

ARTICLE

Missing heritability: is the gap closing? An analysis of 32 complex traits in the Lifelines Cohort Study

Ilja M Nolte^{*1,11}, Peter J van der Most^{1,11}, Behrooz Z Alizadeh¹, Paul IW de Bakker^{2,3}, H Marika Boezen¹, Marcel Bruinenberg⁴, Lude Franke⁵, Pim van der Harst⁶, Gerjan Navis⁷, Dirkje S Postma⁸, Marianne G Rots⁹, Ronald P Stolk¹, Morris A Swertz⁵, Bruce HR Wolffenbuttel¹⁰, Cisca Wijmenga⁵ and Harold Snieder^{*1}

Despite the recent explosive rise in number of genetic markers for complex disease traits identified in genome-wide association studies, there is still a large gap between the known heritability of these traits and the part explained by these markers. To gauge whether this ‘heritability gap’ is closing, we first identified genome-wide significant SNPs from the literature and performed replication analyses for 32 highly relevant traits from five broad disease areas in 13 436 subjects of the Lifelines Cohort. Next, we calculated the variance explained by multi-SNP genetic risk scores (GRSs) for each trait, and compared it to their broad- and narrow-sense heritabilities captured by all common SNPs. The majority of all previously-associated SNPs (median = 75%) were significantly associated with their respective traits. All GRSs were significant, with unweighted GRSs generally explaining less phenotypic variance than weighted GRSs, for which the explained variance was highest for height (15.5%) and varied between 0.02 and 6.7% for the other traits. Broad-sense common-SNP heritability estimates were significant for all traits, with the additive effect of common SNPs explaining 48.9% of the variance for height and between 5.6 and 39.2% for the other traits. Dominance effects were uniformly small (0–1.5%) and not significant. On average, the variance explained by the weighted GRSs accounted for only 10.7% of the common-SNP heritability of the 32 traits. These results indicate that GRSs may not yet be ready for accurate personalized prediction of complex disease traits limiting widespread adoption in clinical practice.

European Journal of Human Genetics (2017) 25, 877–885; doi:10.1038/ejhg.2017.50; published online 12 April 2017

INTRODUCTION

In recent years, many large international consortia have performed meta-analyses of genome-wide association studies (GWASs), identifying numerous associations of common genetic variants (ie, single nucleotide polymorphisms (SNPs)) with a wide variety of diseases and traits.^{1,2} The rapid increase in the number of SNPs identified provides an opportunity to systematically examine the quantitative impact of these common genetic variants, individually or collectively as part of a genetic risk score (GRS). Such GRSs offer promise for personalized prediction of complex disease risk with potential future application in clinical practice.

Lifelines is a multi-disciplinary population-based prospective cohort study examining the health and health-related behaviors of 167 729 persons living in the North East region of the Netherlands in a three-generation family design.^{3,4} The general aim of the Lifelines Cohort Study is to unravel how life-time exposure to environmental, and genetic risk factors and their interaction influences individual susceptibility to multifactorial diseases. Lifelines not only provides an in-

depth characterization of the biomedical, socio-demographic, behavioral, physical and psychological factors that contribute to health and disease in the general population, it also employs a broad disease- and organ-overriding phenotypic characterization of its participants allowing it to validly address questions concerning the multi-morbidity that occurs with ageing, rather than focusing on single-disease conditions. Its representativeness for the population in the Northern Netherlands was recently shown⁵ and the age-related multi-morbidity was documented.⁶ Thus, Lifelines is particularly suited to investigate causes of morbidity across disease domains. Along the same lines, genotype–phenotype associations were assessed for multiple disease domains in the current paper.

The aim of this study was to test the validity and reliability of the Lifelines Cohort Study for genetic research of a wide range of complex disease traits, and to gauge whether the ‘heritability gap’ of these traits is closing. To this end, we collected dedicated data on a wide variety of phenotypes and performed genome-wide genotyping on more than 13 000 unrelated individuals. We selected 32 continuous traits from

¹Unit of Genetic Epidemiology and Bioinformatics, Department of Epidemiology, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands;

²Department of Medical Genetics, Center for Molecular Medicine, University Medical Center Utrecht, Utrecht, The Netherlands; ³Department of Epidemiology, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands; ⁴Lifelines Cohort Study, Groningen, The Netherlands; ⁵Department of Genetics, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands; ⁶Department of Cardiology, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands; ⁷Department of Nephrology, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands; ⁸Department of Pulmonology, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands; ⁹Department of Medical Biology, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands; ¹⁰Department of Endocrinology, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands

*Correspondence: Dr IM Nolte or Dr H Snieder, Unit of Genetic Epidemiology and Bioinformatics, Department of Epidemiology, University of Groningen, University Medical Center Groningen, Hanzplein 1, PO Box 30001, Groningen 9700 RB, The Netherlands. Tel: +31 50 361 0946/+31 50 361 0887; Fax: +31 50 361 4493; E-mail: i.m.nolte@umcg.nl or h.snieider@umcg.nl

¹¹These authors contributed equally to this work.

Received 30 August 2016; revised 3 February 2017; accepted 14 February 2017; published online 12 April 2017

five broad disease areas (musculoskeletal, cardiovascular and renal, metabolic, hematologic and inflammatory, and pulmonary), for which we compiled a list of genome-wide significantly associated index SNPs based on the GWAS catalog² and performed an extensive literature search. For each of the 32 traits we (i) tested whether associations with previously identified index SNPs could be replicated in the Lifelines Cohort; (ii) determined how much of the phenotypic variance could be explained by the known variants when combined in a GRS;^{7,8} and (iii) calculated the percentage of variance explained by additive (h^2_{SNP}) and dominance variation (δ^2_{SNP}) at all common SNPs, because estimates of the narrow-sense (h^2_{SNP}) and the broad-sense heritability ($H^2_{\text{SNP}} = h^2_{\text{SNP}} + \delta^2_{\text{SNP}}$) captured by SNPs⁹ provide an upper bound to the explanatory power of genetic variants that can be discovered by GWAS.^{9–11}

MATERIALS AND METHODS

Trait and SNP selection

Our selection of traits was based on two criteria: (i) it had to be a continuous trait measured in the baseline visit of the Lifelines Cohort Study, and (ii) a meta-GWAS on the trait including at least 10 000 European individuals had to have been published. We searched the GWAS catalog^{1,2} for original papers describing or meta-analyzing GWASs of the selected traits (date: 01/14/2015). In addition, we performed a literature search to identify papers that used a gene-centric genotyping platform for association analysis (Figure 1). From the resulting list of publications, we selected the paper(s) with the largest sample size of European individuals (at least 10 000 individuals). This led to selection of multiple papers for several traits, if the sample sizes of these papers were similarly large, or there were additional papers using a gene-centric array (Table 1). From these papers we identified index SNPs that were significantly associated with the phenotypes of interest. The criterion for statistical significance depended on the genotyping platform. For analyses based on a genome-wide SNP array, the standard genome-wide significance threshold was used ($P = 5 \times 10^{-8}$). For gene-centric genotyping platforms, we used the threshold of significance from the original papers. Preferably, statistical significance was assessed based on the P -value of the combined analysis of discovery and replication samples. If a study did not include a P -value for the combined analysis, we used the P -values from the discovery phase. For each selected SNP, we recorded the effect size (beta + direction of effect), standard error, P -value, and effect allele, when available. For the effect sizes, we used the effect size derived from the combined analysis of discovery and replication cohorts if given. If not, we used the effect size in the replication cohort if available, and otherwise that of the discovery. If multiple papers contributed SNPs for a particular trait and they used different units, transformations or meta-analysis methods, we transformed the effect sizes of the studies to make them comparable.

Genotyping and imputation

A total of 15 638 presumably unrelated individuals of the Lifelines Cohort Study were selected for genome-wide genotyping on the Illumina CytoSNP 12 v2 array and called in GenomeStudio (Illumina, San Diego, CA, USA). A pre-imputation quality control was carried out in PLINK.¹² SNPs with a call rate <95%, Hardy–Weinberg equilibrium P -value <0.0001, or minor allele frequency <1% were excluded, as were samples with a sex mismatch, deviating heterozygosity (>4 s.d. from mean), non-European ancestry, a call rate <95%, or that were duplicates or first-degree relatives. A total of 268 407 SNPs and 13 436 samples remained after quality control. Imputation was carried out using Beagle v3.1.0^(ref. 13) with the HapMap Phase 2 CEU reference panel (release 24, build 36).¹⁴ To determine the genetic relationship matrices needed for the common-SNP heritability estimation (see below) another imputation was performed using IMPUTE2^(ref. 15) with the HapMap Phase 3 reference panel (release 2, build 36), as the HapMap Phase 3 SNP set was optimized to capture common genetic variation in the human genome.¹⁶ This latter imputed dataset was converted to best-guess genotypes: the most likely genotype was assigned for each SNP and for each individual when it had a probability >0.9, otherwise the genotype was set to missing for that individual. After conversion, SNPs with a call rate <0.95 were excluded. This yielded a set of 874 760 SNPs.

Statistical analysis

Correction for the Lifelines' effect size. To obtain unbiased estimates, the validation cohort (ie, Lifelines) needs to be completely independent of the discovery sample.⁷ As some of the selected papers included Lifelines data in their meta-analysis, we corrected their results for the 'Lifelines effect' by recalculating the SNP's effect sizes (betas) and standard errors (SEs) using inverse versions of the formula for an inverse-variance fixed-effects meta-analysis:

$$\beta_{\text{corrected}} = \frac{\left(\frac{1}{\text{SE}_{\text{meta}}^2} \cdot \beta_{\text{meta}}\right) - \left(\frac{1}{\text{SE}_{\text{Lifelines}}^2} \cdot \beta_{\text{Lifelines}}\right)}{\frac{1}{\text{SE}_{\text{meta}}^2} - \frac{1}{\text{SE}_{\text{Lifelines}}^2}}$$

and

$$\text{SE}_{\text{corrected}} = \sqrt{\frac{1}{\frac{1}{\text{SE}_{\text{meta}}^2} - \frac{1}{\text{SE}_{\text{Lifelines}}^2}}}$$

where β_{meta} and SE_{meta} are the beta and SE of the SNP in the original meta-GWAS paper and $\beta_{\text{Lifelines}}$ and $\text{SE}_{\text{Lifelines}}$ are the beta and SE in the Lifelines Cohort, which have been provided to the meta-GWAS consortium. With the corrected beta and SE, we recalculated the P -value. If a SNP no longer met the significance threshold of the original paper, it was excluded.

Only independently associated SNPs were included in our study. If a study did not test for independence between identified SNPs at a locus for the same trait, or if SNPs from multiple studies were included for a single trait, we calculated linkage disequilibrium between the SNPs in the Lifelines dataset. SNP pairs with an $r^2 < 0.1$ were considered to be independent. If multiple dependent SNPs in a locus were reported in a single paper, we selected the most significant one. When multiple papers reported different, dependent SNPs in the same locus for the same trait, we selected the SNP from the study with the largest sample size.

Association analysis. Each trait was associated with the individual SNPs selected for that trait, as well as with two GRSs: the unweighted GRS, ie, the sum of the risk allele dosages of selected SNPs; and the weighted GRS, ie, the sum of the risk allele dosages of selected SNPs weighted by the corresponding effect size from the literature (if applicable corrected for the Lifelines' effect size). To determine whether the inclusion of SNPs with low imputation quality in the GRSs contribute to explained variance, we repeated the analyses for both unweighted and weighted GRS using only SNPs imputed with high quality ($R^2 > 0.5$) in Lifelines. Association was tested using linear regression in R (rms package v4.2-0).^{17,18} For each of the traits, we used the same unit, transformation, exclusion criteria and covariates as described in the original papers (Table 1) to achieve exact replication. Details on the trait measurements, which were all based on either the physical examination or biomaterial collection at the baseline visit of Lifelines, have been described previously.⁴ Ten principal components were added as covariates to correct for population stratification. The genome-wide significant SNPs from the literature were considered replicated in Lifelines if they showed a one-sided P -value <0.05 and the same direction of effect. We chose a significance threshold of 0.05 because the selected SNPs are firmly established associated variants and hence no multiple testing correction was applied. To determine the percentage of variance explained (R^2) by the GRS, we compared nested models (including covariates) with and without the GRS and calculated the difference in R^2 between them. The GRS R^2 values used in the remainder of the paper refer to this difference in R^2 .

Common-SNP heritability. Genomic-relatedness-matrix restricted maximum likelihood was performed using the genome-wide complex trait analysis (GCTA) software package¹¹ to determine the percentage of variance that can be attributed to common SNPs, ie, the common-SNP heritability. Only unrelated individuals (estimated pairwise relationship <0.05) were included in this analysis (max $N = 10\,234$). We also estimated the dominance component using a recent extension of GREML.⁹ The variance explained by additive and dominance variation at all common SNPs are defined as $h^2_{\text{SNP}} = \sigma^2_A / (\sigma^2_A + \sigma^2_D + \sigma^2_e)$ and $\delta^2_{\text{SNP}} = \sigma^2_D / (\sigma^2_A + \sigma^2_D + \sigma^2_e)$, respectively, where h^2_{SNP} is interpreted as the narrow-sense heritability captured by SNPs and $H^2_{\text{SNP}} = h^2_{\text{SNP}} + \delta^2_{\text{SNP}}$ as the broad-sense heritability captured by SNPs. Please note that H^2_{SNP} as defined here does not include an epistatic component, which is expected to be small.⁹

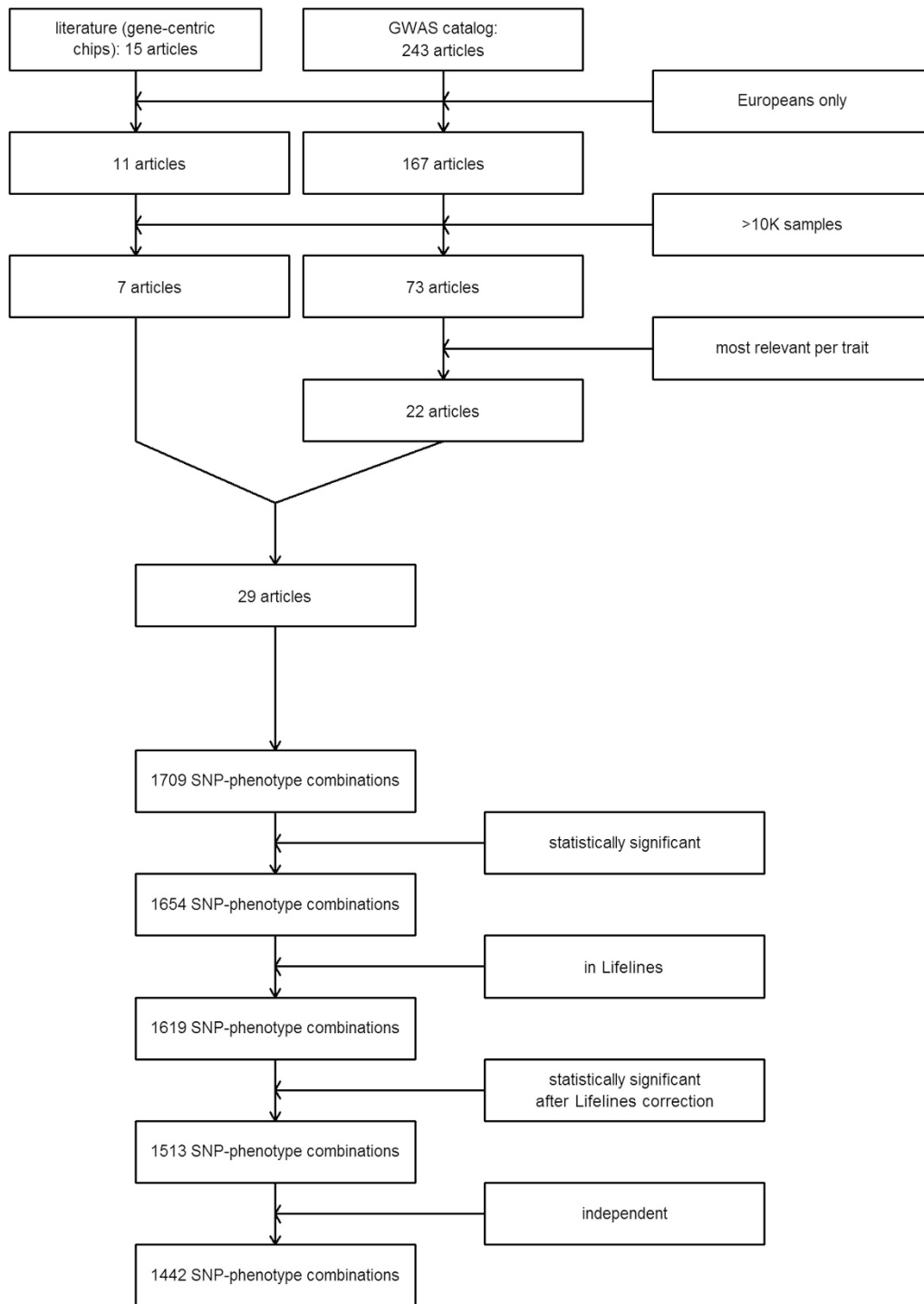


Figure 1 Flow diagram showing the paper and SNP selection process of known SNP-phenotype associations for the 32 traits assessed in this study. For some traits multiple papers were selected (Table 1) if sample sizes of several GWAS papers were similarly large, or there were additional papers using a gene-centric array.

RESULTS

Trait selection

Thirty-two traits fulfilled our criteria, representing five main systems or disease areas that are the focus of Lifelines: musculoskeletal, cardiovascular and renal, metabolic, hematologic and inflammatory, and pulmonary function. Descriptive statistics for the 32 traits and

demographic variables in our cohort are shown in Supplementary Table 1.

SNP selection

Figure 1 shows the results of the paper and SNP selection from the literature for the 32 traits of interest. From the GWAS catalog^{2,19} and

Table 1 List of the 32 selected continuous disease traits in Lifelines from five broad disease areas and details on their transformations, covariates and exclusions used for genetic analysis

Trait (unit)	Abbreviation	Transform	Covariates	Excl.	References
<i>Anthropometrics</i>					
Body mass index (kg/m ²)	BMI	INR	Sex, age, age ²		30 ^a ,31
Height (cm)	Height	Z-scores	Sex, age		22 ^a ,32
Waist-hip ratio	WHRadjBMI	INR	Sex, age, BMI		33 ^a ,34
<i>Cardiovascular and renal</i>					
Systolic blood pressure (mmHg)	SBP	b	Sex, age, age ² , BMI	c	35,36 ^a ,37 ^a
Diastolic blood pressure (mmHg)	DBP	b	Sex, age, age ² , BMI	c	35,36 ^a ,37 ^a
Heart rate (bpm)	HR		Sex, age, age ² , BMI	d	38 ^a
PR interval (ms)	PR		Sex, age, height, BMI, HR, SBP	e	39
QRS interval (ms)	QRS		Sex, age, height, BMI	f	40
QT interval (ms)	QT		Sex, age, HR	g	41 ^a
Serum creatinine (μmol/l)	Creat	Log ₁₀	Sex, age		42
Estimated glomerular filtration rate (creatinine, ml/min per 1.73 m ²)	eGFRcrea	ln	Sex, age		43
Urinary albumin/creatinine ratio (mg/g)	UACR	ln	Sex, age		44
Serum urate (mg/dl)	Urate		Sex, age		45 ^a
<i>Metabolic</i>					
Alkaline phosphatase (IU/l)	ALP	Log ₁₀	Sex, age		46
Alanine transaminase (IU/l)	ALT	Log ₁₀	Sex, age		46
Gamma-glutamyl transferase (IU/l)	GGT	Log ₁₀	Sex, age		46
Fasting glucose (mmol/l)	FG		Sex, age	h	47,48
Glycated hemoglobin (%)	HbA1c		Sex, age	i	49
HDL cholesterol (mg/dl)	HDL	INR	Sex, age, age ²	LLD	50
LDL cholesterol (mg/dl)	LDL	INR	Sex, age, age ²	LLD	50
Total cholesterol (mg/dl)	TC	INR	Sex, age, age ²	LLD	50
Triglycerides (mg/dl)	TG	INR	Sex, age, age ²	LLD	50
<i>Hematology and inflammation</i>					
C-reactive protein (mg/l)	CRP	ln	Sex, age		51
Hemoglobin (g/dl)	Hb		Sex	> 3 s.d.	52 ^a
Mean corpuscular volume (fl)	MCV		Sex	> 3 s.d.	52 ^a
Mean corpuscular hemoglobin (pg)	MCH		Sex	> 3 s.d.	52 ^a
Red blood cell count (10 ¹² /l)	RBC		Sex	> 3 s.d.	52 ^a
White blood cell count (10 ⁹ /ml)	WBC		Sex, age	> 3 s.d.	53
Platelet count (10 ⁹ /l)	PLT		Sex, age	> 6 s.d.	54 ^a ,55
<i>Lung function</i>					
Forced expiratory volume in 1 s (ml)	FEV1	INR	Sex, age, age ² , height, smoking		56
Forced vital capacity (ml)	FVC		age, age ² , sex, height, height ² , weight, smoking		57 ^a
FEV1/forced vital capacity	FEV1/FVC	INR	Sex, age, age ² , height, smoking		56

Abbreviations: Excl., exclusions; ln, natural logarithm; INR, inverse normal of residuals; LLD, lipid-lowering drugs (ATC C10); s.d., standard deviation; s, second.

^aResults of these papers were corrected for the effect of Lifelines for the current study, because Lifelines participated in the original meta-GWAS.

^bFor individuals who took antihypertensive or blood pressure lowering medication (ATC C02, C03, C07, C08, or C09) 10 mmHg was added to their DBP and 15 mmHg to their SBP.

^cIndividuals with a history of myocardial infarction or heart failure, coronary artery disease, a bypass or dotter treatment, or a missing DBP/SBP measurement.

^dIndividuals with atrial fibrillation (missing PR interval), a history of myocardial infarction or heart failure, or drugs of the ATC categories C01AA05, C07, and C08 (calcium-antagonists, beta-blockers, and digoxin).

^eIndividuals with atrial fibrillation (missing PR interval), a history of myocardial infarction or heart failure, use of a pacemaker, drugs of the ATC categories C01B and C01AA05 (class I and III blocking medication and digoxin), or pregnant females.

^fIndividuals with atrial fibrillation (missing PR interval), a history of myocardial infarction or heart failure, use of a pacemaker, or drugs of the ATC category C01B (class I and III blocking medication).

^gIndividuals with atrial fibrillation (missing PR interval), a high (> 120 ms) or absent QRS interval, use of a pacemaker, drugs of the ATC categories C01B, C07, C08, and C01AA (class I-IV anti-arrhythmatics and digitalis), or pregnant females.

^hIndividuals with diagnosed diabetes or receiving diabetes treatment or antidiabetic drugs (ATC A10), non-fasting samples, or pregnant females.

ⁱIndividuals with diagnosed diabetes or receiving diabetes treatment or antidiabetic drugs (ATC A10), high fasting glucose (≥7 mmol/l), non-fasting samples, or pregnant females.

We use letters instead of numbers for the footnotes, as using numbers would likely corrupt any attempt to copy data from the reference column as well as cause confusion for the exponential values used in the units and covariates columns.

the successive literature search, we identified 243 and 15 papers using genome-wide and gene-centric genotyping platforms, respectively. After filtering for ethnicity, sample size, and relevance/suitability, 29 papers (of which 18 used GWAS data, 7 used gene-centric chip

data, and 4 used a combination) were selected as sources of known SNPs for our 32 traits. From these papers, we identified 1709 SNP-phenotype associations. A final number of 1442 index SNP-phenotype associations were included in our analyses (Figure 1; Supplementary

Table 2) after exclusion of the following associations: 55 were not statistically significant according to our criteria, for 35 the respective SNP was not present in the Lifelines HapMap Phase 2 imputed data, 106 lost statistical significance after correcting the meta-GWAS results for the effect size of Lifelines because Lifelines was part of the meta-GWAS (see Materials and Methods), and 73 were in linkage disequilibrium (LD) with another SNP in the SNP list. Some index SNPs were associated with multiple traits; the number of SNPs included in our analyses was 1307, of which 967 (74%; accounting for 1057 SNP-phenotype combinations) had a high imputation quality ($R^2 > 0.5$) in the Lifelines Cohort.

Direct SNP replication

Of the 1442 index SNP-phenotype associations that were tested, 865 (60%; median per trait = 75%, interquartile range = 59.8–88.2%) reached statistical significance (Table 2; Supplementary Table 2). When considering only high-quality imputed SNPs, the replication rate increased to 66.2% (700/1057; median = 83.8%, interquartile

range 63.6–98%). None of the SNPs had a significant effect in the opposite direction, if we would have used a two-sided test. Furthermore, of the non-replicated SNP the direction of effects were highly consistent with the directions from literature. A median of 86.1% (interquartile range 75–100%) of all non-replicated SNPs per trait showed a direction of effect that was consistent with the literature (100% (85.6–100%) for high-quality SNPs; Supplementary Table 3). The replication rates decreased with increasing sample size of the GWAS discovery from which the SNPs were selected, whereas the percentages of non-significant SNPs with a consistent direction of effect increased (Supplementary Figure 1). The Lifelines' effect sizes correlated well with those from the literature (if applicable, after correction for the Lifelines' effect size; Supplementary Figure 2).

Genetic risk score analysis

The number of index SNPs incorporated in the GRSs ranged from 0 (for UACR) to 476 (for height) when only high-quality SNPs were used, and from 1 (for UACR) to 635 (for height), when all SNPs were used (Table 3; Supplementary Table 4). All GRSs were significantly associated with their respective traits. In general, inclusion of low-quality SNPs in the GRS resulted in an increase in phenotypic variance explained by the GRS. For the weighted GRS built from all SNPs, the median of the relative increase was 10.5% (interquartile range = 3.9–19.9%) and for the unweighted GRS it was 11.7% (3.4–20%) compared the GRSs based on only high-quality SNPs.

The most significant GRS was the weighted GRS constructed from all SNPs for height ($P < 1 \times 10^{-320}$), which explained 15.52% of the phenotypic variance. The percentages of explained variance for the other traits ranged between 0.02% (FVC) and 6.67% (HDL). For all but four traits, the weighted GRS model was more significant and explained more phenotypic variance than the unweighted model (Table 3). For the exceptions (BMI-adjusted WHR, heart rate, HbA1c, and FVC) the numbers of SNPs included in the GRSs were small and/or the percentage of variance explained was low.

Common-SNP heritability

The broad-sense common-SNP heritability estimates were statistically significant for all 32 traits except for body mass index-adjusted waist-hip ratio (Table 3). The percentage of phenotypic variance that could be explained by the additive effect of common SNPs was highest for height (48.9%) and ranged from 5.6 to 39.2% for the other traits. The dominance effects, on the other hand, were uniformly small (0–1.5%) and not significant.

DISCUSSION

In this study, we investigated in over 13 000 participants of the Lifelines Cohort Study to which extent the heritability of 32 complex traits can be explained by previously reported, genome-wide significantly associated SNPs. We first demonstrated that the majority of previously reported SNP-phenotype associations (median = 75%) could be replicated and that there was high enrichment of effects in the right direction for the non-replicated ones (median = 86.1%), indicating that power was likely insufficient for those SNPs. These percentages increased to 83.8 and 100%, respectively, when only high-quality SNPs were considered. Second, all unweighted and weighted GRSs combining the information of these SNPs were significantly associated with their respective traits, with weighted GRSs generally explaining more phenotypic variance than the unweighted GRS. Inclusion of poorly-imputed SNPs in GRSs in general still contributed

Table 2 Numbers and percentages of previously reported SNPs that were statistically significantly associated in Lifelines with the trait of interest

Trait	n	All SNPs		High-quality SNPs	
		n	P < 0.05 (%)	n	P < 0.05 (%)
BMI	77	41 (53.2)	63	38 (60.3)	
Height	635	327 (51.5)	476	275 (57.8)	
WHRadjBMI	51	15 (29.4)	40	14 (35)	
HR	15	9 (60)	10	8 (80)	
DBP	32	22 (68.8)	27	18 (66.7)	
SBP	33	21 (63.6)	26	17 (65.4)	
QT	33	31 (93.9)	25	24 (96)	
PR	9	8 (88.9)	7	7 (100)	
QRS	25	22 (88)	17	15 (88.2)	
eGFRcrea	52	34 (65.4)	33	26 (78.8)	
Creat	4	4 (100)	3	3 (100)	
UACR	1	1 (100)	0		
urate	28	26 (92.9)	20	19 (95)	
ALP	14	14 (100)	11	11 (100)	
ALT	4	3 (75)	2	1 (50)	
FG	31	25 (80.6)	23	23 (100)	
GGT	26	22 (84.6)	20	18 (90)	
HbA1c	9	8 (88.9)	3	3 (100)	
HDL	69	42 (60.9)	48	34 (70.8)	
LDL	54	30 (55.6)	34	21 (61.8)	
TC	71	42 (59.2)	48	30 (62.5)	
TG	37	18 (48.6)	25	14 (56)	
CRP	17	13 (76.5)	14	10 (71.4)	
PLT	52	39 (75)	32	28 (87.5)	
WBC	2	1 (50)	1	1 (100)	
Hb	8	5 (62.5)	5	4 (80)	
MCH	15	13 (86.7)	11	11 (100)	
MCV	12	11 (91.7)	11	10 (90.9)	
RBC	7	6 (85.7)	6	6 (100)	
FEV1	4	3 (75)	4	3 (75)	
FVC	5	1 (20)	3	1 (33.3)	
FEV1/FVC	10	7 (70)	9	7 (77.8)	
Total	1442	865 (60)	1057	700 (66.2)	
Median % (IQR)		75 (59.8–88.2)		83.8 (63.6–98)	

Abbreviations: IQR, interquartile range; SNP, single nucleotide polymorphism.

Table 3 Estimates of the explained variances by unweighted and weighted GRSs and the additive (h^2_{SNP}) and dominant (δ^2_{SNP}) variance components captured by common SNPs

Trait	N	SNPs	Unweighted GRS		Weighted GRS		N	GREML-GCTA					
			P	R ² (%)	P	R ² (%)		h ² _{SNP} (%)	h ² _{SNP} SE (%)	P h ² _{SNP}	δ ² _{SNP} (%)	δ ² _{SNP} SE (%)	P δ ² _{SNP}
BMI	13 230	77	2.3E-55	1.81	9.9E-66	2.16	10 233	24.8	3.2	9E-15	1.5	0.9	0.10
Height	13 231	635	<1E-320	13.44	<1E-320	15.52	10 234	48.9	3.2	4.4E-52	0	0.5	1
WHRadjBMI	13 228	51	1.3E-16	0.51	2.9E-16	0.49	10 231	5.6	3.1	6.8E-02	0.1	0.3	0.72
SBP	12 842	33	2.6E-33	0.83	2.9E-35	0.88	9946	17.3	3.2	8E-08	0.6	0.6	0.39
DBP	12 842	32	2.5E-32	0.89	3.1E-33	0.92	9946	17	3.3	2.6E-07	0.3	0.5	0.56
HR	11 477	15	5E-22	0.77	1.5E-21	0.75	8900	9.8	3.6	6.1E-03	0	0.4	1
PR	12 832	9	3.2E-70	2.09	1.4E-85	2.55	9941	16.3	3.2	5.2E-07	0	0.3	1
QRS	12 539	25	9.7E-85	2.40	4.8E-104	2.94	9697	22.5	3.3	1.4E-11	0	0.6	1
QT	11 640	33	4.5E-120	1.62	2.1E-180	2.39	9023	31.7	3.6	1.5E-18	0.1	0.4	0.74
Creat	13 191	4	1.4E-10	0.20	2.1E-11	0.21	10 199	26.8	3.2	6.1E-17	1	0.6	0.10
eGFRcrea	13 191	52	3.4E-64	1.74	6.9E-74	2.01	10 199	27.3	3.2	1.7E-17	1.1	0.6	0.09
UACR	13 152	1	1.9E-05	0.12	1.9E-05	0.12	10 168	7.9	3.1	1.1E-02	0	0.5	1
urate	13 226	28	2.8E-104	2.29	1.2E-210	4.52	10 229	34.5	3.2	3.4E-27	0	0.4	1
ALP	13 227	14	1.5E-129	4.03	9.8E-182	5.58	10 230	22.1	3.2	4E-12	0.7	0.7	0.27
ALT	13 227	4	2.8E-09	0.21	2.6E-12	0.29	10 230	12.1	3.1	1.1E-04	1.2	0.7	0.10
GGT	13 226	26	8.2E-61	1.63	1.2E-72	1.95	10 229	19.7	3.2	6.4E-10	0.3	0.5	0.53
FG	11 954	31	2.1E-86	2.75	5.3E-103	3.28	9244	11.2	3.5	1.6E-03	0.5	0.6	0.41
HbA1c	11 786	9	2E-38	1.17	1.1E-35	1.09	9111	14	3.5	7.3E-05	0	0.4	1
HDL	12 191	69	5.3E-98	3.49	1.1E-191	6.67	9453	18.6	3.4	5E-08	0.3	0.6	0.56
LDL	12 192	54	2.6E-71	2.54	1.3E-133	4.72	9453	26.9	3.5	2.9E-14	0	0.5	0.98
TC	12 192	71	3.6E-86	3.07	3.3E-140	4.95	9453	23.4	3.5	2.2E-11	0	0.5	1
TG	12 164	37	5.8E-52	1.85	5.9E-94	3.35	9428	19.1	3.5	3.8E-08	1.3	0.8	0.11
CRP	12 838	17	1.9E-58	1.94	2.8E-96	3.21	9940	18.9	3.2	6E-09	0.8	0.6	0.20
Hb	13 115	8	1.3E-18	0.33	4.9E-19	0.33	10 147	19	3.2	3.1E-09	0	0.3	1
MCV	12 792	12	6.4E-59	1.98	1.1E-67	2.28	9917	26.8	3.4	1.8E-15	0.1	0.4	0.77
MCH	12 733	15	1.4E-89	2.98	4.3E-102	3.39	9871	23.9	3.4	1.3E-12	0.2	0.4	0.63
RBC	12 876	7	4.5E-17	0.35	6E-18	0.37	9979	24.7	3.3	5.3E-14	0.3	0.4	0.42
WBC	13 114	2	5.8E-11	0.31	4E-13	0.39	10 150	25.2	3.2	2.5E-15	0.8	0.6	0.20
PLT	13 162	52	1.9E-84	2.55	2.5E-98	2.96	10 183	39.2	3.3	5.3E-33	0	0.3	1
FEV1	12 401	4	4.1E-07	0.20	1.6E-07	0.21	9586	25.2	3.4	2.1E-13	0	0.4	1
FVC	12 400	5	2E-03	0.02	2.1E-03	0.02	9585	24.2	3.4	1.5E-12	0	0.3	1
FEV1/FVC	12 401	10	1.5E-14	0.46	1.4E-14	0.47	9586	20.8	3.4	9.6E-10	0.5	0.5	0.33

Abbreviations: P, P-value; R², the percentage of explained phenotypic variance attributable to the GRS.

to the variance explained, advocating for the use of all known SNPs and not only high-quality ones when constructing GRSs. The total variance explained by the weighted GRSs constructed from all SNPs was 15.52% for height and between 0.02 and 6.67% for the other traits. The additive genetic variance at all common SNPs explained a significant proportion of the phenotypic variance for all traits ranging from 5.6% (body mass index-adjusted waist-hip ratio) to 48.9% (height), but none of the traits showed significant dominance genetic variance.

As mentioned above we could not replicate all SNPs that were previously identified in large meta-GWASs. This is likely due to reduced power resulting from a sample size of 'only' 13 000 individuals in comparison to the much larger GWAS discovery samples. Power analysis shows that our study had sufficient power (>80%) to detect effect sizes of 0.04 s.d. for very common SNPs (allele frequency >20%), 0.08 s.d. for SNPs with an allele frequency between 5 and 20%, and 0.16 s.d. for SNPs with an allele frequency between 1 and 5% (Supplementary Figure 3). As expected, SNPs extracted from smaller GWASs were more likely to be replicated in Lifelines as in those GWASs only SNPs with larger effect sizes could be discovered.

Conversely, SNPs from smaller GWASs more often had an effect in the opposite direction, indicating that effect sizes from the large GWASs could be estimated more accurately.

As far as we know, we are the first to determine how much of the phenotypic variance is explained by both known and common SNPs to measure the current heritability gap for a large number of complex disease traits in one homogeneous population. However, explained variances of GRSs and common-SNP heritabilities for individual traits have been estimated before, and our overall results are consistent with this literature. For instance, various studies estimated the common-SNP heritability of height to be 40–50%,^{11,20,21} which matches our finding of 48.9%. The study from which most height-associated SNPs were extracted²² reported that genome-wide significantly associated SNPs explained 16% of the variance in height, which is comparable to our GRS result (15.5%). Only one trait (LDL cholesterol concentration) differed considerably from the literature: we found a common-SNP heritability of 27%, whereas a previous study in the Icelandic population found only 10%.²¹ Zaitlen *et al.* used a slightly different method, which could have caused the difference, but it might also reflect a population specific effect.

In our study, the dominance effect of common SNPs was not significant for any trait. This is consistent with one earlier study that was unable to detect any replicable dominance effects for 79 quantitative complex traits using data from three large European cohorts including Lifelines as a replication sample.⁹

Although all our GRSs composed of known SNPs were significantly associated with their respective outcomes, they explain only a fraction of the total common-SNP heritability of these complex traits. On average, the variance explained by the weighted GRS accounted for only 10.7% of the common-SNP heritability. Only for height (15.5%/48.9% = 31.7%), alanine transaminase (24.5%), fasting glucose (28%), and HDL cholesterol (35.3%), known SNPs accounted for a considerable part of the common-SNP heritability. Our data thus confirms what has been found previously for height, BMI, and QT interval,^{10,23} but now extends these observations to many other complex disease traits. There are a number of potential causes for the large gaps between the common-SNP heritability and the part that can be explained by all identified SNPs. It may be due to errors in the estimation of SNP effects, but it is likely also due to the way in which SNPs are selected for a GRS. Usually the selection of SNPs is restricted to genome-wide significant SNPs only. This is likely too conservative, as low power will result in SNPs with a small effect or low allele frequency not reaching the genome-wide significance threshold. Prediction accuracy will probably increase when less stringent significance thresholds for the selection of SNPs are applied. Polygenic risk score analyses could be applied using various significance thresholds to determine the percentage of variance explained by larger number of SNPs. We should also keep in mind that the markers detected with GWAS are not likely to be the causal variants, but merely in LD with them. As the SNPs on GWAS chips are selected because they are common variants, and low-frequency variants are in low LD with them, power to detect rare causal variants is limited.⁷

Some investigators have argued that a considerable part of the missing heritability may be caused by non-additive effects such as dominance and epistasis.^{9,24} As such, an important result of our analyses, also confirmed by Zhu *et al.*,⁹ was that we were unable to find strong dominance effects. This is further supported by the findings of Zaitlen *et al.*,²¹ who developed a method to include both closely and distantly related individuals, and used this to analyze the inflation of narrow-sense heritability. Their findings supported neither dominance nor epistatic effects, and they attributed the inflation of narrow-sense heritability to shared environmental factors in datasets using close relatives. Epigenetic variation including methylation profiles have been suggested as another source of missing heritability, but a recent paper including samples from Lifelines found that methylation and genetic predictors for BMI did not overlap, suggesting that the former represent environmental effects on this trait. For height, methylation profiles did not explain any variation.²⁵

In the current study, we paid careful attention to the construction of the GRS to determine the percentage of variance explained by known SNPs. The advantage of GRS is that it is conservative, as it only uses verified markers, and that its results are fairly robust. However, a number of issues need to be considered when using a GRS (see also the review by Wray *et al.*⁷). First, there should be no overlap in the samples from which the SNPs were selected and to which the GRS is applied, as this would yield overestimates of the percentage of variance explained. Consequently, for meta-GWAS that included Lifelines, we adjusted the SNP effect sizes by

subtracting the Lifelines' effect size, and excluded the SNPs that were no longer genome-wide significant after this correction. Second, if the replication sample is more closely related to the discovery population than to the target population, or if population stratification patterns of the discovery and replication samples are similar, the prediction accuracy will be overestimated.⁷ For the selection of SNPs we included only papers that used European individuals, to match the ethnicity of Lifelines and increase the probability that the SNPs replicated with a similar effect size. As a consequence our results may be less applicable to other ethnicities and percentages of explained variance are likely to be lower in non-European cohorts. Third, all SNPs included in the GRS need to be independent, because otherwise regions with high LD will contribute more to the GRS than regions with low or no LD, leading to biased results. This check is mainly important when multiple meta-GWAS and/or gene-centric studies are used for the selection of SNPs. For this reason we excluded 71 (4.2%) dependent SNPs from our original SNP set. Fourth, it is advised to check that the effect direction of all individual SNPs in your cohort matches that of the literature, otherwise effects of misaligned and correctly aligned SNPs might cancel each other out in the GRS and the GRS might turn out to be insignificant. As a check on correct modeling of the GRS, when using the original effect sizes as weights, and assuming that the effect sizes in your cohort are similar to those of the meta-GWAS, the regression coefficient of the weighted GRS model should be around 1.

One limitation of our approach is that we focussed on heritability explained by common SNPs only, which is estimated to cover one third to one half of the total heritability found in twin and family studies,^{23,26} with the remaining genetic variance most likely explained by lower frequency variants.^{27,28} As such, our conclusions are limited to the heritability gap for only common SNPs as typically targeted by GWAS. The new method developed by Zaitlen *et al.*²¹ including relatives allows separate estimation of both common-SNP and total heritability. However, this latter estimate is potentially confounded by shared environment. With the increasing availability of whole-genome sequencing data expected in the near future, the recently developed expansion by Yang *et al.*²⁸ of their GREML method, which stratifies for LD and minor allele frequency (GREML-LDMS), may be a good alternative for the estimation of heritability of complex traits as it captures contributions of both rare and common variants. A similar method using both rare and common SNPs was recently applied to 9 common diseases in data from the UK Biobank and explained an average of 57.3% of their total (narrow-sense) heritability based on structural equation modeling.²⁹

In conclusion, we demonstrated that the majority of previously reported SNP associations for 32 continuous disease traits could be replicated in the Lifelines Study Cohort, confirming Lifelines' value and reliability as a resource for genetic epidemiological studies. Although meta-GWAS studies have identified many SNPs that are associated with complex disease traits, these SNPs explain only a small to moderate part of the common-SNP heritability, which in turn explains up to ~50% of the total heritability. Our data suggest that dominance effects are unlikely to fill the gap of the missing heritability. Overall our results showed that none of the GRSs of complex disease traits are sufficiently accurate for personalized prediction limiting successful applications in clinical practice.

DATA AVAILABILITY

The data that support the findings of this study are available from the Lifelines Cohort Study (<https://lifelines.nl/lifelines-research/general>) but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Genotype and phenotype data are available upon request from the Lifelines Cohort, whereas the effect sizes used to correct for the Lifelines effect are available from the authors on reasonable request and with permission from Lifelines.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We would like to acknowledge the services of the Lifelines Cohort Study, the contributing research centers delivering data to Lifelines, and all the study participants. Part of the statistical analyses were carried out on the Genetic Cluster Computer (<http://www.geneticcluster.org>), which is financially supported by the Netherlands Scientific Organization (NWO 480-05-003) along with a supplement from the Dutch Brain Foundation. And informed consent was obtained from all individual participants included in the study.

DISCLAIMER

The Lifelines study was approved by the medical ethics committee of the University Medical Center Groningen and conducted in accordance to the Helsinki Declaration Guidelines.

- 1 Hindorf LA, Sethupathy P, Junkins HA *et al*: Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci USA* 2009; **106**: 9362–9367.
- 2 Welter D, MacArthur J, Morales J *et al*: The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* 2014; **42**: D1001–D1006.
- 3 Stolk RP, Rosmalen JGM, Postma DS *et al*: Universal risk factors for multifactorial diseases-lifeLines: a three-generation population-based study. *Eur J Epidemiol* 2008; **23**: 67–74.
- 4 Scholtens S, Smidt N, Swertz MA *et al*: Cohort profile: LifeLines, a three-generation cohort study and biobank. *Int J Epidemiol* 2015; **44**: 1172–1180.
- 5 Klijs B, Scholtens S, Mandemakers JJ, Snieder H, Stolk RP, Smidt N: Representativeness of the Lifelines Cohort Study. *PLoS ONE* 2015; **10**: e0137203.
- 6 Meems LMG, de Borst MH, Postma DS *et al*: Low levels of vitamin D are associated with multimorbidity: results from the Lifelines Cohort Study. *Ann Med* 2015; **47**: 474–481.
- 7 Wray NR, Yang J, Hayes BJ, Price AL, Goddard ME, Visscher PM: Pitfalls of predicting complex traits from SNPs. *Nat Rev Genet* 2013; **14**: 507–515.
- 8 Jamshidi Y, Nolte IM, Spector TD, Snieder H: Novel genes for QTc interval. How much heritability is explained, and how much is left to find? *Genome Med* 2010; **2**: 35.
- 9 Zhu Z, Bakshi A, Vinkhuyzen AAE *et al*: Dominance genetic variation contributes little to the missing heritability for human complex traits. *Am J Hum Genet* 2015; **96**: 377–385.
- 10 Yang J, Benyamin B, McEvoy BP *et al*: Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* 2010; **42**: 565–569.
- 11 Yang J, Lee SH, Goddard ME, Visscher PM: GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* 2011; **88**: 76–82.
- 12 Purcell S, Neale B, Todd-Brown K *et al*: PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007; **81**: 559–575.
- 13 Browning BL, Browning SR: A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet* 2009; **84**: 210–223.
- 14 Gibbs RA, Belmont JW, Hardenbol P *et al*: The international HapMap project. *Nature* 2003; **426**: 789–796.
- 15 Howie BN, Donnelly P, Marchini J: A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics* 2009; **5**: e1000529.
- 16 Altshuler DM, Gibbs RA, Peltonen L *et al*: Integrating common and rare genetic variation in diverse human populations. *Nature* 2010; **467**: 52–58.
- 17 Harrell FE Jr: rms: Regression Modeling Strategies. R Package Version 4.3-0. Available at: <http://CRAN.R-project.org/package=rms>. 2014.
- 18 R Core Team: *R: A Language and Environment for Statistical Computing*. Foundation for Statistical Computing: Vienna, Austria, 2014.
- 19 Hindorf LA, MacArthur J, Morales J *et al*: *A Catalog of Published Genome-Wide Association Studies* 2014.
- 20 Vattikuti S, Guo J, Chow CC: Heritability and genetic correlations explained by common SNPs for metabolic syndrome traits. *PLoS Genet* 2012; **8**: e1002637.
- 21 Zaitlen N, Kraft P, Patterson N *et al*: Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. *PLoS Genet* 2013; **9**: e1003520.
- 22 Wood AR, Esko T, Yang J *et al*: Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet* 2014; **46**: 1173–1186.
- 23 Yang J, Manolio TA, Pasquale LR *et al*: Genome partitioning of genetic variation for complex traits using common SNPs. *Nat Genet* 2011; **43**: 519–525.
- 24 Zuk O, Hechter E, Sunyaev SR, Lander ES: The mystery of missing heritability: genetic interactions create phantom heritability. *Proc Natl Acad Sci USA* 2012; **109**: 1193–1198.
- 25 Shah S, Bonder MJ, Marioni RE *et al*: Improving phenotypic prediction by combining genetic and epigenetic associations. *Am J Hum Genet* 2015; **97**: 75–85.
- 26 Polderman TJC, Benyamin B, de Leeuw CA *et al*: Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nat Genet* 2015; **47**: 702–709.
- 27 Visscher PM, Brown MA, McCarthy MI, Yang J: Five Years of GWAS Discovery. *Am J Hum Genet* 2012; **90**: 7–24.
- 28 Yang J, Bakshi A, Zhu Z *et al*: Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat Genet* 2015; **47**: 1114–1120.
- 29 Muñoz M, Pong-Wong R, Canela-Xandri O, Rawlik K, Haley CS, Tenesa A: Evaluating the contribution of genetics and familial shared environment to common disease using the UK Biobank. *Nat Genet* 2016; **48**: 980–983.
- 30 Locke AE, Kahali B, Berndt SI *et al*: Genetic studies of body mass index yield new insights for obesity biology. *Nature* 2015; **518**: 197–206.
- 31 Guo Y, Lanktree MB, Taylor KC *et al*: Gene-centric meta-analyses of 108 912 individuals confirm known body mass index loci and reveal three novel signals. *Hum Mol Genet* 2013; **22**: 184–201.
- 32 Lanktree MB, Guo Y, Murtaza M *et al*: Meta-analysis of dense gene-centric association studies reveals common and uncommon variants associated with height. *Am J Hum Genet* 2011; **88**: 6–18.
- 33 Shungin D, Winkler TW, Croteau-Chonka DC *et al*: New genetic loci link adipose and insulin biology to body fat distribution. *Nature* 2015; **518**: 187–196.
- 34 Yoneyama S, Guo Y, Lanktree MB *et al*: Gene-centric meta-analyses for central adiposity traits in up to 57 412 individuals of European descent confirm known loci and reveal several novel associations. *Hum Mol Genet* 2014; **23**: 2498–2510.
- 35 Ehret GB, Munroe PB, Rice KM *et al*: Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature* 2011; **478**: 103–109.
- 36 Ganesh SK, Tragante V, Guo W *et al*: Loci influencing blood pressure identified using a cardiovascular gene-centric array. *Hum Mol Genet* 2013; **22**: 1663–1678.
- 37 Tragante V, Barnes MR, Ganesh SK *et al*: Gene-centric meta-analysis in 87,736 individuals of European ancestry identifies multiple blood-pressure-related loci. *Am J Hum Genet* 2014; **94**: 349–360.
- 38 den Hoed M, Eijgelsheim M, Esko T *et al*: Identification of heart rate-associated loci and their effects on cardiac conduction and rhythm disorders. *Nat Genet* 2013; **45**: 621–631.
- 39 Pfeufer A, van Noord C, Marciante KD *et al*: Genome-wide association study of PR interval. *Nat Genet* 2010; **42**: 153–159.
- 40 Sotoodehnia N, Isaacs A, de Bakker PIW *et al*: Common variants in 22 loci are associated with QRS duration and cardiac ventricular conduction. *Nat Genet* 2010; **42**: 1068–1076.
- 41 Arking DE, Pulit SL, Crotti L *et al*: Genetic association study of QT interval highlights role for calcium signaling pathways in myocardial repolarization. *Nat Genet* 2014; **46**: 826–836.
- 42 Chambers JC, Zhang W, Lord GM *et al*: Genetic loci influencing kidney function and chronic kidney disease. *Nat Genet* 2010; **42**: 373–375.
- 43 Pattaro C, Teumer A, Gorski M *et al*: Genetic associations at 53 loci highlight cell types and biological pathways relevant for kidney function. *Nat Commun* 2016; **7**: 10023.
- 44 Boeger CA, Chen M, Tin A *et al*: CUBN is a gene locus for albuminuria. *J Am Soc Nephrol* 2011; **22**: 555–570.
- 45 Koettgen A, Albrecht E, Teumer A *et al*: Genome-wide association analyses identify 18 new loci associated with serum urate concentrations. *Nat Genet* 2013; **45**: 145–154.
- 46 Chambers JC, Zhang W, Sehmi J *et al*: Genome-wide association study identifies loci influencing concentrations of liver enzymes in plasma. *Nat Genet* 2011; **43**: 1131–1138.
- 47 Dupuis J, Langenberg C, Prokopenko I *et al*: New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nat Genet* 2010; **42**: 105–116.
- 48 Scott RA, Lagou V, Welch RP *et al*: Large-scale association analyses identify new loci influencing glycaemic traits and provide insight into the underlying biological pathways. *Nat Genet* 2012; **44**: 991–1005.
- 49 Soranzo N, Sanna S, Wheeler E *et al*: Common variants at 10 genomic loci influence hemoglobin A1c levels via glycaemic and nonglycaemic pathways. *Diabetes* 2010; **59**: 3229–3239.
- 50 Willer CJ, Schmidt EM, Sengupta S *et al*: Discovery and refinement of loci associated with lipid levels. *Nat Genet* 2013; **45**: 1274–1283.

- 51 Dehghan A, Dupuis J, Barbalic M *et al*: Meta-analysis of genome-wide association studies in >80 000 subjects identifies multiple loci for C-reactive protein levels. *Circulation* 2011; **123**: 731–738.
- 52 van der Harst P, Zhang W, Leach IM *et al*: Seventy-five genetic loci influencing the human red blood cell. *Nature* 2012; **492**: 369–375.
- 53 Nalls MA, Couper DJ, Tanaka T *et al*: Multiple loci are associated with white blood cell phenotypes. *PLoS Genet* 2011; **7**: e1002113.
- 54 Gieger C, Radhakrishnan A, Cvejic A *et al*: New gene functions in megakaryopoiesis and platelet formation. *Nature* 2011; **480**: 201–208.
- 55 Gaunt TR, Zabaneh D, Shah S *et al*: Gene-centric association signals for haemostasis and thrombosis traits identified with the HumanCVD bead chip. *Thromb Haemost* 2013; **110**: 995–1003.
- 56 Soler Artigas M, Loth DW, Wain LV *et al*: Genome-wide association and large-scale follow up identifies 16 new loci influencing lung function. *Nat Genet* 2011; **43**: 1082–1090.
- 57 Loth DW, Artigas MS, Gharib SA *et al*: Genome-wide association analysis identifies six new loci associated with forced vital capacity. *Nat Genet* 2014; **46**: 669–677.

Supplementary Information accompanies this paper on European Journal of Human Genetics website (<http://www.nature.com/ejhg>)