Novel polygenic model (multi-SNP genetic risk factor) helps to find parts of the missing heritability in blood pressure

Student	:	Xiaowen Lu
Student number:		S1802089
Project	:	Master-report for research project
Supervisors	:	Prof. Harold Snieder & Dr. Jingyuan Fu
Organization	:	Unit of Genetic Epidemiology & Bioinformatics, Department of Epidemiology, University Medical Center Groningen
Date	:	01-12-2009 to 30-07-2010

Abstract

Previous study has revealed that diastolic blood pressure (DBP) and systolic blood pressure (SBP) have a high heritability, ranging from 40% to 60%. Several genome-wide association studies have been conducted and discovered 18 associated loci. Compared with the high heritability of DBP and SBP, the proportion of variation explained by these loci is quite small (~0.1%). Researchers have proposed a hypothesis that a large number of single-nucleotide polymorphisms (SNPs) with very small effect carry a large proportion of the missing heritability. In order to testify this hypothesis, we used the polygenic model to quantify the variance explained by multi-SNPs genetic risk factor in DBP and SBP. Due to the fact that GWA result is more accessible than the individual genotype data, we adopted the approximate likelihood method proposed by Toby Johnson, which just needs the GWAS result to quantify the explained variation, instead of using the individual genotype data. First, we confirmed that the approximation multi-SNP model performs equally well as the exact multi-SNP model by using the simulation and the real data from NESDA cohort. Then we used GWA data from Global BPgen and CHARGE consortia and observed that around 0.26% phenotypic variation was contributed by ~2780 SNPs at association $P<0.05$ for both DBP and SBP, much higher than ~0.1% variation by the established association. We also quantified that DBP associated SNP can explain 0.2% variation of SBP and vice verse. This suggested the shared genetic background between DBP and SBP. The highly shared genetic background between two traits promises that we can find some common causal genes or biological pathway of DBP and SBP which underlies hypertension. Our result confirms that to some extent, part of missing heritability is own to many SNPs with small effect.

Introduction

The advent of genome-wide association (GWA) studies, which have mainly focused on common single-nucleotide polymorphisms (SNPs), has helped to discover over 500 common (minor allele frequency (MAF) >~5%) independent SNPs contributing risk to different common complex diseases (all the GWAS can be founded in the National Human Genome Research Institute (NHGRI) Catalog of Published Genome-Wide Association Studies). However, most of these significantly associated genomic variants have small effect size, and the proportion of explained variation is quite low, at best modest[3]. Indeed, most of the heritability is still unaccounted for even in the case of 'well-characterized' diseases for which upwards of 30 predisposing loci have been identified[1,2]. For example, the genome-wide association (GWA) meta-analysis of Global BPgen and CHARGE consortia has discovered 18 SNP loci which are significantly associated with blood pressure[1,2,4,5]. Comparing with the substantial heritability (30-60%) of blood pressure[6-8], these associated loci explained a very small proportion of the total variation in systolic blood pressure (SBP) and diastolic blood pressure (DBP), around 0.05-0.1%[1,2]. This can be own to the small effect sizes of individual causal risk alleles underlying blood pressure and the stringent genome-wide significance threshold, which leave many true signals undetected.

Contrary to quantifying the explained variation by a single SNP, researchers come up with an idea of constructing a genomic risk factor, including a larger number of genetic variants and quantifying the variance explained by this polygenic risk factor, which may increase the variance explained by the genetic factors[9,10]. Janssen et al.[11] have investigated predictive testing for complex diseases using multiple genes by simulation. He and his colleagues demonstrated that the portion of variance explained by genetic factors increased with the increased number of genes which were tested simultaneously. Wray and her co-worker proposed a polygenic model[12] which combines thousands of SNPs with small effect from GWA studies and apply this model to assess the prediction ability of the multi-SNP genetic factor. They conducted this analysis in bipolar disorder (BD), cornary heart disease (CHD), hypertension (HT), Crohn's disease (CD), rheumatoid arthritis (RA), type I diabetes (T1D), type II diabetes (T2D) in WTCCC[6] dataset and showed that relaxing the P-value threshold may increase the portion of explained heritability of BD, CHD, HT, CD and T2D.

The polygenic model can also help to find the shared genetic basis among different diseases. The identified loci by GWA studies also showed that many traits have shared genetic variants, as we have observed in autoimmune diseases[13] and heart-related traits[14]. However, the genetic variants that identified in GWA studies so far generally explained no larger than 5% of heritability. As most of genetic effect remained undefined, it remains unexplored that to what extent that two traits share genetic effect. Recently, Purcell et al conducted a study of schizophrenia, bipolar disorder, coronary artery diseases, Crohn's, hypertensions, rheumatoid arthritis, type 1 diabetes and type 2 diabetes[15] by using the polygenic model. They used this model to assess the shared genetic bases between Schizophrenia and psychiatric disease (i.e. bipolar disorder) and shared genetic bases between Schizophrenia and non-psychiatric diseases (i.e. coronary artery diseases, Crohn's disease, hypertensions, rheumatoid arthritis, type 1 diabetes and type 2 diabetes). They observed that this polygenetic model had better prediction than the model that only used the well-established SNPs and they found the risk effect of schizophrenia can be used to predict the risk for bipolar diseases, but not for the non-psychiatric diseases, suggesting that these two diseases share genetic origins. Highly shared genetic basis indicates the possibility to find the common and specific functional gene or biological pathway for a certain kind of disease, which may help to develop new disease therapy or medicines.

However, there is a difficulty in applying the classic polygenic model. To quantify the proportion of explained heritability, the classic polygenic model needs the genotype data from each individual. Although some studies have tried to make the genotype data available upon request, like WTCCC, dbGaP databases, the access to the genotype data in a large study, especially the study from the various consortiums that perform meta-analysis, is still a problem. Here we use the approximate likelihood method proposed by Toby Johnson to estimate the variance explained by the multi-SNP risk factor, generated from the discovery dataset, in the validation dataset, only based on GWAS results.

In this study, we firstly validated the performance of this approximate polygenic model for a quantitative trait (i.e. blood pressure). Then we assess whether the explained variance of SBP and DBP will increase by including more SNP loci of very small effect. Finally, we use this model to quantify the sharing of the genetic effect between the systolic blood pressure and diastolic blood pressures.

Materials and Methods

Classic Polygenic model

To assess the proportion of variation explained by genetic factors and the sharing of genetic effect between the traits, we first need to construct a risk score based on the genetic effects that are estimated from one study (discovery set) and assess how much variance is explained for the phenotype of interest in another study (validation set). The polygenic model assumes the additive effect of multiple independent genetic effects. Thus the genetic risk score for individual j in the validation set can be described as: $s_j = \sum_{i=1}^{n} \beta_{Di} g_{ij}$, where β_i is the relative risk of the i^{th} SNP that is estimated in the discovery set and g_{ij} is the dose of the coded allele, {0, 1, 2} at the i^{th} SNP for individual j in the validation set. For a quantitative trait, β_{Di} can be the estimated beta-value by the linear regression. For a case-control study, β_{Di} can be the log-transformed odd-ratio. Then the phenotypic variation explained by the risk score can be estimated by fitting a regression model (1):

$$y_j = \mu + as_j + e_j \quad ,$$

where y_j is the phenotype of interest for the j^{th} individual; μ is an intercept; a is the coefficient for the genetic risk score; and e_j is the residual errors that is assumed to have normal distribution $N(0, \sigma_1^2)$. For a quantitative trait, we can use linear regression model and the classical R^2 to measure the fraction of variance explained. For a binary trait, e.g. in the case-control study, we can use logistic regression model and the variance explained can be quantified by a range of different “pseudo-R^2” measures.

Approximate likelihood/Approximation of the polygenic

Combining the genetic risk score $s_j = \sum_{i=1}^{n} \beta_{Di} g_{ij}$ and the regression model (1), the regression model (1) can be written as

$$y_j = \mu + a\sum_{i=1}^{m} \beta_{Di} g_{ij} + e_j \quad (1),$$

where β_{Di} is the relative risk of the i^{th} SNP that is estimated in the discovery set; g_{ij} is the genotype of i^{th} SNP for the j^{th} individuals.

Setting $\beta_i = a\beta_{Di}$, the regression model (2) can be transformed into regression equation (3). The problem of estimating a (2) by maximum likelihood in the regression model has been changed to fitting the regression model (3) in the validation set.

$$y_j = \mu + \sum_{i=1}^{m} \beta_i g_{ij} + e_j \quad (3)$$

where β_i can be estimated $\hat{\beta}_i \sim N\left(\beta_{Vi}, s_{Vi}^2\right)$ by directly fitting the genotype y_j and genotype g_{ij}; β_{Vi} refers the estimated risk effect and s_{Vi} is the standard error that are estimated by the association study in the validate set. Thus the log-likelihood estimation of model (3) is

$$\ln L(\beta_1, \beta_2, \ldots, \beta_m) \approx C - \sum_{i=1}^{i=m} \frac{(\beta_i - \beta_{Vi})^2}{2s_{Vi}^2} = C - \sum_{i=1}^{i=m} \frac{(a\beta_{Di} - \beta_{Vi})^2}{2s_{Vi}^2} \quad (5)$$

where C is a constant that does not depend in the unknown parameter. In order to maximizes the likelihood, the a can be estimated as

$$\hat{a} \approx \frac{\sum_{i=1}^{i=m} \beta_{Di} \beta_{Vi} s_i^{-2}}{\sum_{i-1}^{i=m} \beta_{Di}^2 \; s_i^{-2}} \quad (6)$$

To estimate whether a is significantly derived from zero: H$_0$: a=0 vs H$_1$: $a = \hat{a}$, we can apply the likelihood ratio test for model (1):

$$\Delta = 2[\ln L(a = \hat{a}) - \ln L(a = 0)] = \sum_{i=1}^{i=m} \frac{\beta_{Vi}^2 - (\hat{a}\beta_{Di} - \beta_{Vi})^2}{s_i^2} \quad (7),$$

Furthermore, $\Delta = 2[\ln L(a = \hat{a}) - \ln L(a = 0)] = 2\ln\left[\left(\frac{RSS_{a=\hat{a}}}{RSS_{a=0}}\right)^{\frac{n}{2}}\right] = n\ln\left(\frac{1}{1-R^2}\right) \approx nR^2$ (8),

where R^2 refers to correlation coefficient that is defined as $R^2 = \frac{RSS_{a=0} - RSS_{a=\hat{a}}}{RSS_{a=0}}$.

Combing the equation (7) and (8) we can obtain

$$R^2 = \frac{1}{n}\sum_{i=1}^{m}\frac{\beta_{Vi}^2 - (\hat{a}\beta_{Di} - \beta_{Vi})^2}{s_i^2} \quad (9)$$

Simulation study in model validation

In order to test the performance of the approximate likelihood method, we simulated a quantitative phenotype, which has a genetic composition of 10 unlinked SNPs. The sample size of discovery set and validation set is 100. The effect size of i$^{\text{th}}$ SNP in discovery set is different from that of i$^{\text{th}}$ SNP in validation set. The simulated phenotype for discovery data and validation data is generated by using the mathematical model $y_j = \mu + \sum_{i=1}^{i=10}\beta_i g_{ij} + e_j$, where μ is the overall mean; β_i is the effect of i$^{\text{th}}$ SNP which is randomly generated for the normal distribution with mean 0 and standard deviation 1; g_{ij} is the genotype value at i$^{\text{th}}$ SNP for j$^{\text{th}}$ individual, which has value of 0, 1 or 2; e_j is the residual which is randomly and independently drawn from normal distribution with a mean of 0 and variance of 10. Then we simply regressed the simulated phenotype value on each SNP in both discovery and validation data and estimated the effect size for 10 SNPs $\beta = (\beta_1, \beta_2, \ldots \beta_{10})$, just like GWAS. After generating the GWAS result, we applied the classic polygenic model to calculate the exact variation explained by these 10 SNPs and the approximate likelihood method to calculate the approximate explained variation.

Data for Model Validation

We used the GWA data from Global Blood Pressure Genetics (Global BPgen) consortium as the discovery set, where meta-analysis was conducted for 13 cohorts of 34,433 individuals of European ancestry with SBP and DBP measurement[2] using inverse-variance weighting. We used the cohort from The Netherlands Study of Depression and Anxiety (NESDA) as the validation set which the genotype and phenotype information for 1591 samples were available. The detailed demographic of cohorts used in model validation is showed in Table 1.

Table 1 Demographic of cohorts in discovery sets and validation set

Discovery set						
Study	N	% women	Age (SD) in years	SBP (SD) in mmHg	DBP (SD) in mmHg	BMI (SD) in kg/m^2
Population-based cohorts						
BLSA	708	44	42.4 (13.2)	119.5 (15.0)	77.3 (10.2)	24.5 (3.6)
B58C – T1DGC	2,580	51	44.3 (0.3)	121.7 (15.3)	79.3 (10.5)	27.4 (4.9)
B58C - WTCCC	1,4 73	50	44.9 (0.4)	126.7 (15.2)	79.1 (10.2)	27.4 (4.7)
CoLaus	4,969	53	51.7 (9.5)	127.3 (17.4)	79.4 (10.8)	25.8 (4.6)
EPIC- Norfolk-GWAS	2,100	54	57.2 (7.8)	136.7 (19.1)	83.9 (11.9)	26.3 (3.9)
Fenland	1,401	56	45.0 (7.3)	122.8 (16.3)	75.5 (10.7)	27.1 (4.9)
InCHIANTI	562	55	56.9 (14.5)	138.4 (20.1)	81.4 (10.1)	27.1 (4.2)
KORA	1,664	51	52.5 (10.1)	133.4 (18.5)	81.8 (10.9)	27.3 (4.1)
NFBC1966	4,761	52	31*	125.2 (13.8)	77.5 (11.7)	24.6 (4.2)
SardiNIA	3,998	57	40.8 (15.3)	128.7 (28.4)	79.9 (17.3)	25.1 (4.6)
SHIP	3,310	53	45.0 (13.9)	133.1 (20.2)	83.5 (11.3)	26.9 (4.7)
SUVIMAX	1,823	60	50.5 (6.2)	120.9 (12.3)	78.0 (8.1)	23.5 (3.3)
TwinsUK	873	100	45.8 (11.9)	122.9 (15.4)	78.2 (10.3)	24.8 (4.6)
Controls from case-control studies						
DGI controls	1,277	51	56.1 (8.7)	133.3 (18.4)	80.1 (10.0)	26.7 (3.8)
FUSION NGT controls	1,038	49	58.2 (10.7)	139.4 (19.3)	81.5 (10.3)	27.1 (4.0)
MIGen controls	1,121	38	48.9 (8.3)	127.1 (17.8)	80.2 (11.6)	27.1 (5.2)
PROCARDIS controls	795	37	58.9 (6.9)	134.7 (18.6)	82.8 (10.0)	25.9 (3.7)

Validation set						
Study	N	% women	Age (SD) in years	SBP (SD) in mmHg	DBP (SD) in mmHg	BMI (SD) in kg/m^2
NESDA	1591	69	41.6 (12.4)	134.9 (19.4)	81.6 (11.5)	25.5 (4.9)

* Subject in NFBC1966 were examined at age 31.

Data for identifying sharing genetic basis between SBP and DBP

International Consortium for Blood Pressure GWAS (ICBP-GWAS) includes the GWAS result of cohort in Global BPgen Consortium[2], CHARGE Consortium[1] and other 5 new cohorts. The total sample size of ICBP-GWAS is 67,806 and the detailed demographic of ICBP-GWAS listed in Table 2. For each analysis, we divided the ICBP-GWAS dataset cohort-wisely into a discovery set, containing ~80% samples, and a non-overlapping validation set, containing ~20% samples. It shows an example of cohort-wisely splitting ICBP into 80% (discovery set) and 20% (validation set) in Table 2. The cohorts whose cell is light green filled composed the validation set with a sample size of 13,773 and the rest composed the discovery set with a sample size of 54,033. This process was repeated 50 times.

Table 2 Demographic of cohorts in ICBP

Study	N	% women	Age (SD) in years	SBP (SD) in mmHg	DBP (SD) in mmHg	BMI (SD) in kg/m^2
Population-based cohorts						
AGES	3,164	58	51 (6)	132 (17)	83 (10)	25.2 (3.5)
ARIC	8,052	53	54 (6)	118 (17)	71 (10)	27.0 (4.9)
BLSA	708	44	42.4 (13.2)	119.5 (15.0)	77.3 (10.2)	24.5 (3.6)
B58C – T1DGC	2,580	51	44.3 (0.3)	121.7 (15.3)	79.3 (10.5)	27.4 (4.9)
B58C - WTCCC	1,4 73	50	44.9 (0.4)	126.7 (15.2)	79.1 (10.2)	27.4 (4.7)
CHS	3,277	61	72 (5)	135 (21)	70 (11)	26.3 (4.4)
CoLaus	4,969	53	51.7 (9.5)	127.3 (17.4)	79.4 (10.8)	25.8 (4.6)
EPIC- Norfolk-GWAS	2,100	54	57.2 (7.8)	136.7 (19.1)	83.9 (11.9)	26.3 (3.9)
Eurospan Croatia	697	58	57.8 (15.6)	137.6 (24.5)	80.5 (11.5)	27.3 (4.3)
Eurospan Erf	1,300	60	50.5 (15.8)	139.7 (20.8)	80.0 (10.0)	26.7 (4.7)
Eurospan Orkney	700	53	53.6 (15.7)	130.2 (19.3)	76.3 (10.1)	27.8 (4.9)
Eurospan Sami	644	53	47.0 (20.7)	122.8 (18.7)	74.1 (7.9)	26.3 (4.8)
Eurospan Tyrol	1,096	57	45.3 (16.1)	132.9 (20.2)	79.9 (11.1)	25.6 (4.8)
Fenland	1,401	56	45.0 (7.3)	122.8 (16.3)	75.5 (10.7)	27.1 (4.9)
Framingham Heart Study	8,096	54	38 (9)	119 (15)	77 (10)	25.9 (4.9)
InCHIANTI	562	55	56.9 (14.5)	138.4 (20.1)	81.4 (10.1)	27.1 (4.2)
KORA	1,664	51	52.5 (10.1)	133.4 (18.5)	81.8 (10.9)	27.3 (4.1)
NFBC1966	4,761	52	31*	125.2 (13.8)	77.5 (11.7)	24.6 (4.2)
Rotterdam study	4,737	60	68 (8)	139 (22)	74 (11)	26.2 (3.6)
Rotterdam Extended Study	1,760	56	64 (7)	143 (21)	79 (11)	27.2 (4.2)
SardiNIA	3,998	57	40.8 (15.3)	128.7 (28.4)	79.9 (17.3)	25.1 (4.6)
SHIP	3,310	53	45.0 (13.9)	133.1 (20.2)	83.5 (11.3)	26.9 (4.7)
SUVIMAX	1,823	60	50.5 (6.2)	120.9 (12.3)	78.0 (8.1)	23.5 (3.3)
TwinsUK	873	100	45.8 (11.9)	122.9 (15.4)	78.2 (10.3)	24.8 (4.6)
Controls from case-control studies						
DGI controls	1,277	51	56.1 (8.7)	133.3 (18.4)	80.1 (10.0)	26.7 (3.8)
FUSION NGT controls	1,038	49	58.2 (10.7)	139.4 (19.3)	81.5 (10.3)	27.1 (4.0)
MIGen controls	1,121	38	48.9 (8.3)	127.1 (17.8)	80.2 (11.6)	27.1 (5.2)
PROCARDIS controls	795	37	58.9 (6.9)	134.7 (18.6)	82.8 (10.0)	25.9 (3.7)

* A example of splitting the ICBP cohort wisely into a discovery set, containing ~80% of the total samples, and a validation set, containing the rest ~20% of the total samples. The cohorts whose cell is light green filled composed the validation set with a sample size of 13,773 and the rest composed the discovery set with a sample size of 54,033.

Phenotype modeling, GWAS and Meta-analysis

Phenotype modeling of NESDA cohort. The NESDA cohorts were cleaned and preprocessed using the same protocols as the study of Global BPgen Consortium. Firstly, an adjustment of continuous traits for medication was conducted. Individuals had undergone a heart operation or had been diagnosed with a heart condition, such as coronary heart disease, angina pectoris and heart failure were excluded from the analysis. By this way, we excluded 49 out of 1788 individuals in the NESDA cohort. Then blood pressure was corrected by adding 15mmHg and 10mmHg for SBP and DBP for the individuals who took antihypertension medication. In the former study, we have discovered that the antidepressant medication use do have a significant effect on both SBP and DBP. So we further excluded the 58 individuals who undertook antidepressant medication (i.e. tricyclic antidepressants). Finally we selected the individuals who have both genotype and phenotype data and ended up with a sample size of 1591. In the adjusted NESDA cohorts with 1591 individuals, we calculated the residuals of SBP and DBP, after the adjustment for age, age2, body mass index (BMI) in the sex-specific way.

GWAS of NESDA cohorts. genome-wide association (GWA) between the SBP and DBP adjusted for age, age2, body mass index (BMI) and SNPs was conducted for females and males respectively, using the linear regression model under the assumption of an additive model of genotypic effect. The software used for the cohorts containing unrelated samples is SNPTEST V1 (http://www.stats.ox.ac.uk/~marchini/software/gwas/snptest.html).

Meta-analysis. The regression results of female and male cohorts were meta-analyzed using inverse variance weighting. The meta-analysis was conducted by METAL (http://www.sph.umich.edu/csg/abecasis/metal/).

Results

Model validation by simulated data

We simulated a discovery set and a validation set, where the phenotype has a genetic composition of 10 independent SNPs. Each dataset had a sample size of 100. We applied the classic polygenic model to calculate the exact *a* coefficient in regression model $y_j = \mu + as_j + e_j$ and the proportion of explained variance (i.e. R^2). The approximate *a* coefficient and R^2 were calculated by using the approximate likelihood methods. We repeated the simulation 500 times and tested the similarity and correlation between exact and approximate *a* coefficient and R^2. The simulated phenotypes in two dataset (discovery and validation data) are uncorrelated (i.e. correlation coefficient = 0.0817739). Figure 1 shows that the approximations (coefficient a and R^2) are mostly located along the diagonal line, which proved that the approximate results are highly corrected to the exact estimates.

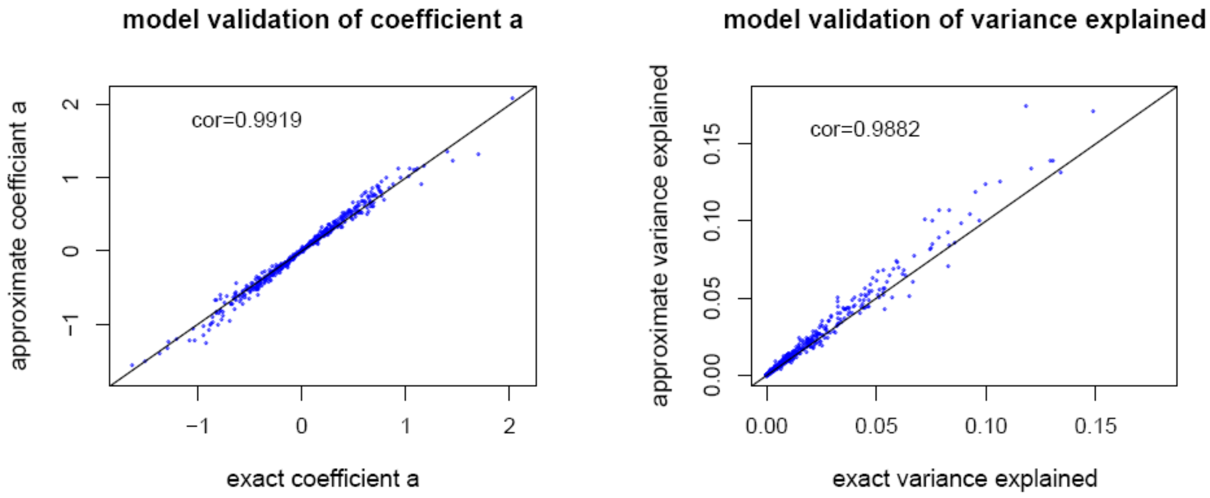

Figure 1 Results of model validation by simulated data. The right plotting is the exact coefficient a , calculated by classic polygenic model, against the approximate a , calculated by approximate likelihood method. They have a statistically significant correlation around 0.9919. The left plotting is for proportion of explained heritability (R^2) with a correlation around 0.9882. The line is the diagonal line.

When the variance explained values is large (>=0.07), the values calculated by approximate likelihood and classic polygenic model are a little bit different from each other (i.e. the dots deviated from the diagonal line). This is due to the failure of satisfying the assumption in approximate likelihood method. The approximate likelihood assumes that the likelihood function is approximately proportional to a Gaussian density with zero covariance when the explanatory variables (i.e. SNPs) are independent and the total fraction of variance explained is small. In our cases, the previous research has demonstrated that the variance explained by genetic factors is quite small, less than 0.001[1,2]. Thus, it is safe to use the approximate likelihood method, which just need GWAS result to calculate the proportion of explained heritability, to substitute the classic one, which requires the both the genotype data and GWAS result.

Model validation by real data

We also test the performance of the approximate likelihood method by using the real data. Global BPgen, containing 17 cohorts with a total sample size of 34,433 is the discovery set. The GWAS results of different cohorts in Global BPgen consortium were meta-analyzed using inverse variance weighted fixed-effects models. The validation set is NESDA with a sample size of 1591, which has both the GWAS result and the genotype data.

We selected SNPs with p-value below a range of thresholds from meta-analysis result of all Global BPgen GWAS data for DBP and SBP, and then applying stringent LD pruning with a threshold of 0.05 for pair-wise r^2 from Hapmap CEU by using PLINK (http://pngu.mgh.harvard.edu/~purcell/plink/) and SNAP (http://www.broadinstitute.org/mpg/snap/).

For each associated p-value threshold, we calculated the multi-SNP risk score in 1591 individuals in the NESDA cohort using genotype and the relative risk β_i at i^{th} SNP and fitted the risk score to the residual DBP/SBP data after sex –specific adjustment for age, age2 and BMI. We then compared the estimated coefficient a and the resulting R^2 from this analysis of individual data, with the values obtained using the approximate likelihood method. Figure 2 is the plotting of coefficient a and the resulting R^2 generated by classic polygenic model and the approximate likelihood method. The dots are close to the diagonal line in figure 2. This demonstrated that the parameter a and resulting R^2 by two methods has a statistically significant high correlation and close enough. From both the simulation and real data validation, we concluded that the approximation of coefficient a and R^2 are close enough to be acceptable.

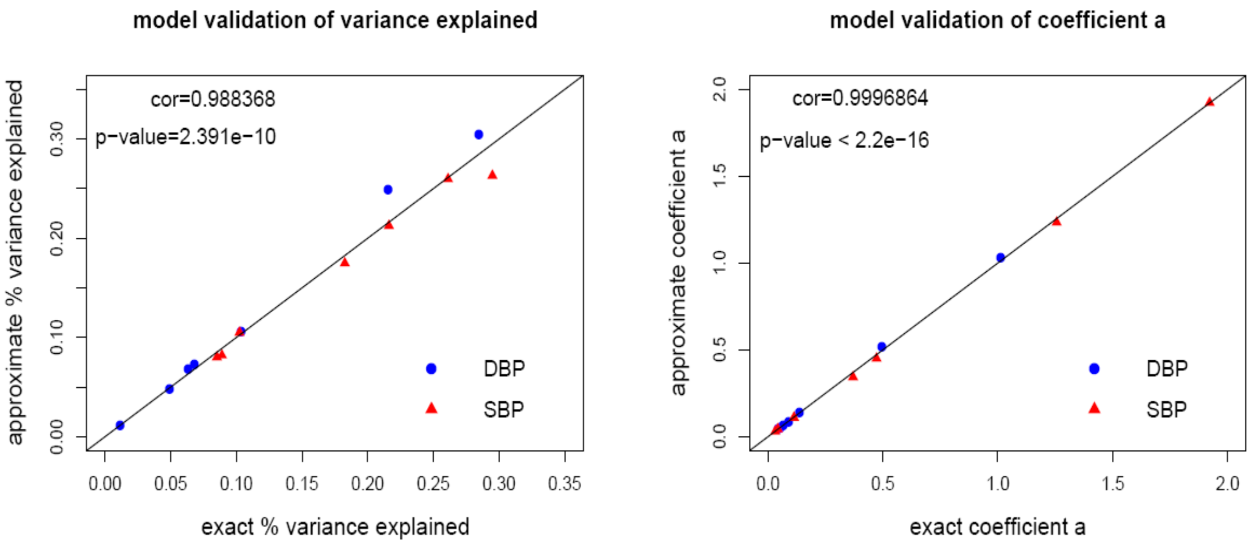

Figure 2 Result of model validation by real data. The left plotting is the exact coefficient a, calculated by classic polygenic model, against the approximate a, calculated by approximate likelihood method. The right plotting is for proportion of explained heritability (R^2). The blue dot is the result for DBP trait and the red triangle is the result for SBP trait. The line is the diagonal line.

Finding missing heritability and sharing genetic basis between SBP & DBP

After confirming that the approximation is close enough to be acceptable, we applied the approximate likelihood method to estimate the variance of DBP/SBP explained by multi-SNP risk factors. ICBP is divided into two sets in a cohort-wise way. One containing ~80% of samples served as the discovery dataset and the other one, containing the rest ~20% is the validation dataset. The GWAS result of different cohorts in each dataset were meta-analyzed using inverse variance weighted fixed-effects models. The DBP multi-SNP risk factor for each sample in the validation set is calculated with the relative risk and genotype data for each SNP selected from all SNPs with association p-value below different threshold in discovery set. Then, we applied stringent LD pruning with a threshold of 0.05 for pairwise r^2 from Hapmap CEU by using PLINK and SNAP. We quantified how much of heritability of DBP and SBP can be explained by this DBP multi-SNP risk factor. The comparison between the heritability of DBP and SBP by the DBP multi-SNP risk factor indicates the extent to which two traits have a shared genetic basis. We also calculated the SBP multi-SNP risk factor in the same way. Then we quantify the variance of both SBP and DBP explained by this SBP multi-SNP risk factor. The analysis was repeated for 50 time to dilute the bias caused by the cohort-wisely data separation.

SNPs with small effect helps to find some missing heratibality. Figure 3 displays proportion of variance explained by multi-SNP risk factor in DBP and SBP. We used 9 different p-value thresholds to select the associated variants in the discovery set and quantify the explained variance in the validation set. The values next to the dots in figure 3 are the number of selected SNPs by using different p-value cutoff.

For both SBP and DBP, there was a trend for proportion of explained variance to increase as the p-value threshold for including SNPs in calculation of a genetic risk factor became less stringent. Proportion of variance explained by significantly associated SNPs (i.e. p-value threshold less than 1e-04) is quite small, not exceeding 0.07%. When loosing the p-value threshold, more SNPs were included in the genetic risk factor and more phenotypic variance explained by the multi-SNP risk factor has been found. We achieved the most significant increase at the p-value threshold of 0.01. For example, the proportion of explained DBP variance by 627 SNPs reached 0.216% at p-value of 0.01 from 0.121% at p-value of 0.001, an increase of 0.095%. Correspondingly, the number of SNPs at p-value of 0.01 is almost eight times of the one at p-value of 0.001. This is the largest comparative difference of SNP number for two neighboring p-value. The increase of the explained variance indicates that those SNPs which are with small effect size convey some information of the missing heritability and it is quite possible these SNPs behave in an additive mode.

The increase rate of the explained variance at p-value of 0.05 and 0.01 is flatter than the one at p-value of 0.01 and 0.001 even although including around 2,100 SNPs extrally. This indicates that the SNPs in this threshold period (i.e. 0.01~0.05) contains some not significantly associated or even unassociated SNPs, which bring noise to the genetic factor. In another word, the contamination caused by these SNPs overweighs the contribution of the additional association SNPs with small effect size. This hypothesis is also supported by the result that the explained variance of DBP reached the apex at the liberal p-value threshold (i.e. p-value = 0.05) for DBP (~ 0.23%) and SBP (~0.24%) in stead of at the least stringent p-value threshold of 0.1.

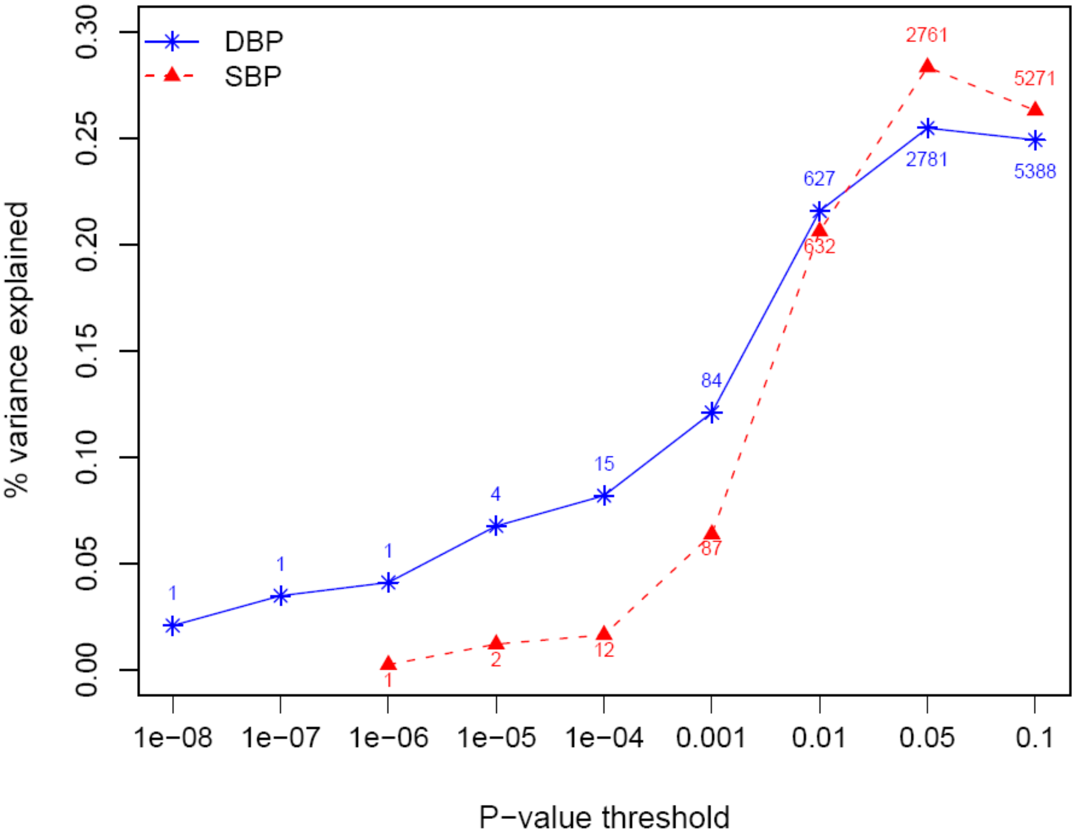

Figure 3 Plotting of explained variance against different p-value threshold. The blue stars are the explained DBP variance in the validation set by the multi-SNP risk factor generated in the discovery set and the red triangles are for the SBP trait. The values next to the blue stars and the red triangles are the numbers of SNPs corresponding to different p-value thresholds.

Interestingly, the explained variance of SBP are smaller than that of DBP at the p-value <= 0.01. However, at two most loosing p-value thresholds, the explained variance of SBP exceeded that of DBP even with fewer SNPs. This result is concordance with the recent GWA study of blood pressure[1,2]; the discovered SNPs associated with SBP are fewer than the SNPs associated with DBP. Moreover, the effect size of those SNPs of SBP are even smaller than those of DBP. The GWAs result indicates that comparing with DBP, SBP has a higher possibility that there exists some undiscovered real causal loci of modest or small effect. The exceeding action of explained SBP variance by large number of SNPs at less stringent p-value thresholds prove our guessing more or less.

DBP and SBP have a highly shared genetic basis. Besides providing a substantial polygenic component to the risk of DBP involving thousands of common allele of small effect, we also show that the DBP multi-SNP risk factor component also contributes to the risk of SBP, which reveal that DBP and SBP share a lot genetic risk factors. We quantified how much of the SBP variance was explained by the mutli-SNP DBP risk score in the validation set. The comparison of the cyan and blue bar-charts (Figure 4 upper) at a certain p-value threshold, for example p-value=0.001, shows that SBP and DBP have 85% genetic background in common. The significant sharing of a genetic background indicates that two diseases may have some same causal genes, or share a common regulatory pathway.

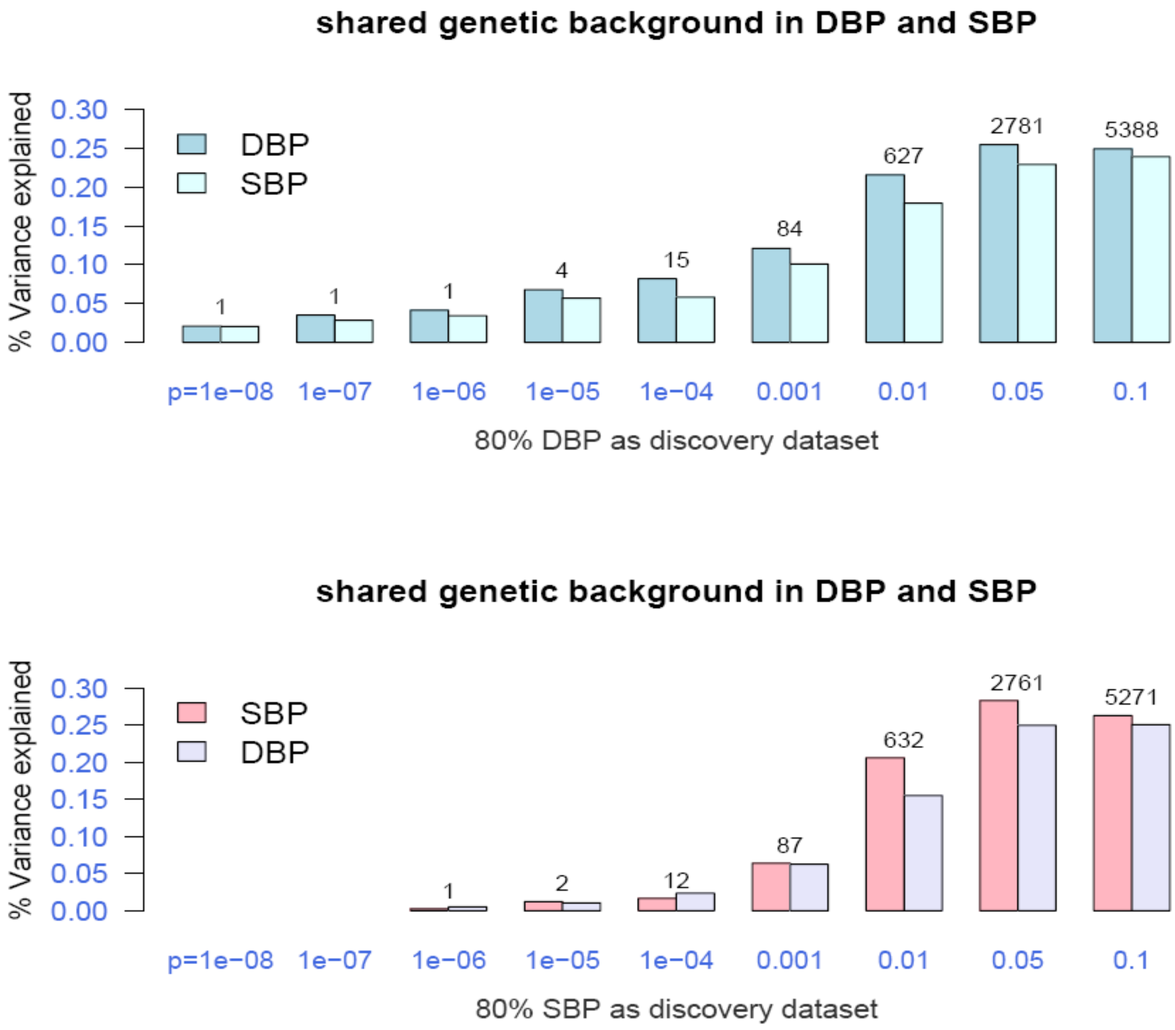

Figure 4 The upper bar chart is the proportion of DBP and SBP heritability explained by the DBP multi-SNP risk score, which is generated from the discovery set containing 80% samples in ICBP. And the low one is the proportion of DBP and SNP heritability

explained by SBP multi-SNP risk score. The values above the bar are the numbers of SNPs corresponding to different p-value thresholds.

When we use the multi-SNP SBP risk score to quantify the explained DBP variance, we found that the risk score even explained more variance in DBP than in SBP at the stringent p-value of 1e-06 (i.e. 0.00522% in DBP and 0.00274% in SBP). We are not surprised to at this result. Because it showed most significantly associated SNPs for SBP were also discovered to be associated with DBP in the GWAs results[1,2]. For example, all the top10 hits for SBP in CHARGE consortium study (all p-values < 1e-10) are significantly associated with DBP (all p-values < 1e-06)[1]. In another word, the selected SNPs for SBP at a stringent p-value possess a high potential to be the true causal SNPs for DBP.

Discussion

In the recent research by P. Visscher and his colleagues[16], their result showed that including all the common SNPs can explain 45% of heritability of height and concluded that the heritability is not missing but failed to be detected because of the small effect size of single genetic factor. In our case, even though the increased detected heritability of SBP and DBP by multi-SNP risk factor is not as much as they discovered in height (increasing 0.14% of heritability), our results still proved that genome-wide risk scores including more associated SNPs with small effect helps to detected the phenotypic variance of complex disease. The reason that the increased explained heritability is so small compared with the explained height heritability is possible the SNPs we including are much fewer (5387 in our work vs. 294,831 in P. Visscher's work).

Compared with explained phenotypic variance at p-value=0.05 (2781 SNPs), there is almost no increase of explained phenotypic variance at the p-value=0.1 (5388), even though the number of included SNPs doubled. One possible reason is that as the p-value threshold became less stringent, the contamination caused by the unassociated SNPs overweighed the contribution of the associated SNPs of small effect-size. While including the genuinely associated loci, the genome-wide risk score includes hundreds or thousands of loci that are not significantly associated or not associated with the disease.

This indicates that we can find increase of explained phenotypic variance by adding genome-wide information (i.e. including genome-wide scale SNPs) in those diseases for which there are not many confirmed causal loci have been discovered. If the causal loci, which is with large effect, of a complex disease have not been known, it shows some possibility that there exists some undiscovered loci of modest or small effect.

SBP and DBP are two highly correlated phenotypes (i.e. the correlation between SBP and DBP in NESDA is around 0.74), which is coordinated with the strong shared genetic basis around 85%. Provided with such a high genetic sharing, it is possible to identify genes or biological systems which can be shared or unique to SBP and DBP by conducting pathway analysis or clustering gene with candidate SNPs. This may help to find the treatment target of hypertension.

Even though the genome-wide risk score at a less severe threshold results in a 3-fold increase of the explained variance. It is still far from the SBP and DBP heritability, ranging from 40% -60%[8]. There are several possible reasons that may explain for the huge gap between the explained and remaining heritability. First, our multi-SNP risk score failed to capture the gene-gene interactions, which has been revealed to contribute to the risk of human disease[17,18]. If a genetic factor functions through a complex mechanism which involves multiple other genes, even some environmental factors, it is quite possible that the effect of this genetic factor is missed if we examine a factor in a one by one way (just as GWAS). Secondly, the SNP pool we used has been filtered with a MAF > 1%, which means that we did not take the rare SNPs into consideration in the multi-SNP model. Researches have discovered some rare SNPs associated with complex disease[19]. The different between the chance of harboring rare variants in individual with and without a certain disease[19-22] indicates that single or multiple rare variants tends to be associated with some disease or disease-related phenotypes. Dickson et al. has proposed there is a synthetic association between common SNPs and rare SNPs to explain how rare SNPs contribute to the association signal owing to common SNPs in GWAS study[23]. They found that the rare variants can easily create a statistically significant genome-wide association signal to a common variant (i.e. a common variant could convey a diluted association signal owing to a rare variant with large effect.) by both simulation and real data study. Third, the genetic variation buffering mechanism may attenuate the effect size of DNA sequence variation on the targeted phenotype. C.H. Waddington proposed that organisms tend to be adjusted in order to result in one definite end result regardless of the minor variation, both genetic and environmental variants[24]. Other genes functioning in the same biochemical pathway can buffer the effect of some SNPs on a certain gene[25,26]. Buffering can also caused by regulatory feedbacks. Negative feedback is common in biology and it acts to minimizing the variation propagation[27,28]. In order to trace back the buffered SNPs, it is wise to include the SNPs which are co-localized with the quantitative trait loci (QTL) of those gene-expression, protein or metabolism traits, which are associated with complex diseases. There are researches showed that variants associated with gene expression have little overlap with those associated with complex disease[29].

In the future work, we want to cluster the SNPs according to gene catalogue which helps to discovery the functional gene, or we may cluster the SNPs according to the chromosome site which enable us with the annotation of new function gene. And the clustered SNPs may help to exclude the unassociated SNPs. For example, we can cluster the SNPs at p-value threshold of 0.1 by using the cardiovascular related regulatory pathway. Those SNPs clustered in a less correlated pathway bear less potential to be the true causal variants. In addition, we want to apply this multi-SNP model to find the common polygenic variation within the same kind of diseases, such as type I diabetes and celiac, both of them are autoimmune-related disease. If we can find a moderate to high share genetic background, it is promising that we may conduct a meta-analysis between two phenotype or we may find some shared targeted genes in pathway analysis.

Recently, the nature of 'missing heritability' is the main problem in many complex diseases. For DBP and SBP, our data points to a genetic architecture that includes many variants of small to moderate effect. Our result proves that there exists some 'missing heritability' in the large amount of SNPs with small to moderate effect.

References

1 Levy, D. *et al.* Genome-wide association study of blood pressure and hypertension. *Nat Genet* **41**, 677-687 (2009).

2 Newton-Cheh, C. *et al.* Genome-wide association study identifies eight loci associated with blood pressure. *Nat Genet* **41**, 666-676 (2009).

3 Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747-753 (2009).

4 Zeggini, E. *et al.* Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat Genet* **40**, 638-645 (2008).

5 Barrett, J. C. *et al.* Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat Genet* **40**, 955-962 (2008).

6 Consortium, T. W. T. C. C. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661-678 (2007).

7 Levy, D. *et al.* Evidence for a Gene Influencing Blood Pressure on Chromosome 17 : Genome Scan Linkage Results for Longitudinal Blood Pressure Phenotypes in Subjects From the Framingham Heart Study. *Hypertension* **36**, 477-483 (2000).

8 Evans, A. The Genetics of Coronary Heart Disease: The Contribution of Twin Studies. *Twin Research and Human Genetics* **6**, 432-441 (2003).

9 Pharoah, P. D. P. *et al.* Polygenic susceptibility to breast cancer and implications for prevention. *Nat Genet* **31**, 33-36 (2002).

10 Yang, Q., Khoury, M. J., Botto, L., Friedman, J. M. & Dana, F. W. Improving the Prediction of Complex Diseases by Testing for Multiple Disease-Susceptibility Genes. *Am. J. Hum. Genet* **72**, 636-649 (2003).

11 Janssens, A. C. J. W. *et al.* Predictive testing for complex diseases using multiple genes: Fact or fiction? *Genetics in Medicine* **8** (2006).

12 Evans, D. M., Visscher, P. M. & Wray, N. R. Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. *Hum. Mol. Genet.* **18**, 3525-3531 (2009).

13 Smyth, D. J. *et al.* Shared and Distinct Genetic Variants in Type 1 Diabetes and Celiac Disease. *New England Journal of Medicine* **359**, 2767-2777, doi:doi:10.1056/NEJMoa0807917 (2008).

14 Schaefer, A. S. *et al.* Identification of a Shared Genetic Susceptibility Locus for Coronary Heart Disease and Periodontitis. *PLoS Genet* **5**, e1000378 (2009).

15 Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748-752 (2009).

16 Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* **42**, 565-569 (2010).

17 Neuman, R. J. *et al.* Gene-Gene Interactions Lead to Higher Risk for Development of Type 2 Diabetes in an Ashkenazi Jewish Population. *PLoS ONE* **5**, e9903 (2010).

18 Wiltshire, S. *et al.* Epistasis Between Type 2 Diabetes Susceptibility Loci on Chromosomes 1q21-25 and 10q23-26 in Northern Europeans. *Annals of Human Genetics* **70**, 726-737 (2006).

19 Nejentsev, S., Walker, N., Riches, D., Egholm, M. & Todd, J. A. Rare Variants of IFIH1, a Gene Implicated in Antiviral Responses, Protect Against Type 1 Diabetes. *Science* **324**, 387-389, doi:10.1126/science.1167728 (2009).

20 Cohen, J. *et al.* Multiple rare variants in *NPC1L1* associated with reduced sterol absorption and plasma low-density lipoprotein levels. *Proc Natl Acad Sci* **103**, 1810-1815 (2006).

21 Cohen, J. *et al.* Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9. *Nat Genet* **37**, 161-165 (2005).

22 Ji, W. *et al.* Rare independent mutations in renal salt handling genes contribute to blood pressure variation. *Nat Genet* **40**, 592-599 (2008).

23 Dickson, S. P., Wang, K., Krantz, I., Hakonarson, H. & Goldstein, D. B. Rare Variants Create Synthetic Genome-Wide Associations. *PLoS Biol* **8**, e1000294 (2010).

24 Waddington, C. H. canalization of development and the inheritance of acquired characters. *Nature*, 563-565, doi:10.1038/150563a0 (1942 November).

25 Guo, B., Styles, C. A., Feng, Q. & Fink, G. R. A Saccharomyces gene family involved in invasive growth, cell–cell adhesion, and mating. *Proc. Natl. Acad. Sci. U.S.A.* **97**, 12158-12163 (2000).

26 Wagner, A. Robustness against mutations in genetic networks of yeast. *Nat Genet* **24**, 355-361 (2000).

27 Alon, U., Surette, M. G., Barkai, N. & Leibler, S. Robustness in bacterial chemotaxis. *Nature* **397**, 168-171 (1999).

28 Yi, T. M. e. a. Robust perfect adaption in bacterial chemotaxis through integral feedback control. *PNAS* **97**, 4649-4653 (2000).

29 Heinzen, E. L. *et al.* Tissue-Specific Genetic Control of Splicing: Implications for the Study of Complex Traits. *PLoS Biol* **6**, e1000001 (2008).