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

Sara Lindström[1,2,3], Stephanie Loomis[4], Constance Turman[1,2], Hongyan Huang[1,2], Jinyan Huang[1,2], Hugues Aschard[1,2], Andrew T. Chan[5], Hyon Choi[6], Marilyn Cornelis[7], Gary Curhan[8,9], Immaculata De Vivo[1,2,8], A. Heather Eliassen[2,8], Charles Fuchs[8,10], Michael Gaziano[11], Susan E. Hankinson[2,7,12], Frank Hu[2,13], Majken Jensen[8,13], Jae H. Kang[8], Christopher Kabrhel[8,14], Liming Liang[1,2,15], Louis R. Pasquale[4,8], Eric Rimm[2,8,13], Meir J. Stampfer[2,8,13], Rulla M. Tamimi[2,8], Shelley S. Tworoger[2,8], Janey L. Wiggs[4], David J. Hunter[1,2,8,13] and Peter Kraft[1,2,15]

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

Keywords: GWAS, cohorts, imputation, secondary analysis, association

Correspondence to
Sara Lindstrom
University of Washington
Office F247B
Box 357236
Department of Epidemiology
Health Sciences Building
Tel: 206-221-3148
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=2.70x10^{-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).

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.

154 **Table 1: GWAS datasets in HPFS, NHS, NHSII and PHS**
155

Cohort	Outcome	Subjects (cases/controls)	Platform	GWAS dataset
HPFS	Coronary Heart Disease	435/878	Affymetrix 6.0	AffyMetrix
HPFS	Type 2 Diabetes	1,189/1,298	Affymetrix 6.0	AffyMetrix
HPFS	Pancreatic Cancer	54/52	Illumina 550k	Illumina HumanHap
HPFS	Kidney Stone	315/238	Illumina 610k	Illumina HumanHap
HPFS	Prostate Cancer	218/205	Illumina 610k	Illumina HumanHap
HPFS	Glaucoma	178/299	Illumina 660W	Illumina HumanHap
HPFS	Glioma	26/0	Illumina 660W	Illumina HumanHap
HPFS	Colon Cancer	229/230	Illumina OmniExpress	Illumina OmniExpress
HPFS	Gout	717/699	Illumina OmniExpress	Illumina OmniExpress
	SUBTOTAL			**7,018 (1,511 Illumina Human Hapmap, 3,634 Affymetrix, 1,873 Illumina OmniExpress)**
NHS	Type 2 Diabetes	1,532/1,754	Affymetrix 6.0	AffyMetrix
NHS	Coronary Heart Disease	342/804	Affymetrix 6.0	AffyMetrix
NHS	Ovarian Cancer	36/0	Illumina 317k	Illumina HumanHap
NHS	Breast Cancer	1,145/1,142	Illumina 550k	Illumina HumanHap
NHS	Pancreatic Cancer	82/84	Illumina 550k	Illumina HumanHap
NHS	Kidney Stone	328/166	Illumina 610k	Illumina HumanHap
NHS	Glaucoma	313/497	Illumina 660W	Illumina HumanHap
NHS	Glioma	38/0	Illumina 660W	Illumina HumanHap
NHS	Endometrial Cancer	396/348	Illumina OmniExpress	Illumina OmniExpress
NHS	Colon Cancer	394/774	Illumina OmniExpress	Illumina OmniExpress
NHS	Mammographic density	153/641	Illumina OmniExpress	Illumina OmniExpress
NHS	Gout	319/392	Illumina OmniExpress	Illumina OmniExpress
	SUBTOTAL			**11,522 (3,711 Illumina Human Hapmap, 4,413 Affymetrix, 3,380 Illumina OmniExpress)**
NHSII	Breast Cancer	289/0	Illumina 610k	Illumina HumanHap
NHSII	Kidney Stone	341/294	Illumina 610k	Illumina HumanHap
	SUBTOTAL			**924 (924 Illumina Human Hapmap, 0 Affymetrix, 0 Illumina OmniExpress)**
PHS	Pancreatic Cancer	49/54	Illumina 550k	Illumina HumanHap
PHS	Prostate Cancer	312/363	Illumina 610k	Illumina HumanHap
PHS	Colon Cancer	331/333	Illumina OmniExpress	Illumina OmniExpress
	SUBTOTAL			**1,305 (641 Illumina Human Hapmap, 0 Affymetrix, 664 Illumina OmniExpress)**
	TOTAL			**20,769 (6,787 Illumina Human Hapmap, 8,065 Affymetrix, 5,917 Illumina OmniExpress)**

156 Dataset merging

157 Successfully merging genotype data for different individuals requires complete overlap in SNPs.
158 SNPs that are missing by design (due to different genotyping platforms) from some studies will
159 be correlated with the primary phenotype for that dataset. This might cause spurious results in
160 any secondary analysis on related traits. Although a missing SNP can be imputed, it will have a
161 higher degree of inaccuracy in imputed compared with genotyped SNPs, potentially creating
162 differential measurement error that could also lead to bias [41,46,47]. Therefore, we first looked
163 at the overlap of SNPs between different genotyping arrays and identified three broad platform
164 families with high degree of overlap within category but low overlap across categories – the
165 earlier generation of Illumina arrays (HumanHap), the Illumina OmniExpress array and
166 Affymetrix 6.0 array. The HumanHap platform had a total of 459,999 SNPs compared with
167 565,810 SNPs for OmniExpress and 668,283 SNPs for Affymetrix 6.0. However, the intersection
168 among all three platform families was only 75,285 SNPs (**Figure 1**). To achieve the largest GWAS
169 datasets as possible without losing SNP information, we created three datasets – HumanHap
170 comprising six GWAS datasets, OmniExpress comprising four GWAS datasets and Affymetrix 6.0
171 comprising two GWAS datasets. In the merging process, we removed any SNPs that were not in
172 all studies for a specific platform or had a missing call rate>5%. We flipped strands where
173 appropriate and removed A/T and C/G SNPs to create the final compiled datasets.
174
175 **Figure 1:** Overlap in SNPs across genotype platforms
176
177 We ran a pairwise identity by descent (IBD) analysis within and across the combined dataset to
178 detect duplicate and related individuals based on resulting IBD probabilities Z_0, Z_1 and Z_2 (Z_k is
179 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 were flagged for removal from 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.

228 "Proof of Principle" GWAS– BMI and VTE

229 To validate our merged GWAS datasets, we conducted two proof-of principle GWAS of one
230 quantitative trait (BMI) and one binary trait (VTE). We defined BMI as weight (kg)/height^2 (cm)
231 and obtained it by extracting information on weight from the accompanying questionnaire
232 collected at time of blood draw. If weight information was missing, we extracted it from the
233 questionnaire closest in time to time of blood draw. Height was extracted from the baseline
234 questionnaire. We obtained data on BMI for 20,283 participants. VTE is a spectrum of disease
235 that includes pulmonary embolism (PE) and deep vein thromboembolism (DVT). Physician-
236 diagnosed PE has been asked on every biennial NHS questionnaire since 1982, and every NHSII
237 and HPFS questionnaire since cohort inception. In the NHS, DVT without PE is captured when a
238 nurse answers that she has had phlebitis or thrombophlebitis (ICD-9=453.x). In NHS, NHSII and
239 HPFS cohorts through 2010 (we did not have VTE data for PHS), we identified 6,041 individuals
240 who reported VTE. Self-reported PE was verified through medical records review by a trained
241 physician (CK). DVT cases are based on self-report, though a validation study of 100 DVT cases
242 found self-reports to be highly consistent (>96%) with medical record review. In total, we
243 identified 1,364 VTE cases with GWAS data. We treated all non-VTE cases with GWAS data as
244 controls (n=17,628). Since we did not have data on VTE in PHS, we excluded PHS from this
245 analysis.
246
247 Statistical analysis – GWAS
248 SNPs and indels with an imputation quality score <0.3 (as defined by the RSQR_HAT value in
249 MACH) or a minor allele frequency (MAF) <0.01 were excluded. Primary association analysis was
250 performed separately within each platform family (HumanHap, OmniExpress and Affymetrix).
251 For imputed SNPs, the estimated number of effect alleles (ranging from 0 to 2) was used as a

252 covariate. For BMI, we conducted linear regression adjusting for study (indicator variables
253 including cohort as well as primary GWAS outcome), age at blood draw and the top four
254 principal components. For VTE, we conducted logistic regression adjusting for study as above
255 and the top four principal components. For both BMI and VTE, we combined platform family-
256 specific results with fixed-effects meta-analysis using the METAL [52] software. We used the
257 Cochran's Q statistic to test for heterogeneity across studies.
258
259

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<0.01. The average imputation quality score by minor frequency for each platform family is shown in **Figure 2**. The imputation quality was very similar across all three datasets (**Suppl Figures 1-3**) with 49-51% of markers having an imputation quality score ≥0.3. When restricting to markers with MAF>0.01 (~10 million), 92-94% of the markers had a quality score ≥0.3. After filtering markers based on MAF (>0.01) and imputation r-sq (≥0.3), approximately 9.8 million markers were available for analysis.

Figure 2: Imputation quality score by minor allele frequency for the three platform families

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 (**Suppl Figures 4a-c**) showed no indication of systematic bias (genomic inflation factor λ=1.00-1.02). The results from the meta-analysis are shown in **Figures 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, β=0.45, p=3.48x10^{-22}). We also observed genome-wide significant associations for the previously identified *TMEM18* (strongest associated SNP: rs7563362, β=-0.36, p=1.76x10^{-8}) and *FANCL* loci (strongest associated SNP: rs980183, β=-0.26, p=2.73x10^{-8}). 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 **(Figure 5).**

Figure 3: QQ-plot for GWAS analysis of body mass index based on 20,283 individuals.

Figure 4: Manhattan plot for GWAS analysis of body mass index based on 20,283 individuals.

Figure 5: Associations for known body mass index.

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 5,356 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 (**Suppl Figures 5a-c**) showed no indication of systematic bias (genomic inflation factor λ=1.00-1.01). The results from the meta-analysis are shown in **Figures 6 and 7** (genomic inflation factor λ=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.

307 **Figure 6:** QQ-plot for GWAS analysis of venous thromboembolism based on 1,364 cases and

308 17,628 controls.

309 **Figure 7:** Manhattan plot for GWAS analysis of venous thromboembolism based on 1,364 cases

310 and 17,628 controls.

311

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 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.

335 Our pipeline for combining and imputing twelve different GWAS datasets can overcome both
336 technical and methodological issues. We chose to create three different datasets defined by
337 platform family (in our case, Illumina HumanHap, Illumina OmniExpress and AffyMetrix) since
338 the SNP overlap across platforms was low on a genome-wide scale (75,285 SNPs). An attempt to
339 impute a genome-wide dataset comprising only 75,000 SNPs as starting point would have
340 resulted in decreased imputation accuracy in regions of the genome with sparse genotype data.
341 Moreover, it has been shown that different platforms might call SNPs differently and that SNP-
342 specific allele frequencies can differ between platforms (see [41] for further discussion). We
343 conducted multiple case-control GWAS among control subjects within each dataset (i.e. running
344 multiple "null" GWAS) and identified and excluded more than 100 SNPs that showed spurious
345 associations. These results emphasize that although datasets are merged by platform family,
346 problematic SNPs giving rise to spurious associations might still exist and it is important to
347 carefully check for these.
348
349 To assess the validity of our data, we conducted two proof-of-principle GWAS. The first trait we
350 studied was BMI, and in line with what expected, we observed strong evidence of associations
351 with known BMI loci including *FTO* and *TMEM18* that both reached genome-wide significance
352 ($P<5x10^{-8}$). In addition, out of 32 known BMI SNPs we observed nominal significance ($P<0.05$) for
353 19 of them, all in the same direction as expected from previous reports. Of note, our sample size
354 (n=20,823) is less than 10% of the original GWAS that had a total sample size of 249,766
355 individuals. Therefore, we would not expect to observe significant associations for all BMI SNPs
356 due to limited power. For VTE, we observed genome-wide significant associations for the *F5* and
357 *ABO* loci that are both known to be associated with VTE. In addition, we also observed a nominal
358 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 [54] 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 [55-58].

Accumulation of these GWAS data is ongoing and we expect to generate new GWAS data for an

383 additional 15,000 participants within the next two years, almost doubling our total GWAS
384 sample size. This growing resource will be a core component of future studies aiming to
385 elucidate how genes and the environment impact public health.
386
387
388 **ACKNOWLEDGEMENTS**
389 This work was supported by National Institute of Health (P01CA87969, P01CA055075,
390 P01DK070756, U01HG004728, UM1CA186107, UM1CA176726, R01CA49449, R01CA50385,
391 R01CA67262, R01CA131332, R01HL034594, R01HL088521, R01HL35464, R01HL116854,
392 R01EY015473, R01EY022305, P30EY014104, R03DC013373 and R03CA165131). Dr. Pasquale is
393 also supported by a Harvard Medical School Distinguished Ophthalmology Scholar award. The
394 funders had no role in study design, data collection and analysis, decision to publish, or
395 preparation of the manuscript.
396
397
398
399
400

401 **REFERENCES**

402 1. Doll R, Hill AB. The mortality of doctors in relation to their smoking habits: a preliminary
403 report. 1954. Bmj. 2004;328(7455):1529-33; discussion 33. doi: 10.1136/bmj.328.7455.1529.
404 PubMed PMID: 15217868; PubMed Central PMCID: PMC437141.

405 2. Doll R, Hill AB. Lung cancer and other causes of death in relation to smoking; a second
406 report on the mortality of British doctors. British medical journal. 1956;2(5001):1071-81.
407 PubMed PMID: 13364389; PubMed Central PMCID: PMC2035864.

408 3. Kannel WB, Dawber TR, Kagan A, Revotskie N, Stokes J, 3rd. Factors of risk in the
409 development of coronary heart disease--six year follow-up experience. The Framingham Study.
410 Annals of internal medicine. 1961;55:33-50. PubMed PMID: 13751193.

411 4. Willett WC, Stampfer MJ, Manson JE, Colditz GA, Speizer FE, Rosner BA, et al. Intake of
412 trans fatty acids and risk of coronary heart disease among women. Lancet. 1993;341(8845):581-
413 5. PubMed PMID: 8094827.

414 5. Mayers JR, Wu C, Clish CB, Kraft P, Torrence ME, Fiske BP, et al. Elevation of circulating
415 branched-chain amino acids is an early event in human pancreatic adenocarcinoma
416 development. Nature medicine. 2014;20(10):1193-8. doi: 10.1038/nm.3686. PubMed PMID:
417 25261994; PubMed Central PMCID: PMC4191991.

418 6. Zhang X, Tworoger SS, Eliassen AH, Hankinson SE. Postmenopausal plasma sex hormone
419 levels and breast cancer risk over 20 years of follow-up. Breast cancer research and treatment.
420 2013;137(3):883-92. doi: 10.1007/s10549-012-2391-z. PubMed PMID: 23283524; PubMed
421 Central PMCID: PMC3582409.

422 7. Manolio TA. Bringing genome-wide association findings into clinical use. Nature reviews
423 Genetics. 2013;14(8):549-58. doi: 10.1038/nrg3523. PubMed PMID: 23835440.

8. Ridker PM, Chasman DI, 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 initially healthy american women. Clinical chemistry. 2008;54(2):249-55. doi: 10.1373/clinchem.2007.099366. PubMed PMID: 18070814.

9. Banda Y, Kvale MN, Hoffmaknn TJ, Hesselson SE, Ranatunga D, Tang H, et al. Characterizing Race/Ethnicity and Genetic Ancestry for 100,000 Subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) Cohort. Genetics. 2015;200(4):1285-95. doi: 10.1534/genetics.115.178616. PubMed PMID: 26092716; PubMed Central PMCID: PMCPMC4574246.

10. Colditz GA, Hankinson SE. The Nurses' Health Study: lifestyle and health among women. Nat Rev Cancer. 2005;5(5):388-96. Epub 2005/05/03. doi: 10.1038/nrc1608. PubMed PMID: 15864280.

11. Tworoger SS, Sluss P, Hankinson SE. Association between plasma prolactin concentrations and risk of breast cancer among predominately premenopausal women. Cancer research. 2006;66(4):2476-82. doi: 10.1158/0008-5472.CAN-05-3369. PubMed PMID: 16489055.

12. Giovannucci E, Pollak M, Liu Y, Platz EA, Majeed N, Rimm EB, et al. Nutritional predictors of insulin-like growth factor I and their relationships to cancer in men. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology. 2003;12(2):84-9. PubMed PMID: 12582016.

13. Sesso HD, Gaziano JM, VanDenburgh M, Hennekens CH, Glynn RJ, Buring JE. Comparison of baseline characteristics and mortality experience of participants and nonparticipants in a randomized clinical trial: the Physicians' Health Study. Controlled clinical trials. 2002;23(6):686-702. PubMed PMID: 12505246.

14. Qi L, Cornelis MC, Kraft P, Stanya KJ, Linda Kao WH, Pankow JS, et al. Genetic variants at 2q24 are associated with susceptibility to type 2 diabetes. Human molecular genetics. 2010;19(13):2706-15. doi: 10.1093/hmg/ddq156. PubMed PMID: 20418489; PubMed Central PMCID: PMC2883345.

15. Jensen MK, Pers TH, Dworzynski P, Girman CJ, Brunak S, Rimm EB. Protein interaction-based genome-wide analysis of incident coronary heart disease. Circulation Cardiovascular genetics. 2011;4(5):549-56. doi: 10.1161/CIRCGENETICS.111.960393. PubMed PMID: 21880673; PubMed Central PMCID: PMC3197770.

16. Hunter DJ, Kraft P, Jacobs KB, Cox DG, Yeager M, Hankinson SE, et al. A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nature genetics. 2007;39(7):870-4. doi: 10.1038/ng2075. PubMed PMID: 17529973; PubMed Central PMCID: PMC3493132.

17. De Vivo I, Prescott J, Setiawan VW, Olson SH, Wentzensen N, Australian National Endometrial Cancer Study G, et al. Genome-wide association study of endometrial cancer in E2C2. Human genetics. 2014;133(2):211-24. doi: 10.1007/s00439-013-1369-1. PubMed PMID: 24096698; PubMed Central PMCID: PMC3898362.

18. Schumacher FR, Berndt SI, Siddiq A, Jacobs KB, Wang Z, Lindstrom S, et al. Genome-wide association study identifies new prostate cancer susceptibility loci. Human molecular genetics. 2011;20(19):3867-75. doi: 10.1093/hmg/ddr295. PubMed PMID: 21743057; PubMed Central PMCID: PMC3168287.

19. Peters U, Jiao S, Schumacher FR, Hutter CM, Aragaki AK, Baron JA, et al. Identification of Genetic Susceptibility Loci for Colorectal Tumors in a Genome-Wide Meta-analysis. Gastroenterology. 2013;144(4):799-807 e24. doi: 10.1053/j.gastro.2012.12.020. PubMed PMID: 23266556; PubMed Central PMCID: PMC3636812.

20. Stevens KN, Lindstrom S, Scott CG, Thompson D, Sellers TA, Wang X, et al. Identification of a novel percent mammographic density locus at 12q24. Human molecular genetics. 2012;21(14):3299-305. doi: 10.1093/hmg/dds158. PubMed PMID: 22532574; PubMed Central PMCID: PMC3384385.

21. Lindstrom S, Thompson DJ, Paterson AD, Li J, Gierach GL, Scott C, et al. Genome-wide association study identifies multiple loci associated with both mammographic density and breast cancer risk. Nature communications. 2014;5:5303. doi: 10.1038/ncomms6303. PubMed PMID: 25342443; PubMed Central PMCID: PMC4320806.

22. Lettre G, Jackson AU, Gieger C, Schumacher FR, Berndt SI, Sanna S, et al. Identification of ten loci associated with height highlights new biological pathways in human growth. Nature genetics. 2008;40(5):584-91. doi: 10.1038/ng.125. PubMed PMID: 18391950; PubMed Central PMCID: PMC2687076.

23. Loos RJ, Lindgren CM, Li S, Wheeler E, Zhao JH, Prokopenko I, et al. Common variants near MC4R are associated with fat mass, weight and risk of obesity. Nature genetics. 2008;40(6):768-75. doi: 10.1038/ng.140. PubMed PMID: 18454148; PubMed Central PMCID: PMC2669167.

24. Heid IM, Jackson AU, Randall JC, Winkler TW, Qi L, Steinthorsdottir V, et al. Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. Nature genetics. 2010;42(11):949-60. doi: 10.1038/ng.685. PubMed PMID: 20935629; PubMed Central PMCID: PMC3000924.

25. Han J, Kraft P, Nan H, Guo Q, Chen C, Qureshi A, et al. A genome-wide association study identifies novel alleles associated with hair color and skin pigmentation. PLoS genetics. 2008;4(5):e1000074. doi: 10.1371/journal.pgen.1000074. PubMed PMID: 18483556; PubMed Central PMCID: PMC2367449.

496 26. He C, Kraft P, Chen C, Buring JE, Pare G, Hankinson SE, et al. Genome-wide association
497 studies identify loci associated with age at menarche and age at natural menopause. Nature
498 genetics. 2009;41(6):724-8. doi: 10.1038/ng.385. PubMed PMID: 19448621; PubMed Central
499 PMCID: PMC2888798.
500 27. Caporaso N, Gu F, Chatterjee N, Sheng-Chih J, Yu K, Yeager M, et al. Genome-wide and
501 candidate gene association study of cigarette smoking behaviors. PloS one. 2009;4(2):e4653.
502 doi: 10.1371/journal.pone.0004653. PubMed PMID: 19247474; PubMed Central PMCID:
503 PMC2644817.
504 28. Prescott J, Kraft P, Chasman DI, Savage SA, Mirabello L, Berndt SI, et al. Genome-wide
505 association study of relative telomere length. PloS one. 2011;6(5):e19635. doi:
506 10.1371/journal.pone.0019635. PubMed PMID: 21573004; PubMed Central PMCID:
507 PMC3091863.
508 29. Lindstrom S, Vachon CM, Li J, Varghese J, Thompson D, Warren R, et al. Common
509 variants in ZNF365 are associated with both mammographic density and breast cancer risk.
510 Nature genetics. 2011;43(3):185-7. doi: 10.1038/ng.760. PubMed PMID: 21278746; PubMed
511 Central PMCID: PMC3076615.
512 30. Nan H, Xu M, Zhang J, Zhang M, Kraft P, Qureshi AA, et al. Genome-wide association
513 study identifies nidogen 1 (NID1) as a susceptibility locus to cutaneous nevi and melanoma risk.
514 Human molecular genetics. 2011;20(13):2673-9. doi: 10.1093/hmg/ddr154. PubMed PMID:
515 21478494; PubMed Central PMCID: PMC3110001.
516 31. Hek K, Demirkan A, Lahti J, Terracciano A, Teumer A, Cornelis MC, et al. A genome-wide
517 association study of depressive symptoms. Biological psychiatry. 2013;73(7):667-78. doi:
518 10.1016/j.biopsych.2012.09.033. PubMed PMID: 23290196; PubMed Central PMCID:
519 PMC3845085.

32. The C, Caffeine Genetics C, Cornelis MC, Byrne EM, Esko T, Nalls MA, et al. Genome-wide meta-analysis identifies six novel loci associated with habitual coffee consumption. Molecular psychiatry. 2014. doi: 10.1038/mp.2014.107. PubMed PMID: 25288136; PubMed Central PMCID: PMC4388784.

33. Hazra A, Kraft P, Selhub J, Giovannucci EL, Thomas G, Hoover RN, et al. Common variants of FUT2 are associated with plasma vitamin B12 levels. Nature genetics. 2008;40(10):1160-2. doi: 10.1038/ng.210. PubMed PMID: 18776911; PubMed Central PMCID: PMC2673801.

34. Hazra A, Kraft P, Lazarus R, Chen C, Chanock SJ, Jacques P, et al. Genome-wide significant predictors of metabolites in the one-carbon metabolism pathway. Human molecular genetics. 2009;18(23):4677-87. doi: 10.1093/hmg/ddp428. PubMed PMID: 19744961; PubMed Central PMCID: PMC2773275.

35. Prescott J, Thompson DJ, Kraft P, Chanock SJ, Audley T, Brown J, et al. Genome-wide association study of circulating estradiol, testosterone, and sex hormone-binding globulin in postmenopausal women. PloS one. 2012;7(6):e37815. doi: 10.1371/journal.pone.0037815. PubMed PMID: 22675492; PubMed Central PMCID: PMC3366971.

36. Ahn J, Yu K, Stolzenberg-Solomon R, Simon KC, McCullough ML, Gallicchio L, et al. Genome-wide association study of circulating vitamin D levels. Human molecular genetics. 2010;19(13):2739-45. doi: 10.1093/hmg/ddq155. PubMed PMID: 20418485; PubMed Central PMCID: PMC2883344.

37. Major JM, Yu K, Wheeler W, Zhang H, Cornelis MC, Wright ME, et al. Genome-wide association study identifies common variants associated with circulating vitamin E levels. Human molecular genetics. 2011;20(19):3876-83. doi: 10.1093/hmg/ddr296. PubMed PMID: 21729881; PubMed Central PMCID: PMC3168288.

544 38. Mondul AM, Yu K, Wheeler W, Zhang H, Weinstein SJ, Major JM, et al. Genome-wide
545 association study of circulating retinol levels. Human molecular genetics. 2011;20(23):4724-31.
546 doi: 10.1093/hmg/ddr387. PubMed PMID: 21878437; PubMed Central PMCID: PMC3209826.
547 39. Qi L, Cornelis MC, Kraft P, Jensen M, van Dam RM, Sun Q, et al. Genetic variants in ABO
548 blood group region, plasma soluble E-selectin levels and risk of type 2 diabetes. Human
549 molecular genetics. 2010;19(9):1856-62. doi: 10.1093/hmg/ddq057. PubMed PMID: 20147318;
550 PubMed Central PMCID: PMC2850622.
551 40. Monsees GM, Tamimi RM, Kraft P. Genome-wide association scans for secondary traits
552 using case-control samples. Genetic epidemiology. 2009;33(8):717-28. doi: 10.1002/gepi.20424.
553 PubMed PMID: 19365863; PubMed Central PMCID: PMC2790028.
554 41. Sinnott JA, Kraft P. Artifact due to differential error when cases and controls are
555 imputed from different platforms. Human genetics. 2012;131(1):111-9. doi: 10.1007/s00439-
556 011-1054-1. PubMed PMID: 21735171; PubMed Central PMCID: PMC3217156.
557 42. Genomes Project C, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. An
558 integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491(7422):56-65.
559 doi: 10.1038/nature11632. PubMed PMID: 23128226; PubMed Central PMCID: PMC3498066.
560 43. Amundadottir L, Kraft P, Stolzenberg-Solomon RZ, Fuchs CS, Petersen GM, Arslan AA, et
561 al. Genome-wide association study identifies variants in the ABO locus associated with
562 susceptibility to pancreatic cancer. Nature genetics. 2009;41(9):986-90. doi: 10.1038/ng.429.
563 PubMed PMID: 19648918; PubMed Central PMCID: PMC2839871.
564 44. Wiggs JL, Kang JH, Yaspan BL, Mirel DB, Laurie C, Crenshaw A, et al. Common variants
565 near CAV1 and CAV2 are associated with primary open-angle glaucoma in Caucasians from the
566 USA. Human molecular genetics. 2011;20(23):4707-13. doi: 10.1093/hmg/ddr382. PubMed
567 PMID: 21873608; PubMed Central PMCID: PMC3209825.

568 45. Rajaraman P, Melin BS, Wang Z, McKean-Cowdin R, Michaud DS, Wang SS, et al.
569 Genome-wide association study of glioma and meta-analysis. Human genetics.
570 2012;131(12):1877-88. doi: 10.1007/s00439-012-1212-0. PubMed PMID: 22886559; PubMed
571 Central PMCID: PMC3761216.
572 46. Johnson EO, Hancock DB, Levy JL, Gaddis NC, Saccone NL, Bierut LJ, et al. Imputation
573 across genotyping arrays for genome-wide association studies: assessment of bias and a
574 correction strategy. Human genetics. 2013;132(5):509-22. doi: 10.1007/s00439-013-1266-7.
575 PubMed PMID: 23334152; PubMed Central PMCID: PMC3628082.
576 47. Uh HW, Deelen J, Beekman M, Helmer Q, Rivadeneira F, Hottenga JJ, et al. How to deal
577 with the early GWAS data when imputing and combining different arrays is necessary. European
578 journal of human genetics : EJHG. 2012;20(5):572-6. doi: 10.1038/ejhg.2011.231. PubMed PMID:
579 22189269; PubMed Central PMCID: PMC3330212.
580 48. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal
581 components analysis corrects for stratification in genome-wide association studies. Nature
582 genetics. 2006;38(8):904-9. doi: 10.1038/ng1847. PubMed PMID: 16862161.
583 49. Yu K, Wang Z, Li Q, Wacholder S, Hunter DJ, Hoover RN, et al. Population substructure
584 and control selection in genome-wide association studies. PloS one. 2008;3(7):e2551. doi:
585 10.1371/journal.pone.0002551. PubMed PMID: 18596976; PubMed Central PMCID:
586 PMC2432498.
587 50. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: using sequence and genotype data
588 to estimate haplotypes and unobserved genotypes. Genetic epidemiology. 2010;34(8):816-34.
589 doi: 10.1002/gepi.20533. PubMed PMID: 21058334; PubMed Central PMCID: PMC3175618.
590 51. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate
591 genotype imputation in genome-wide association studies through pre-phasing. Nature genetics.

2012;44(8):955-9. doi: 10.1038/ng.2354. PubMed PMID: 22820512; PubMed Central PMCID: PMC3696580.

52. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics. 2010;26(17):2190-1. doi: 10.1093/bioinformatics/btq340. PubMed PMID: 20616382; PubMed Central PMCID: PMC2922887.

53. Speliotes EK, Willer CJ, Berndt SI, Monda KL, Thorleifsson G, Jackson AU, et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nature genetics. 2010;42(11):937-48. doi: 10.1038/ng.686. PubMed PMID: 20935630; PubMed Central PMCID: PMC3014648.

54. Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J, Schizophrenia Working Group of the Psychiatric Genomics C, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nature genetics. 2015;47(3):291-5. doi: 10.1038/ng.3211. PubMed PMID: 25642630.

55. Qi Q, Kilpelainen TO, Downer MK, Tanaka T, Smith CE, Sluijs I, et al. FTO genetic variants, dietary intake and body mass index: insights from 177 330 individuals. Human molecular genetics. 2014. doi: 10.1093/hmg/ddu411. PubMed PMID: 25104851.

56. Qi Q, Chu AY, Kang JH, Huang J, Rose LM, Jensen MK, et al. Fried food consumption, genetic risk, and body mass index: gene-diet interaction analysis in three US cohort studies. Bmj. 2014;348:g1610. doi: 10.1136/bmj.g1610. PubMed PMID: 24646652; PubMed Central PMCID: PMC3959253.

57. Ahmad S, Rukh G, Varga TV, Ali A, Kurbasic A, Shungin D, et al. Gene x physical activity interactions in obesity: combined analysis of 111,421 individuals of European ancestry. PLoS genetics. 2013;9(7):e1003607. doi: 10.1371/journal.pgen.1003607. PubMed PMID: 23935507; PubMed Central PMCID: PMC3723486.

616 58. Qi Q, Chu AY, Kang JH, Jensen MK, Curhan GC, Pasquale LR, et al. Sugar-sweetened
617 beverages and genetic risk of obesity. The New England journal of medicine. 2012;367(15):1387-
618 96. doi: 10.1056/NEJMoa1203039. PubMed PMID: 22998338; PubMed Central PMCID:
619 PMC3518794.

620

621

622 **Supplemental Figure 1:** Proportion of sucessfully imputed markers on the Affymetrix platform

623 **Supplemental Figure 2:** Proportion of sucessfully imputed markers on the Illumina HumanHap

624 platform.

625 **Supplemental Figure 3:** Proportion of sucessfully imputed markers on the Illumina Omniexpress

626 platform.

627 **Supplemental Figure 4a:** QQ-plot for GWAS analysis of body mass index on the Illumina

628 Omniexpress platform (n=5,844).

629 **Supplemental Figure 4b:** QQ-plot for GWAS analysis of body mass index on the Affymetrix

630 platform (n=7,677).

631 **Supplemental Figure 4c:** QQ-plot for GWAS analysis of body mass index on the Illumina

632 HumanHap platform (n=6,762).

633 **Supplemental Figure 5a:** QQ-plot for GWAS analysis of venous on the Illumina Omniexpress

634 platform (406 cases and 4,786 controls).

635 **Supplemental Figure 5b:** QQ-plot for GWAS analysis of venous on the Illumina Omniexpress

636 platform (406 cases and 4,786 controls).

637 **Supplemental Figure 5c:** QQ-plot for GWAS analysis of venous on the Affymetrix platform (532

638 cases and 7,147 controls).

639

640

641