this threshold. Moreover, omitting either the head or the tail of each topic (ie, the most- or least-weighted diagnosis) eliminates the association, suggesting that the observed effect does not arise from a small number of codes.

The identification of robust associations with loci implicated in prior genome-wide association studies demonstrates convergent validity (27,28). We demonstrate that this approach more efficiently detects these known associations (based on magnitude of p value) than single-code association. That is, simply incorporating a single ICD9/PheWAS code yielded weaker evidence of association. These loci and the sensitivity analysis associated with topic pruning (head and tail distributions; Table 1, Figure 1) function as positive controls and illustrations of assay sensitivity.

In nearly all cases, we note that the strongest associations are identified by incorporating all codes loading on a topic, rather than limiting the analysis to only the most strongly loading. Indeed, we observe that omitting such strongly loaded codes does not necessarily reduce the magnitude of association. Interestingly, there is only one example where a single code yielded an association nearly as robust as that observed with the topic, rheumatoid arthritis-plus, which may reflect the distinct genetic architecture of this disorder compared with some others.

In lieu of looking across phenotypes, a recent report describes a method to identify disease subtypes based upon network analysis (29). Our approach is

Table 1. Loci associated with a topic with $p < 1e\text{-}15$

Topic #	Topic name	CHR	SNP	*p* value	A1	A2	MAF	SNPs in locus [A]	Locus range	Genes in locus [B]	GWAS catalog [C]
19	Pulmonary disease and cystic fibrosis-plus	7	7:117277554	1.049E-42	T	A	0.023	25	chr7:117041941..117357607	(*ASZ1*, *CFTR*, *CTTNBP2*)	Cystic fibrosis; esophageal cancer (PMID: 27527254)
19	Pulmonary disease and cystic fibrosis-plus	7	7:116967838	3.005E-32	A	C	0.026	18	chr7:116806103..117217725	(*ASZ1*, *CFTR*, *ST7*, *ST7-OT3*, *WNT2*)	Cystic fibrosis; esophageal cancer (PMID: 27527254)
09	Anemia and fracture-plus	4	4:12459496	5.985E-29	G	C	0.010	10	chr4:12459496..12585393	()	None
24	Rheumatoid arthritis-plus	6	6:32570417	2.14E-28	C	A	0.167	759	chr6:32326531..32685550	(HLA region)	Rheumatoid arthritis
16	Pregnancy complications-plus	5	5:32198317	2.852E-20	C	T	0.010	4	chr5:32006737..32198317	(*GOLPH3*, *PDZD2*)	Myocardial infarction (PMID: 26708285)
19	Pulmonary disease and cystic fibrosis-plus	7	7:117220403	2.055E-19	G	A	0.108	69	chr7:116971883..117223764	(*ASZ1*, *CFTR*)	Cystic fibrosis; esophageal cancer (PMID: 27527254)
22	Uterine neoplasm-plus	14	14:62132727	3.294E-19	A	G	0.018	11	chr14:62109231..62160023	(*FLJ22447*, *HIF1A-AS1*)	None
07	Viral-plus	14	14:57647875	7.58E-19	G	T	0.010	2	chr14:57647875..57649586	()	None
22	Uterine neoplasm-plus	6	6:148020784	7.105E-18	G	A	0.010	54	chr6:147998800..148177349	()	None
24	Rheumatoid arthritis-plus	6	6:32681992	9.832E-18	C	T	0.202	211	chr6:32433759..32682043	(HLA region)	Rheumatoid arthritis
40	Neoplasm-plus	3	3:1145301	1.962E-17	A	G	0.016	10	chr3:1131967..1145857	(*CNTN6*)	Gut microbiome (PMID: 27723756)
22	Uterine neoplasm-plus	2	2:209079758	7.645E-17	C	G	0.017	4	chr2:209077907..209084036	()	None
01	Adrenal and electrolyte-plus	4	4:41059660	1.813E-16	A	C	0.012	3	chr4:40970912..41064140	(*APBB2*)	None
22	Uterine neoplasm-plus	4	4:84272944	1.993E-16	G	T	0.015	2	chr4:84272944..84273025	()	None
25	Pituitary and adrenal disease-plus	18	18:25741737	3.269E-16	C	A	0.012	5	chr18:25724075..25753257	(*CDH2*)	Resting heart rate (PMID: 27798624)
24	Rheumatoid arthritis-plus	6	6:32303848	3.558E-16	A	G	0.204	263	chr6:32114515..32389305	(HLA region)	Rheumatoid arthritis
07	Viral-plus	8	8:58987251	4.165E-16	T	C	0.012	5	chr8:58987251..59037450	(*FAM110B*)	None

[A] Total number of SNPs within $r^2 = 0.2$ and 250 kb with $p < 0.01$

[B] Annotated mRNA in locus

[C] Genes annotated in GWAS catalog

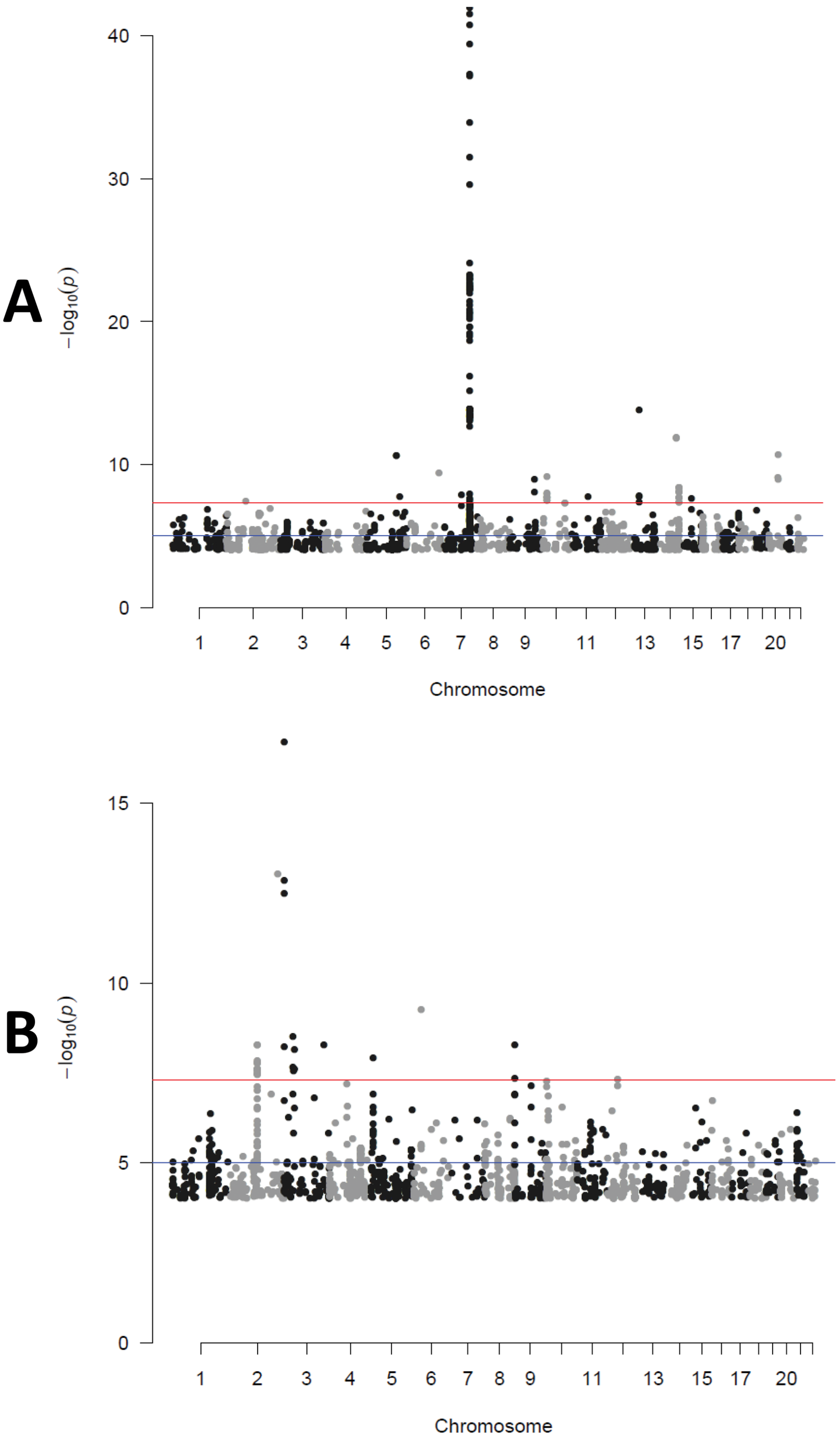

Figure 1. Manhattan plots for two example topics: (A) pulmonary disease/cystic fibrosis-plus and (B) neoplasm-plus.

not directly comparable, but may also be valuable in identifying subtypes in the case where a given diagnosis is genetically heterogeneous but the presence of comorbidities helps to define more homogeneous groups. A major advantage of the present approach compared with other unsupervised methods (eg, deep learning) is inspectability: it yields a weighted list of diagnostic codes. This inspectability enhances biological utility, as it allows *post hoc* clarification of the results, as illustrated by the sensitivity analysis. While we describe its application for genomics, it may be useful for other approaches drawing on coded EHR data where diagnostic codes do not definitively identify a diagnosis or subtype. Notably, it should also be possible to further extend the utility of our approach by incorporating additional coded or uncoded data – concepts extracted by NLP, for example – where such data are available.

The gain in statistical power afforded by this approach is apparent. For a genotypic risk ratio of 1.5 with a minor allele frequency of 25% and a disease prevalence of 5%, nearly 80,000 cases are required to achieve 80% power after Bonferroni correction for 500 PheWAS phenotypes, versus nearly 800 cases with 50 topics, if each is analyzed as a dichotomous outcome. In reality, the increased case reliability that arises from integrating across related codes likely renders these estimates conservative, in some cases markedly so. Empirically, our results show that in no case would a single-code association have yielded stronger nominal association, independent of Bonferroni correction, than the topic-based association, and in most cases the association was markedly less.

Still, we note several important limitations. From a modern genomics perspective, the present cohorts are likely insufficient to robustly detect all but the largest associations. They are, however, large enough to demonstrate the feasibility and efficiency of using aggregated groups of diagnoses as an efficient complement to phenome-wide association. Additionally, the biobank cohorts

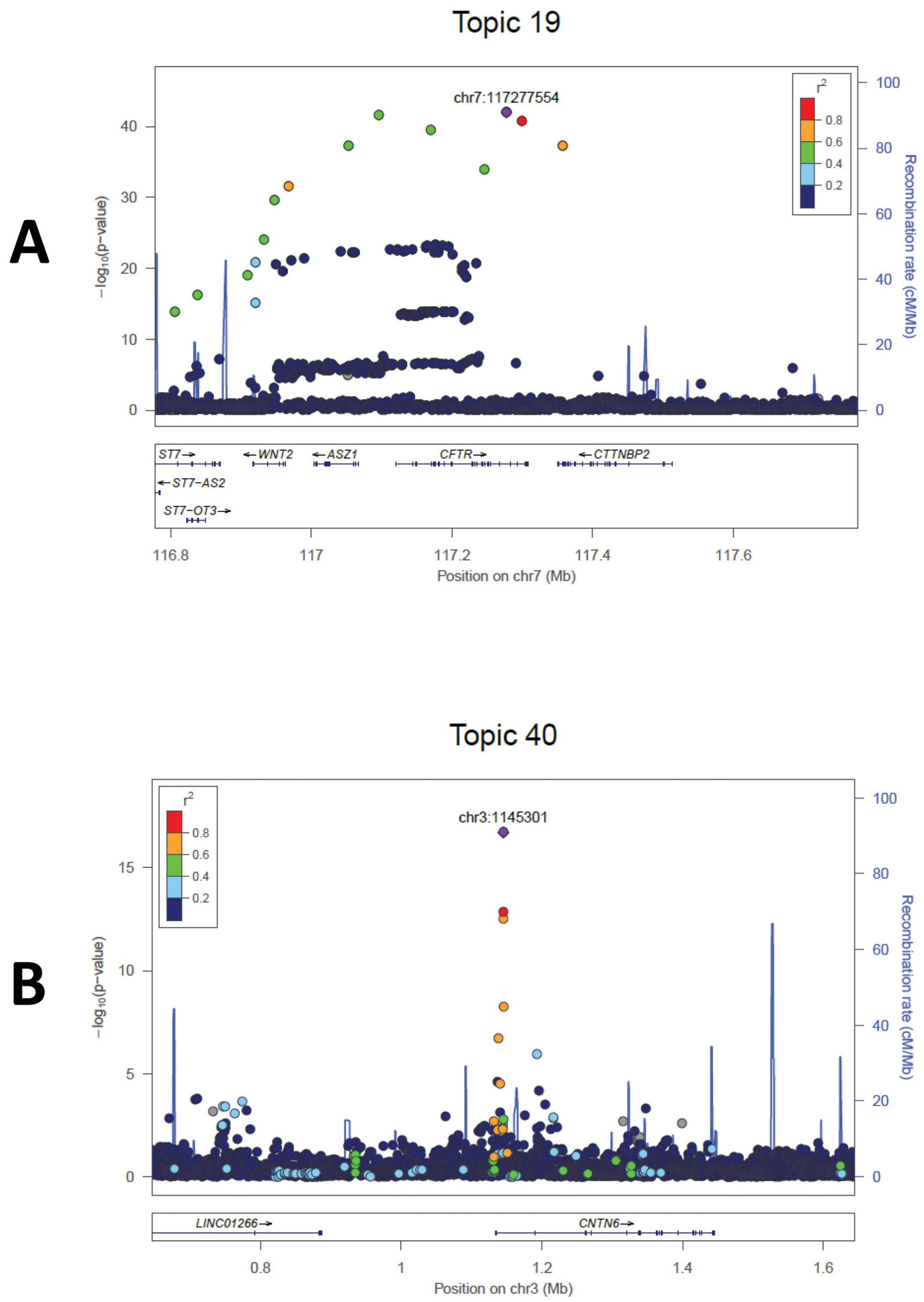

Figure 2. Locus plots for two example topics: (A) pulmonary disease/cystic fibrosis-plus and (B) neoplasm-plus.

studied here are insufficient to examine these associations in non–northern European populations; replication in other populations would be informative.

We note that the topics in many cases include codes seemingly unrelated to the predominant diagnosis; while these may represent type I error or noise, they may also illustrate the power of topic modeling to detect co-occurring diagnoses where the physiologic relationship is not otherwise recognized.

Table 2. Diagnostic codes loading on topics associated with $p < 1e$-

Topic 01	Topic 07	Topic 09	Topic 16	Topic 19	Topic 22	Topic 24	Topic 25	Topic 40
Disorders of adrenal glands	Psoriasis and related disorders	Other anemias	Erythematous conditions	Chronic airway obstruction	Uterine cancer	Rheumatoid arthritis and related inflammatory polyarthropathies	Disorders of the pituitary gland and its hypothalamic control	Secondary malignant neoplasm
Fracture of ankle and foot	Human immunodeficiency virus	Fracture of upper limb	Other conditions of the mother complicating pregnancy	Diseases of pancreas	Benign neoplasm of other parts of digestive system	Inflammatory spondylopathies	Benign neoplasm of other endocrine glands	Colorectal cancer
Disorders of fluid electrolyte and acid-base balance	Viral infection	Disorders of mineral metabolism	Indications for care or intervention related to labor and delivery (not elsewhere classified)	Cystic fibrosis	Endometrial hyperplasia		Radiotherapy	Cancer of other female genital organs
Symptoms involving skin and other integumentary tissue	Disorders of the autonomic nervous system	Iron deficiency anemias not otherwise specified	Known or suspected fetal abnormality	Pneumonia	Benign neoplasm of colon		Neoplasm of uncertain behavior	Chemotherapy
Pernicious or B12 deficiency anemia	Viral warts and human papillomavirus	Myeloproliferative disease	Hypertension complicating pregnancy	Tobacco use disorder	Cancer of other female genital organs		Disorders of adrenal glands	Radiotherapy
Other disorders of the kidney and ureters	Tobacco use disorder	Gastrointestinal hemorrhage	Disorders of menstruation	Other diseases of lung	Diseases of esophagus		Cancer of other endocrine glands	Cancer of the digestive organs and peritoneum
Urinary tract infection	Hereditary and idiopathic peripheral neuropathy	Malaise and fatigue	Other high-risk pregnancy	Abdominal pain	Nausea and vomiting		Effects of radiation not otherwise specified	Cancer suspected or other
Abdominal pain		Diseases of blood and blood-forming organs	Early or threatened labor; hemorrhage in early pregnancy	Osteoporosis osteopenia and pathological fractures	Postoperative infection		Testicular dysfunction	Pancreatic cancer

Continued on next page

Table 2. *Continued.*

Topic 01	Topic 07	Topic 09	Topic 16	Topic 19	Topic 22	Topic 24	Topic 25	Topic 40
Genitourinary congenital anomalies		Chromosomal anomalies and genetic disorders	Parkinson's disease		Other disorders of the kidney and ureters		Benign neoplasm of brain and other parts of nervous system	Diseases of blood and blood-forming organs
Disorders of function of stomach		Pernicious or B12-deficiency anemia	Miscarriage; stillbirth		Neoplasm of unspecified nature of digestive system		Hypothyroidism	Neoplasm of unspecified nature of digestive system
Musculoskeletal symptoms referable to limbs		Fracture of hand or wrist	Other complications of pregnancy (not elsewhere classified)		Other disorders of intestine		Other headache syndromes	Disorders of fluid electrolyte and acid-base balance
Hydronephrosis		Pneumonia	Problems associated with amniotic cavity and membranes					
Urinary calculus		Immune disorders	Abnormality pelvic soft tissues and organs complicating pregnancy					
Anxiety phobic and dissociative disorders		Disorders of fluid electrolyte and acid-base balance	Symptoms affecting skin					
Vitamin deficiency			Hemorrhage during pregnancy, childbirth and postpartum					
Tobacco use disorder								

Table 3. Sensitivity analysis examining alternate phenotypes omitting the most- or least-strongly loading diagnoses

Topic #	Topic name	CHR	SNP	*P* value	*P* values in sensitivity analysis		
				Full topic *(all diagnoses)*	Topic head *(diagnoses loading ≥0.01)*	Topic tail *(diagnoses loading ≤0.05)*	Single diagnosis *(only top diagnosis)*
19	Pulmonary disease and cystic fibrosis-plus	7	7:117277554	1.049E-42	1.121E-4	7.178E-11	0.02882
19	Pulmonary disease and cystic fibrosis-plus	7	7:116967838	3.005E-32	1.686E-3	4.765E-3	0.0518
09	Anemia and fracture-plus	4	4:12459496	5.985E-29	0.3539	6.102E-51	*
24	Rheumatoid arthritis-plus	6	6:32570417	2.14E-28	2.895E-20	0.2878	7.586E-19
16	Pregnancy complications-plus	5	5:32198317	2.852E-20	0.1513	6.545E-44	*
19	Pulmonary disease and cystic fibrosis-plus	7	7:117220403	2.055E-19	4.268E-4	1.114E-08	0.05133
22	Uterine neoplasm-plus	14	14:62132727	3.294E-19	0.2401	0.7799	6.231E-3
07	Viral-plus	14	14:57647875	7.58E-19	0.3954	0.3554	*
22	Uterine neoplasm-plus	6	6:148020784	7.105E-18	0.5184	3.387E-3	5.874E-3
24	Rheumatoid arthritis-plus	6	6:32681992	9.832E-18	4.413E-14	0.2	4.409E-11
40	Neoplasm-plus	3	3:1145301	1.962E-17	9.212E-4	3.945E-6	0.01377
22	Uterine neoplasm-plus	2	2:209079758	7.645E-17	0.1993	0.8465	*
01	Adrenal and electrolyte-plus	4	4:41059660	1.813E-16	0.01646	0.1058	*
22	Uterine neoplasm-plus	4	4:84272944	1.993E-16	0.02548	0.8802	*
25	Pituitary and adrenal disease-plus	18	18:25741737	3.269E-16	0.06459	0.078	0.07633
24	Rheumatoid arthritis-plus	6	6:32303848	3.558E-16	5.129E-14	0.118	3.344E-13
07	Viral-plus	8	8:58987251	4.165E-16	0.01557	0.8416	0.09888

*Sparseness of single diagnosis code and low MAF precludes estimate.

Our analysis of topic "heads" and "tails" suggests that, in most cases, topics are not well captured by a single code or small number of codes. As such, the names we apply represent a best guess at interpretation, and investigation of the mechanism of overlap (*in vivo* or *in silico*) represents an important next step. In particular, consideration of orthogonal biological data, such as investigating pathways or expression quantitative trait loci, could further clarify the way in which groups of associated diagnoses relate to one another mechanistically.

CONCLUSION

In sum, our results indicate the utility of an approach to large-scale biobank data that aggregates over groups of diagnostic codes by treating groups of codes as relating to underlying topics. This approach is superior to single-code association for diagnoses with shared liability or groups of diagnostic codes that more reliably identify an underlying phenotype. It identifies multiple apparently novel disease loci while replicating existing associations, and suggests multiple other regions as well as phenotypes that merit further investigation in biobank cohorts or registries.

ACKNOWLEDGMENTS

This study was funded by the National Human Genome Research Institute and the National Institute of Mental Health. The authors acknowledge the Partners HealthCare Biobank for providing genomic data and health information data.

DISCLOSURE

RHP has served on advisory boards for, or provided consulting to, Genomind, Healthrageous, Perfect Health, Pfizer, Psy Therapeutics, and RIDVentures.

REFERENCES

1. Antony A, *et al.* (2004) Translational upregulation of folate receptors is mediated by homocysteine via RNA-heterogeneous nuclear ribonucleoprotein E1 interactions. *J. Clin. Invest.* 113:285–301.
2. Perlis R, *et al.* (2012) Using electronic medical records to enable large-scale studies in psychiatry: treatment resistant depression as a model. *Psychol. Med*. 42:41–50.
3. Castro VM, *et al.* (2015) Validation of electronic health record phenotyping of bipolar disorder cases and controls. *Am. J. Psychiatr*. 172:363–72.
4. Hallberg P, Sjöblom V. (2005) The use of selective serotonin reuptake inhibitors during pregnancy and breast-feeding: a review and clinical aspects. *J. Clin. Psychopharmacol*. 25:59–73.
5. American Psychiatric Association. (2010) *Practice guidelines for the treatment of major depression*. Washington, DC: American Psychiatric Press.
6. Cross-Disorder Group of the Psychiatric Genomics Consortium, *et al.* (2013) Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nat. Genet.* 45:984–94.
7. Cho JH, Feldman M. (2015) Heterogeneity of autoimmune diseases: pathophysiologic insights from genetics and implications for new therapies. *Nat. Med*. 21:730–38.
8. Elkin I, *et al.* (1995) Initial severity and differential treatment outcome in the National Institute of Mental Health Treatment of Depression Collaborative Research Program. *J. Consult. Clin. Psychol*. 63:841.
9. Blei DM, Ng AY, Jordan MI. (2003) Latent dirichlet allocation. *J. Mach. Learn. Res*. 3:993–1022.
10. Gainer VS, *et al.* (2016) The biobank portal for Partners Personalized Medicine: a query tool for working with consented biobank samples, genotypes, and phenotypes using i2b2. *J. Pers. Med*. 6:11.
11. Kent WJ. BLAT—the BLAST-like alignment tool. *Genome Res* 2002;12:656–64. doi:10.1101/gr.229202. Article published online before March 2002.
12. Henn BM, *et al.* (2012) Cryptic distant relatives are common in both isolated and cosmopolitan genetic samples. *PLoS One*. 7(4): e34267.

13. Fuchsberger C, Abecasis GR, Hinds DA. (2015) minimac2: faster genotype imputation. *Bioinformatics*. 31:782–84.
14. Minimac3. Available from http://genome.sph.umich.edu/wiki/Minimac3. Accessed May 1, 2017.
15. Michigan imputation server. Available from https://imputationserver.sph.umich.edu. Accessed May 1, 2017.
16. Delaneau O, Marchini J, Zagury JF. (2012) A linear complexity phasing method for thousands of genomes. *Nat. Methods*. 9:179–81.
17. Price AL, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 38:904–09.
18. Chang CC, et al. (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*. 4:1.
19. Purcell S, *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81:559–75.
20. Denny JC, *et al.* (2010) PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics.* 26:1205–10.
21. Rehurek R, Sojka P. (2010) Software framework for topic modelling with large corpora In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. 45–50.
22. Hoffman M, Bach FR, Blei DM. (2010) Online learning for latent Dirichlet allocation. In *Advances in Neural Information Processing systems*. Cambridge, MA: MIT Press; 856–64.
23. S Purcell, C Chang. (2015) PLINK 1.9. Available from www.cog-genomics.org/plink2. Downloaded May 15, 2017.
24. Bulik-Sullivan BK, *et al.* (2015) LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* 47:291–95.
25. Prom-Wormley E, *et al.* (2015) Genetic and environmental contributions to the relationships between brain structure and average lifetime cigarette use. *Behav. Genet.* 45:157–70.
26. Gene Page online. Cambridge, MA: Broad Institute. Available from www.gtexportal.org/home/gene/CLPX. Accessed March 8, 2017.
27. Plenge RM, *et al.* (2017) TRAF1-C5 as a risk locus for rheumatoid arthritis – a genomewide study. *N. Engl. J. Med.* 357:1199–1209.
28. International Multiple Sclerosis Genetics Consortium. (2007) Risk alleles for multiple sclerosis identified by a genomewide study. *N. Engl. J. Med.* 357:851–62.
29. Pavlova B, Perlis RH, Alda M, Uher R. (2015) Lifetime prevalence of anxiety disorders in people with bipolar disorder: a systematic review and meta-analysis. *Lancet Psychiatr.* 2:710–17.

Cite this article as: McCoy TH, *et al.* (2017) Efficient genome-wide association in biobanks using topic modeling identifies multiple novel disease loci. *Mol. Med.* 23:285–94.