A comprehensive survey of genetic variation in 20,691 subjects from four large cohorts

Sara Lindström1,2,3, Stephanie Loomis4, Constance Turman1,2, Hongyan Huang1,2, Jinyan Huang1,2, Hugues Aschard1,2, Andrew T. Chan5, Hyon Choi5, Marilyn Cornelis6, Gary Curhan7,8,9, Immaculata De Vivo1,2,8, A. Heather Eliassen2,8, Charles Fuchs8,10, Michael Gaziano11, Susan E. Hankinson2,8,12, Frank Hu2,13, Majken Jensen8,13, Jae H. Kang8, Christopher Kabrabel8,14, Liming Liang1,2,15, Louis R. Pasquale6,8, Eric Rimm2,8,13, Meir J. Stampfer2,8,13, Rulla M. Tamimi2,8, Shelley S. Tworoger2,8, Janey L. Wiggs4, David J. Hunter1,2,8,13, Peter Kraft1,2,15

1 Program in Genetic Epidemiology and Statistical Genetics, Harvard T.H. Chan School of Public Health, Boston, MA, United States of America, 2 Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, United States of America, 3 Department of Epidemiology, University of Washington, Seattle, WA, United States of America, 4 Department of Ophthalmology, Harvard Medical School, Massachusetts Eye and Ear Infirmary, Boston, MA, United States of America, 5 Gastrointestinal Unit, Massachusetts General Hospital, Boston, MA, United States of America, 6 Section of Rheumatology and Clinical Epidemiology Unit, Boston University School of Medicine, Boston, MA, United States of America, 7 Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL, United States of America, 8 Channing Division of Network Medicine, Brigham and Women’s Hospital and Harvard Medical School, Boston, MA, United States of America, 9 Renal Division, Department of Medicine, Brigham and Women’s Hospital, Boston, MA, United States of America, 10 Department of Medical Oncology, Dana-Farber Cancer Institute and Harvard Medical School, Boston, MA, United States of America, 11 Division of Aging, Department of Medicine, Brigham and Women’s Hospital, Boston, MA, United States of America, 12 Department of Biostatistics and Epidemiology, University of Massachusetts, Amherst, MA, United States of America, 13 Department of Nutrition, Harvard T.H. Chan School of Public Health, Boston, MA, United States of America, 14 Department of Emergency Medicine, Center for Vascular Emergencies, Massachusetts General Hospital, Harvard Medical School, Boston, MA, United States of America, 15 Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, United States of America.

* saralind@uw.edu

Abstract

The Nurses’ Health Study (NHS), Nurses’ Health Study II (NHSII), Health Professionals Follow Up Study (HPFS) and the Physicians Health Study (PHS) have collected detailed longitudinal data on multiple exposures and traits for approximately 310,000 study participants over the last 35 years. Over 160,000 study participants across the cohorts have donated a DNA sample and to date, 20,691 subjects have been genotyped as part of genome-wide association studies (GWAS) of twelve primary outcomes. However, these studies utilized six different GWAS arrays making it difficult to conduct analyses of secondary phenotypes or share controls across studies. To allow for secondary analyses of these data, we have created three new datasets merged by platform family and performed imputation using a common reference panel, the 1,000 Genomes Phase I release. Here, we describe the methodology behind the data merging and imputation and present imputation quality statistics and association results from two GWAS of secondary phenotypes (body mass index (BMI) and venous thromboembolism (VTE)). We observed the strongest BMI association for the FTO SNP rs55872725 (β = 0.45, p = 3.48x10^{-22}), and using a significance level of p = 0.05,
we replicated 19 out of 32 known BMI SNPs. For VTE, we observed the strongest association for the rs2040445 SNP (OR = 2.17, 95% CI: 1.79–2.63, p = 2.70 x 10^{-15}), located downstream of F5 and also observed significant associations for the known ABO and F11 regions. This pooled resource can be used to maximize power in GWAS of phenotypes collected across the cohorts and for studying gene-environment interactions as well as rare phenotypes and genotypes.

### Introduction

Large, well-phenotyped cohort studies have constituted the backbone of epidemiology for several decades. Prospectively collected longitudinal information on exposures and outcomes enables a broad spectrum of analyses and has led to novel insights into disease etiology, such as the link between smoking and lung cancer [1,2] as well as the link between both high cholesterol levels and trans fatty acids with coronary heart disease [3,4]. Many existing cohorts collect biological specimens from their participants, allowing for studies of inherited genetic variation as well as prospectively measured biomarkers such as metabolomic profiles [5] and circulating hormone levels [6]. Genome-wide association studies (GWAS) are currently a main engine of genetic epidemiology and have led to the identification of thousands of loci for hundreds of traits (for an overview and its clinical applications, see Manolio [7]). When designing a GWAS, cost is still the determining factor and consequently, GWAS within cohorts are often conducted within nested case-control studies or sub-cohorts. In contrast, the Women’s Genome Health Study (WGHS) [8] genotyped the entire cohort of 27,000 women and the Genetic Epidemiology Research on Adult Health and Aging (GERA) Cohort has generated GWAS data on almost 100,000 individuals [9]. However, in many instances, GWAS are tied to specific funding sources acquired for studying a pre-defined outcome and only a small fraction of the cohort is genotyped at a specific time.

Within the Nurses’ Health Study (NHS) [10], Nurses’ Health Study II (NHSII) [11], Health Professional Follow Up Study (HPFS) [12] and the Physicians’ Health Study (PHS) [13], since 2007, we have conducted twelve GWAS of different traits including type 2 diabetes [14], coronary heart disease [15], several cancer types [16–19], and mammographic density [20,21]. In total, we have assembled GWAS data for 20,769 individuals across the cohorts, creating unprecedented opportunities to conduct secondary analyses on other collected outcomes. Indeed, we have used one or many of these GWAS to analyze secondary phenotypes including but not limited to body anthropometrics [22–24], hair color [25], reproductive aging [26], smoking behavior [27], telomere length [28], mammographic density [29], cutaneous nevi [30], melanoma [30], depressive symptoms [31], coffee consumption [32] as well as circulating levels of B12 [33], folate [34], hormones [35], vitamins [36,37], retinol [38] and e-selectin [39]. However, GWAS of secondary traits face practical issues in terms of different genotyping arrays, low variability in the phenotype of interest within a single GWAS (e.g. rare diseases where only a handful of cases may occur in the original GWAS), and theoretical issues including ascertainment bias due to oversampling of cases [40] or differential genotype/imputation quality between studies [41] (e.g. if controls are “utilized” from GWAS data generated on a different genotype platform).

Here, we describe our pipeline for merging and imputing the individual GWAS datasets within NHS, NHSII, HPFS and PHS. Datasets were merged based on genotype platform family...
and all data were subsequently imputed to a common reference panel (the 1,000 Genomes Phase I release [42]). We present proof-of-principle results from genome-wide analysis of body mass index (BMI) and venous thromboembolism (VTE).

Materials and methods
Description of NHS, NHSII, HPFS and PHS
In 1976, the Nurses’ Health Study (NHS) was launched with the goal of studying women’s health [10]. Since that time, 121,700 nurse participants have answered biennial questionnaires (response rate >90% over time) about personal and physical characteristics, physical activity and ability, reproductive history, family history of disease, environmental/personal exposures, diet and dietary supplements, screening, disease and health conditions, prescription and over-the-counter medications, and psychosocial history. In addition, 32,826 blood and 29,684 cheek cell samples have been collected since the late 1980s. An additional 116,430 nurses were recruited in 1989 as a part of Nurses’ Health Study II (NHSII) and have returned biennial questionnaires similar to those used for NHS [11]. For NHSII, we have collected blood samples for 29,612 women and cheek cell samples for an additional 29,859 women. The Health Professional Follow-Up Study (HPFS) began in 1986 with the aim of studying men’s health [12]. A total of 51,529 men in health professions were recruited, and every two years, members of the study receive questionnaires similar to the ones used in NHS. In HPFS, we have collected blood samples from 18,159 participants and cheek cell samples from an additional 13,956 men. The Physicians’ Health Study (PHS) is a randomized primary prevention trial of aspirin and supplements among 29,067 United States physicians followed with annual questionnaires since 1982 [13]. A total of 14,916 men provided a baseline blood sample.

Ethics statement
Each GWAS study was approved by the Brigham and Women’s Hospital Institutional Review Board. Return of the mailed self-administered questionnaires was voluntary. Thus, receipt of a completed questionnaire was considered as evidence of a desire to participate in the study and was taken as a formal indication of consent.

Description of GWAS studies and genotyping
Since 2007, twelve separate GWAS have been conducted within these four cohorts (Table 1). The primary traits are breast cancer [16], pancreatic cancer [43], glaucoma [44], endometrial cancer [17], colon cancer [19], glioma [45], prostate cancer [18], type 2 diabetes [14], coronary heart disease [15], kidney stones, gout and mammographic density [20]. These studies were genotyped on six different arrays (Table 1) at four different genotyping centers (National Cancer Institute, Broad Institute, University of Southern California and Rosetta/Merck). Standard quality control filters for call rate, Hardy-Weinberg equilibrium, and other measures were applied to the genotyped SNPs and/or samples. In total, these GWAS data sets comprise 20,769 participants including 11,522 from NHS, 934 subjects from NHSII, 7,018 subjects from HPFS and 1,305 subjects from PHS.

Dataset merging
Successfully merging genotype data for different individuals requires complete overlap in SNPs. SNPs that are missing by design (due to different genotyping platforms) from some studies will be correlated with the primary phenotype for that dataset. This might cause spurious results in any secondary analysis on related traits. Although a missing SNP can be
imputed, it will have a higher degree of inaccuracy in imputed compared with genotyped SNPs, potentially creating differential measurement error that could also lead to bias [41,46,47]. Therefore, we first looked at the overlap of SNPs between different genotyping arrays and identified three broad platform families with high degree of overlap within category.
but low overlap across categories—the earlier generation of Illumina arrays (HumanHap), the Illumina OmniExpress array and Affymetrix 6.0 array. The HumanHap platform had a total of 459,999 SNPs compared with 565,810 SNPs for OmniExpress and 668,283 SNPs for Affymetrix 6.0. However, the intersection among all three platform families was only 75,285 SNPs (Fig 1). To achieve the largest GWAS datasets as possible without losing SNP information, we created three datasets—HumanHap comprising six GWAS datasets, OmniExpress comprising four GWAS datasets and Affymetrix 6.0 comprising two GWAS datasets. In the merging process, we removed any SNPs that were not in all studies for a specific platform or had a missing

Fig 1. Overlap in SNPs across genotype platforms.
https://doi.org/10.1371/journal.pone.0173997.g001
We flipped strands where appropriate and removed A/T and C/G SNPs to create the final compiled datasets.

We ran a pairwise identity by descent (IBD) analysis within and across the combined dataset to detect duplicate and related individuals based on resulting IBD probabilities $Z_0$, $Z_1$ and $Z_2$ ($Z_k$ is probability that a pair of subjects share $k$ alleles identical by descent, estimated from genome-wide SNP data). If $0 \leq Z_0 \leq 0.1$ and $0 \leq Z_1 \leq 0.1$ and $0.9 \leq Z_2 \leq 1.1$ then a pair was flagged as being identical twins or duplicates. Pairs were considered full siblings if $0.17 \leq Z_0 \leq 0.33$ and $0.4 \leq Z_1 \leq 0.6$ and $0.17 \leq Z_2 \leq 0.33$. Half siblings or avunculars were defined as having $0.4 \leq Z_1 \leq 0.6$ and $0 \leq Z_2 \leq 0.1$. Some of the duplicates flagged were expected, having been genotyped in multiple datasets and hence having the same cohort identifiers. In this case, one of each pair was randomly chosen for removal from the dataset. In instances where pairs showed pairwise genotype concordance rate $> 0.999$ but were not expected duplicates, both individuals were removed. Related individuals (full siblings, half siblings/avunculars) were not removed from the final datasets. In the HumanHap dataset, 107 individuals were removed because they were duplicates or flagged for removal in the genotyping step, leaving 6,787 subjects. In addition, 8 pairs of individuals were flagged as related. In the OmniExpress dataset, we removed 39 subjects leaving 5,917 IDs and 5 pairs of related subjects. In the Affymetrix dataset, 167 individuals were removed because they were duplicates or flagged for removal in secondary genotype data cleaning, leaving a total of 8,065 individuals. Across all three datasets, we identified 444 duplicate pairs (406 expected) and thus removed additional 482 individuals from analysis across all three platform families.

After removing duplicate and related pairs of IDs, we used EIGENSTRAT [48] to run principal component analysis (PCA) on each dataset, removing one member from each flagged pair of related individuals. For Affymetrix and HumanHap, we used approximately 12,000 SNPs from Yu et al [49] that were filtered to ensure low pairwise linkage disequilibrium (LD). For the OmniExpress dataset we used approximately 33,000 SNPs that were similarly filtered. The top principal components were manually checked for outliers.

To identify any SNPs that created spurious associations, we ran several logistic regression analyses among subjects that were selected as controls in the initial GWAS (i.e. excluding all case subjects). For each regression, we used cohort-specific controls from one original GWAS as cases and the rest of the controls in that dataset as controls. For example, in the OmniExpress dataset, we considered NHS controls from the gout GWAS as “cases” while treating controls from the gout (HPFS), endometrial cancer (NHS), colon cancer (NHS, HPFS and PHS), and mammographic density (NHS) as “controls”. We repeated this, treating each cohort-specific “controls set” as “cases” and all other controls as “controls”. For each GWAS, we extracted genome-wide significant SNPs ($p < 10^{-8}$) and examined QQ plots. In the Affymetrix dataset, 100 SNPs were flagged and removed. In the HumanHap dataset, 8 SNPs had $p < 10^{-8}$ in at least one of the QC regressions and were removed. No SNPs in the OmniExpress dataset had $p < 10^{-8}$ and hence, no SNP was removed.

### Imputation

After the datasets were combined and appropriate SNP and subjects filters applied, the compiled datasets were separately imputed. We used the 1000 Genomes Project ALL Phase I Integrated Release Version 3 Haplotypes excluding monomorphic and singleton sites (2010–11 data freeze, 2012-03-14 haplotypes) as the reference panel. SNP and indel genotypes were imputed in three steps. First, genotypes on each chromosome were split into chunks to facilitate windowed imputation in parallel using ChunkChromosome (v.2011-08-05). Then each chunk of chromosome was phased using MACH [50,51] (v.1.0.18.c). In the final step, Minimac
(v.2012-08-15) was used to impute the phased genotypes to approximately 31 million markers in the 1000 Genomes Project.

“Proof of Principle” GWAS–BMI and VTE

To validate our merged GWAS datasets, we conducted two proof-of-principle GWAS of one quantitative trait (BMI) and one binary trait (VTE). We defined BMI as weight \( \frac{\text{kg}}{\text{height}^2 \text{cm}} \) and obtained it by extracting information on weight from the accompanying questionnaire collected at time of blood draw. If weight information was missing, we extracted it from the questionnaire closest in time to time of blood draw. Height was extracted from the baseline questionnaire. We obtained data on BMI for 20,283 participants. VTE is a spectrum of disease that includes pulmonary embolism (PE) and deep vein thromboembolism (DVT). Physician-diagnosed PE has been asked on every biennial NHS questionnaire since 1982, and every NHSII and HPFS questionnaire since cohort inception. In the NHS, DVT without PE is captured when a nurse answers that she has had phlebitis or thrombophlebitis (ICD-9 = 453.x). In NHS, NHSII and HPFS cohorts through 2010 (we did not have VTE data for PHS), we identified 6,041 individuals who reported VTE. Self-reported PE was verified through medical records review by a trained physician (CK). DVT cases are based on self-report, though a validation study of 100 DVT cases found self-reports to be highly consistent (>96%) with medical record review. In total, we identified 1,364 VTE cases with GWAS data. We treated all non-VTE cases with GWAS data as controls (\( n = 17,628 \)). Since we did not have data on VTE in PHS, we excluded PHS from this analysis.

Statistical analysis–GWAS

SNPs and indels with an imputation quality score <0.3 (as defined by the RSQR_HAT value in MACH) or a minor allele frequency (MAF) <0.01 were excluded. Primary association analysis was performed separately within each platform family (HumanHap, OmniExpress and Affymetrix). For imputed SNPs, the estimated number of effect alleles (ranging from 0 to 2) was used as a covariate. For BMI, we conducted linear regression adjusting for study (indicator variables including cohort as well as primary GWAS outcome), age at blood draw and the top four principal components. For VTE, we conducted logistic regression adjusting for study as above and the top four principal components. For both BMI and VTE, we combined platform family-specific results with fixed-effects meta-analysis using the METAL [52] software. We used the Cochran’s Q statistic to test for heterogeneity across studies.

Results

Imputation statistics

We imputed a total of 31,326,389 markers (29,890,747 SNPs and 1,435,642 indels) and the majority (69%) of these had a MAF\( \leq 0.01 \). The average imputation quality score by minor frequency for each platform family is shown in Fig 2 and the distribution of imputation quality score for rare (MAF\( \leq 0.01 \)) variants is shown in S1 Table. The imputation quality was very similar across all three datasets (S1–S3 Figs) with 49–51% of markers having an imputation quality score \( \geq 0.3 \). When restricting to markers with MAF>0.01 (~10 million), 92–94% of the markers had a quality score \( \geq 0.3 \), compared to 29–32% of markers with MAF\( \leq 0.01 \). After filtering markers based on MAF (>0.01) and imputation r-sq \( \geq 0.3 \), approximately 9.8 million markers were available for analysis.
BMI results

We had BMI and GWAS data for 20,283 individuals (n = 6,762 for HumanHap, n = 5,844 for OmniExpress, n = 7,677 for AffyMetrix) within NHS, NHSII, HPFS and PHS. Platform-specific QQ-plots (S4A–S4C Fig) showed no indication of systematic bias (genomic inflation factor λ = 1.00–1.02). The results from the meta-analysis are shown in Figs 3 and 4. We observed a tail of strongly associated SNPs with the top SNPs located in the known BMI FTO locus (strongest associated SNP: rs55872725, \( \beta = 0.45, p = 3.48 \times 10^{-22} \)). Given that the FTO locus has also been associated with Type 2 Diabetes, we reran the analysis excluding all Type 2 Diabetes cases (n = 2,540), The association for the FTO SNP rs55872725 remained strongly significant.
Fig 3. QQ-plot for GWAS analysis of body mass index based on 20,283 individuals.
https://doi.org/10.1371/journal.pone.0173997.g003

Fig 4. Manhattan plot for GWAS analysis of body mass index based on 20,283 individuals.
https://doi.org/10.1371/journal.pone.0173997.g004
We also observed genome-wide significant associations for the previously identified TMEM18 (strongest associated SNP: rs7563362, $\beta = -0.36$, $p = 1.76 \times 10^{-8}$) and FANCL loci (strongest associated SNP: rs980183, $\beta = -0.26$, $p = 2.73 \times 10^{-8}$). None of the SNPs that were originally reported were the top SNP in our data. However, for these three regions (S5A and S5B Fig, S2 Table), our top SNPs showed strong LD with the original reported SNPs (FTO locus: $r^2 = 0.97$ for rs55872725 and rs1558902; TMEM18 locus: $r^2 = 1.00$ for rs7563362 and rs13021737; FANCL locus: $r^2 = 0.72$ for rs980183 and rs1016287). Using a significance level of $p = 0.05$, 59% (19/32) known BMI SNPs [53], showed association with BMI in our data. In addition, 31 out of the 32 known SNPs showed associations in the same direction as the original BMI study (Fig 5).

VTE results

We had information on VTE status and GWAS data for 1,364 cases and 17,628 controls within NHS, NHSII and HPFS. The median number of case subjects by dataset was 87.5 and ranged from 16 in the NHSII breast cancer GWAS dataset (total of 289 individuals) to 417 in the type 2 diabetes GWAS dataset (total of 5,773 individuals). The small number of cases in many individual GWAS data sets led to unstable study-specific association statistics. Restricting to studies with an expected case minor allele count $> 10$ for SNPs with a MAF of 0.05 (i.e. studies with at least 200 cases) reduced the sample size to 417 cases and 5,556 controls. However, within each compiled imputed GWAS dataset, VTE case numbers ranged from 406 (OmniExpress) to 532 (Affymetrix). Thus, combining the individual GWAS datasets into three main datasets enabled association analysis of hundreds of cases rather than tens, leading to more stable estimates in the regression analysis. Platform-specific QQ-plots (S5A–S5C Fig) showed no indication of systematic bias (genomic inflation factor $\lambda = 1.00$–1.01). The results from the meta-analysis are shown in Figs 6 and 7 (genomic inflation factor $\lambda = 1.00$). We observed a strong association located downstream of the F5 gene (strongest associated SNP: rs2040445, OR = 2.17, 95% CI: 1.79–2.63, $p = 2.70 \times 10^{-15}$). We also observed genome-wide significant associations for the ABO locus (strongest associated SNP: rs2519093, OR = 1.36, 95% CI: 1.23–1.49, $p = 1.51 \times 10^{-10}$) and a nominal association ($P = 0.007$ with the previously VTE-associated F11 locus. For both the F5 and ABO regions (S6A and S6B Fig, S2 Table), our top SNPs showed moderate correlation with previously reported top SNP (ABO locus: $r^2 = 0.53$ for rs529565 and rs2519093 and F5 locus: $D^2 = 1.00$, $r^2 = 0.00$ for rs6025 and rs2040445 and $D^2 = 1.00$, $r^2 = 0.03$ for rs4524 and rs2040445). Using a significance level of $p = 0.05$, three of nine known VTE SNPs [54], showed association with VTE in our data, however, the directions of association were the same as previously observed for all SNPs (S3 Table).

Discussion

Thousands of genetic loci associated with hundreds of complex traits have been identified through GWAS and as sample sizes continue to increase, more loci will be discovered. Although the cost of GWAS has dropped, lack of financial resources is still the limiting factor for generating new data. Most GWAS have been conducted in case-control studies, and this has led to the creation of disease-specific consortia in which power can be maximized. However, there is usually only one disease phenotype available from these cases, and little capacity to follow cases or controls to collect information on additional phenotypes that develop over time. Cohort studies are designed to collect multiple endpoints on individuals, but often suffer from limited power for a specific disease. To maximize the utility of existing cohort data resources, it is important to explore associations with additional traits and outcomes that have been collected for individuals in multiple cohorts. In particular, the accumulation of GWAS
Fig 5. Associations for known body mass index SNPs based on 20,283 individuals.

https://doi.org/10.1371/journal.pone.0173997.g005
Fig 6. QQ-plot for GWAS analysis of venous thromboembolism based on 1,364 cases and 17,628 controls.
https://doi.org/10.1371/journal.pone.0173997.g006

Fig 7. Manhattan plot for GWAS analysis of venous thromboembolism based on 1,364 cases and 17,628 controls.
https://doi.org/10.1371/journal.pone.0173997.g007
data within large cohorts with rich environmental and outcome data creates new opportunities to assess novel hypotheses. In addition, cohort studies provide unique opportunities to prospectively assess biomarker-disease associations, thereby minimizing bias due to reverse causation or treatment effects. However, “borrowing” GWAS data between traits is not straightforward. Known issues that can cause bias include technical artifacts due to different genotyping platforms, differences in imputation accuracy and ascertainment bias. Thus, careful data management, imputation procedures and quality checks are needed. Furthermore, if the secondary trait is rare, there will be low phenotypic variability within each GWAS dataset. For example, we observed fewer than 100 VTE cases within the majority of individual GWAS, compared to more than 400 cases within each combined dataset.

Our pipeline for combining and imputing twelve different GWAS datasets can overcome both technical and methodological issues. We chose to create three different datasets defined by platform family (in our case, Illumina HumanHap, Illumina OmniExpress and AffyMetrix) since the SNP overlap across platforms was low on a genome-wide scale (75,285 SNPs). An attempt to impute a genome-wide dataset comprising only 75,000 SNPs as starting point would have resulted in decreased imputation accuracy in regions of the genome with sparse genotype data. Moreover, it has been shown that different platforms might call SNPs differently and that SNP-specific allele frequencies can differ between platforms (see [41] for further discussion). We conducted multiple case-control GWAS among control subjects within each dataset (i.e. running multiple “null” GWAS) and identified and excluded more than 100 SNPs that showed spurious associations. These results emphasize that although datasets are merged by platform family, problematic SNPs giving rise to spurious associations might still exist and it is important to carefully check for these.

To assess the validity of our data, we conducted two proof-of-principle GWAS. The first trait we studied was BMI, and in line with what expected, we observed strong evidence of associations with known BMI loci including FTO and TMEM18 that both reached genome-wide significance ($P < 5 \times 10^{-8}$). In addition, out of 32 known BMI SNPs we observed nominal significance ($P < 0.05$) for 19 of them, all in the same direction as expected from previous reports. Of note, our sample size ($n = 20,823$) is less than 10% of the original GWAS that had a total sample size of 249,766 individuals. Therefore, we would not expect to observe significant associations for all BMI SNPs due to limited power. For VTE, we observed genome-wide significant associations for the F5 and ABO loci that are both known to be associated with VTE. In addition, we also observed a nominal association ($P = 0.007$) with the F11 region. Our BMI and VTE results confirm that GWAS analysis of secondary traits in this data is valid and provides a platform for future studies of secondary traits. We ran the BMI and VTE analyses twice, the first time without removing duplicates between the datasets (total of 444 pairs), and the second time with the duplicates removed. Although the 444 pairs constitute less than 5% of our total sample size, including them had an impact on the genomic inflation factor (for BMI, the genomic inflation factor went from 1.09 to 1.05 and for VTE, the genomic inflation factor went from 1.02 to 1.00). These results are especially interesting as it is often difficult to identify duplicates across studies when raw data from all participating studies are not available. Care should be taken to remove overlapping subjects across GWAS contributing to a meta-analysis, but any remaining cryptic overlap may inflate association statistics. In that case, statistical adjustment procedures like LD score regression [55] can be used to account for cryptic overlap.

One of the main benefits with collecting comprehensive genetic information on cohort subjects is the opportunity to assess interactions between genetic factors and prospectively collected environmental data. To date, few gene-environment interactions have been identified and although their extent and clinical impact remain an open empirical question, the current
lack of homogenous large datasets with both genetic and environmental data has precluded comprehensive investigation. Capitalizing on this GWAS resource, we will be able to explore gene-environment interactions for a plethora of outcomes including complex traits such as height and BMI, but also disease outcomes. It will also allow us to study the impact of environmental factors within genetic strata to identify individuals for whom a particular intervention might be especially important [56–59].

Accumulation of these GWAS data is ongoing and we expect to generate new GWAS data for an additional 15,000 participants within the next two years, almost doubling our total GWAS sample size. This growing resource will be a core component of future studies aiming to elucidate how genes and the environment impact public health.

Supporting information

S1 Fig. Proportion of successfully imputed markers on the Affymetrix platform. Different colors correspond to different imputation quality score r-sq thresholds. Data is categorized by minor allele frequency.
(PDF)

S2 Fig. Proportion of successfully imputed markers on the Illumina HumanHap platform. Different colors correspond to different imputation quality score r-sq thresholds. Data is categorized by minor allele frequency.
(PDF)

S3 Fig. Proportion of successfully imputed markers on the Illumina Omniexpress platform. Different colors correspond to different imputation quality score r-sq thresholds. Data is categorized by minor allele frequency.
(PDF)

S4 Fig. A: QQ-plot for GWAS analysis of body mass index on the Illumina Omniexpress platform (n = 5,844). B: QQ-plot for GWAS analysis of body mass index on the Affymetrix platform (n = 7,677). C: QQ-plot for GWAS analysis of body mass index on the Illumina HumanHap platform (n = 6,762).
(PDF)

S5 Fig. A: LocusZoom plot for the BMI FTO locus. B: LocusZoom plot for the BMI TMEM18 locus. C: LocusZoom plot for the BMI FANCL locus.
(PDF)

S6 Fig. A: QQ-plot for GWAS analysis of venous on the Illumina Omniexpress platform (406 cases and 4,786 controls). B: QQ-plot for GWAS analysis of venous on the Illumina Omniexpress platform (406 cases and 4,786 controls). C: QQ-plot for GWAS analysis of venous on the Affymetrix platform (532 cases and 7,147 controls).
(PDF)

S7 Fig. A: LocusZoom plot for the VTE F5 locus. B: LocusZoom plot for the VTE ABO locus.
(PDF)

S1 Table. Number of SNPs (N) with MAF \( \leq 0.01 \) overall and by imputation quality score (r-sq) threshold for the three platforms Illumina HumanHap, AffyMetrix 6.0 and Illumina Omniexpress.
(PDF)

S2 Table. Associations for previously reported lead SNPs in regions that were genome-wide significant in analysis based on 20,283 individuals (BMI) and 1,364 cases and 17,628
controls (VTE) in NHS, NHSII, HPFS and PHS.
(PDF)

S3 Table. Association with known VTE SNPs (Germain et al, AJHG 2015) based on our analysis including 1,364 cases and 17,628 controls from NHS, NHSII and HPFS
(PDF)

Author Contributions

Conceptualization: PK.
Data curation: S. Loomis CT HH JH MC MJ.
Formal analysis: S. Lindström S. Loomis CT HH JH HA MC MJ.
Funding acquisition: S. Lindström ATC HC GC IDV AHE CF MG SEH FH JHK CK LRP ER MJS RMT SST JLV DJH.
Methodology: PK.
Resources: ATC HC GC IDV AHE CF MG SEH FH JHK CK LRP ER MJS RMT SST JLV DJH.
Supervision: PK.
Visualization: S. Lindström CT HH.
Writing – original draft: S. Lindström PK.
Writing – review & editing: S. Lindström S. Loomis CT HH HA ATC HC MC GC IDV AHE CF MG SEH FH MJ JHK CK LL LRP ER MJS RMT SST JLW DJH PK.

References

8. Ridker PM, Chasan-Taber S, Zee RY, Parker A, Rose L, Cook NR, et al. Rationale, design, and methodology of the Women's Genome Health Study: a genome-wide association study of more than 25,000


