LD Score regression for estimating and partitioning heritability of lipid levels in the Finnish population

Heidi Marika Hautakangas

Master's thesis
Statistics
Department of Mathematics and Statistics
University of Helsinki
August 2018

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta/Osasto — Fakultet/Sektion — Faculty	Laitos — Institution — Department
Faculty of Science	Department of Mathematics and Statistics

Tekijä — Författare — Author
Heidi Hautakangas

Työn nimi — Arbetets titel — Title
LD Score regression for estimating and partitioning heritability of lipid levels in the Finnish population

Oppiaine — Läroämne — Subject
Statistics

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Master's thesis	August 2018	67 s.

Tiivistelmä — Referat — Abstract

To better understand biology of complex traits, quantifying the contribution of different genetic factors is essential. Heritability is a population parameter that estimates the proportion of phenotypic variance explained by genetic factors. A recent goal in statistical genetics has been to estimate heritability from genome-wide association study (GWAS) data. GWAS have shown that a large number of genetic variants with small effects together affect complex traits. Because the individual effects are so small, a challenge of the GWAS is to achieve enough statistical power to detect the true associations. Statistical power has been increased by increasing the GWAS sample size, typically by a meta-analysis. In a meta-analysis, summary association statistics from multiple study cohorts are jointly analysed, and therefore it is often impossible to get access to the original individual-level data underlying the meta-analysis.

In this thesis, I will study linkage disequilibrium score regression (LDSC), that estimates heritability by regressing GWAS summary statistics on linkage disequilibrium (LD) scores, that measure how much genetic variation each variant tags. Importantly, LD Scores can be estimated from a reference panel without requiring any individual-level data. Furthermore, I will study stratified LD Score regression (S-LDSC), that is an extension of LDSC for partitioning heritability by functional annotations.

This thesis has three aims. First, to explain the statistics behind LDSC. Second, to evaluate the effect of LD reference panel on heritability estimation of lipid levels in the Finnish population by comparing an in-sample LD reference panel to external LD reference panels. Third, to partition the heritability of lipid levels in the Finnish population by functional annotations using S-LDSC. I applied LDSC and S-LDSC to the National FINRISK Study and used four lipid levels as quantitative phenotypes: high-density lipoprotein cholesterol (HDL-C), low-density lipoprotein cholesterol (LDL-C), triglycerides (TG) and total cholesterol (TC).

As results, I observed that LDSC was robust to the choice of LD reference panel when applied to the Finnish population. Heritability estimates were consistent between different LD reference panels regardless of the LD mismatch. The highest heritability point estimates and the lowest point estimates of confounding biases were produced by the Finnish specific panels, though the differences were not statistically significant. In the heritability enrichment analyses, I replicated several previous findings: for example, I observed enriched heritability for many histone marks in all four lipid traits and enriched heritability for super enhancers for HDL-C, TC and TG.

Avainsanat — Nyckelord — Keywords
Linear regression model, Linkage disequilibrium, Genome-wide association study, Heritability, Lipids

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — Övriga uppgifter — Additional information

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta/Osasto — Fakultet/Sektion — Faculty	Laitos — Institution — Department
Matemaattis-luonnontieteellinen	Matematiikan ja tilastotieteen osasto

Tekijä — Författare — Author
Heidi Hautakangas

Työn nimi — Arbetets titel — Title
Kytkentäepätasapainopistemääräregressio lipiditasojen periytyvyyden estimoinnissa ja osituksessa suomalaisessa populaatiossa

Oppiaine — Läroämne — Subject
Tilastotiede

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Pro gradu -tutkielma	Elokuu 2018	67 s.

Tiivistelmä — Referat — Abstract

Erilaisten geneettisten tekijöiden vaikutusten määrittäminen on tärkeää, jotta voidaan ymmärtää biologiaa monitekijäisten ominaisuuksien taustalla. Periytyvyys on populaatioparametri, jolla estimoidaan geneettisten tekijöiden osuutta ilmiasun vaihtelussa. Viime aikoina tilastollisen genetiikan tavoitteena on ollut periytyvyyden estimoiminen genominlaajuisten assosiaatiotutkimusten (GWAS) tuloksista. GWAS:t ovat osoittaneet, että monitekijäisten ominaisuuksien ilmenemiseen vaikuttaa suuri joukko geenimuotoja, joilla on yksittäin pieni vaikutus. Koska yksittäiset vaikutukset ovat pieniä, on GWAS:n haasteena saavuttaa riittävä tilastollinen voima niiden havaitsemiseen. Tilastollista voimaa voidaan kasvattaa otoskokoa kasvattamalla, tyypillisesti meta-analyysiä hyödyntäen. Meta-analyysissä yhdistetään usean yksittäisen tutkimuskohortin tulokset, minkä seurauksena pääsy meta-analyysin pohjana oleviin yksilötason aineistoihin on useimmiten mahdotonta.

Tutkielmassa perehdytään kytkentäepätasapainopistemääräregressio (LDSC) -menetelmään, joka arvioi periytyvyyttä regressoimalla GWAS-tulokset kytkentäepätasapaino (LD) -pistemääriä vasten. LD-pistemäärät mittaavat kuinka paljon geneettistä vaihtelua kukin geenimerkki ilmentää. Erityisesti LD-pistemäärät voidaan estimoida verrokkipaneelista, eikä yksilötason aineistoa tarvita. Lisäksi tutkielmassa perehdytään ositettuun LD-pistemääräregressioon (S-LDSC), jolla periytyvyys voidaan osittaa erilaisiin genomin toiminnallisiin luokkiin.

Tutkielmalla on kolme tavoitetta. Ensimmäinen tavoite on kuvailla LDSC:ssa käytettävä tilastotiede. Toinen tavoite on arvioida eri LD-pistemäärien verrokkipaneelien vaikutusta lipiditasojen periytyvyyden estimointiin suomalaisessa populaatiossa vertailemalla GWAS-otoksen sisäistä paneelia ulkopuolisiin paneeleihin. Kolmas tavoite on osittaa eri lipiditasojen periytyvyys erilaisiin genomin toiminnallisiin luokkiin käyttäen S-LDSC-menetelmää. Tutkielmassa menetelmiä sovelletaan FINRISKI-tutkimukseen käyttäen neljää lipiditasomuuttujaa: HDL-, LDL- ja kokonaiskolesterolia sekä triglyseridejä.

Tutkielmassa havaittiin, että LDSC oli vakaa LD-pistemäärän verrokkipaneelin valinnan suhteen, kun menetelmää sovellettiin suomalaiseen populaatioon. Periytyvyyden estimaatit olivat yhteneviä eri LD-pistemäärän verrokkipaneelien välillä huolimatta eroista LD-pistemäärissä. Suomalaiset paneelit tuottivat sekä suurimmat periytyvyyden piste-estimaatit että pienimmät sekoittavista tekijöistä johtuvaa harhaa kuvaavat piste-estimaatit, joskaan erot eivät olleet tilastollisesti merkitseviä. Analyysit periytyvyyden rikastumisesta toistivat useita aiemmin raportoituja tuloksia: rikastumista havaittiin esimerkiksi useissa histonimerkeissä kaikkien lidipitasojen kohdalla sekä supervahvistajissa HDL-kolesteroli-, kokonaiskolesteroli- ja triglyseriditasojen kohdalla.

Avainsanat — Nyckelord — Keywords
Lineaarinen regressiomalli, Kytkentäepätasapaino, Genominlaajuinen assosiaatiotutkimus, Periytyvyys, Lipidit

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — Övriga uppgifter — Additional information

Contents

Chapter 1

Introduction

1.1 Motivation

Genetics and environment together influence the human phenotypes, such as cholesterol levels or heart disease. The role of genetics varies between different traits, for example eye colour is determined entirely by genetics, whereas for educational attainment environmental factors have a greater role. An inheritance pattern of monogenic diseases, such as Huntington's disease, that are caused by a single genetic mutation, can be seen from a pedigree that shows the ancestral relationships and the presence or absence of the disease for each family member. For complex diseases and traits that are caused by both genetic and environmental factors, the pattern is not as clear as with monogenic diseases, and instead of a pedigree analysis, a quantitative definition of *heritability* is used. Heritability describes the proportion of the phenotypic variation that can be explained by genetic factors. [Klug et al., 2012]

Traditionally, heritability has been estimated from twin studies where phenotypic differences between pairs of monozygotic (MZ) and dizygotic (DZ) twins are compared. MZ twins are almost 100% identical in their genotypes, whereas DZ twins share approximately 50% of their genotypic variation, the same amount as other full siblings. Therefore, phenotypic differences in MZ twins can be assumed to represent the effect of environmental factors and heritability can be estimated by comparing trait correlation between MZ and DZ twin pairs. [Visscher et al., 2008]

More recently, new methods [Yang et al., 2010] have been developed that estimate heritability by comparing both genotypic and phenotypic similarities between seemingly unrelated individuals. The development of genotyping chips made it possible to easily measure hundreds of thousands of variants from genomes of thousands of individuals. For example, heritability can be estimated by partitioning phenotypic variance into variance components with linear mixed models (LMM) that apply restricted maximum-likelihood (REML). However, the heritability estimates from unrelated individuals correspond only to the heritability tagged by those variants that have been explicitly genotyped, often referred to as chip heritability. Therefore, the estimates obtained from methods utilizing the genotyping chip technologies have been substantially lower compared to those obtained from twin studies [Zuk et al., 2012], [Gusev et al., 2013]. A gap between estimates of twin heritability and chip heritability is called 'missing heritability' [Zuk et al., 2012].

A novel method called *linkage disequilibrium* (LD) Score regression (LDSC) [Bulik-Sullivan et al., 2015] estimates chip heritability by regressing summary statistics from a genome-wide association study (GWAS) on LD scores. LD is defined as a non-random association of *alleles* (i.e. alternative genetic variants at the same genomic position) between different genomic positions [Slatkin, 2008]. LD Score of a variant measures how much genetic variation the variant tags because of LD. The GWAS studies association between a genetic variant and a trait by comparing the genomes of a large number of individuals with varying phenotypes. LDSC uses information from all variants and LD which increases statistical power to explain the variance of the trait. For example, I compared the performance of LDSC to the heritability estimation based on only independent lead variants from a GWAS using four circulating lipid levels from the National FINRISK Study data: high-density lipoprotein cholesterol (HDL-C), low-density lipoprotein cholesterol (LDL-C), triglycerides (TG) and total cholesterol (TC). As a result, LDSC could explain more of phenotype variation in all four traits: LDSC explained 109 percent more of the variance of TG than lead variants, 53 percent more of the variance of TC, 50 percent more of the variance of LDL-C and 7 percent more of the variance of HDL-C than lead variants. The results are presented in Figure 1.1.

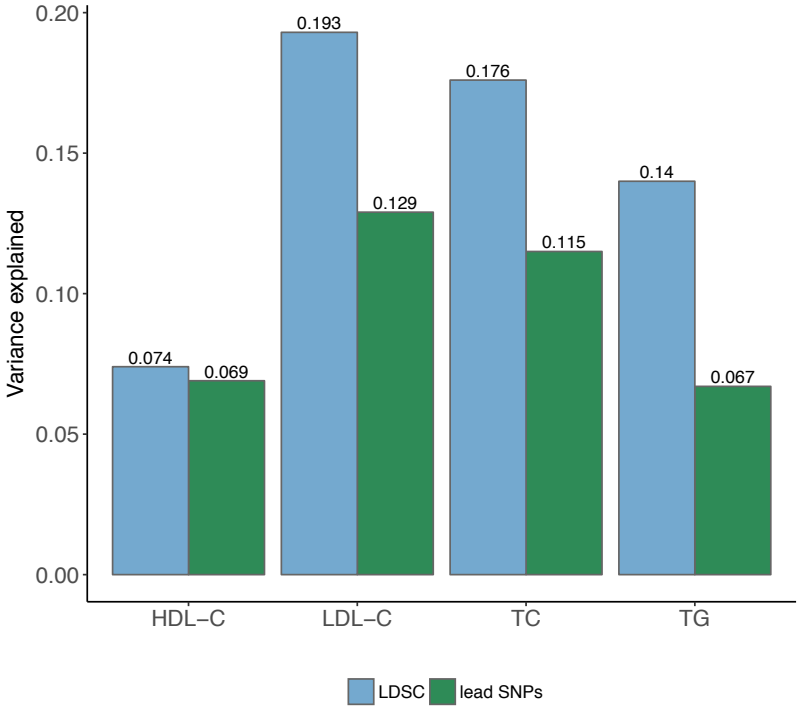


Figure 1.1: Proportion of total trait variance explained by lead variants (green bars) or by LD Score regression (LDSC) (blue bars) from the summary statistics of four genome-wide association studies of the lipid levels from the National FINRISK Study data.

The major advantage of LDSC compared to more traditional methods such as LMM, is that LDSC does not require any individual-level phenotype-genotype data. Instead, it uses only summary association statistics from the GWAS and an external LD reference panel to estimate heritability. However, an external LD reference panel has to match to the population used in the GWAS, and a mismatch between LD estimates and the GWAS sample can bias LDSC estimates [Bulik-Sullivan et al., 2015]. Summary-level data has three great advantages over the individual-level data. First, individual-level data is sensitive and privacy concerns often limit access to it, whereas from the summary-level data no single individual can be identified. Second, to increase power of the GWAS, many of the largest studies are conducted as a meta-analysis, where summary association statistics from multiple separate study cohorts are jointly analysed, and access to the original individual-level data is therefore usually not possible. Third, summary-level data is more compact and reduces the computational burden massively compared to individual-level data. For example, consider a meta-analysis of k studies including $N = N_1 + \ldots + N_k$ individuals and M variants. The size of the original individual-level data from all cohorts

would be $N \times M$ and the GWAS summary association statistics (for example regression coefficients and their standard errors) from all cohorts would be $2 \times k \times M$, whereas the size of the meta-analysis summary data would be only $2 \times M$, see Figure 1.2. The reduction in size from the individual-level data to the meta-analysis summary statistics would be $N/2$ -fold.

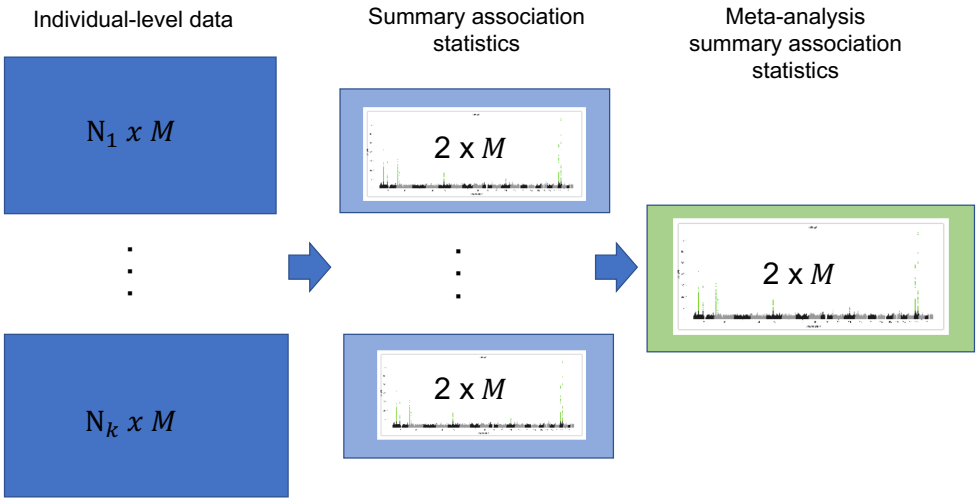


Figure 1.2: Data size reduction from individual-level data to summary-level data. Consider a meta-analysis of k study cohorts including $N = N_1 + \ldots + N_k$ individuals and M variants. The size of the original individual-level data would be $N \times M$, and the size of the summary association statistics (e.g. regression coefficients and their standard errors) from the cohorts would be reduced to $2 \times k \times M$. Whereas, the size of the meta-analysis summary association statistics would be only $2 \times M$.

Previous studies [Maurano et al., 2012], [Trynka et al., 2013] have shown that heritability of quantitative traits and complex diseases does not distribute uniformly across the whole genome, instead different functional parts of the genome contribute disproportionally to the heritability. LDSC model can be extended to partition the heritability by functional annotations. Stratified LD Score regression (S-LDSC) [Finucane et al., 2015] can be used to examine if some regions of the genome are enriched for heritability, which can improve the understanding of the genetic architecture behind quantitative traits and complex diseases.

1.1.1 Aims of the study

This study has three aims:

1. Explaining the statistics behind LD Score regression
2. Evaluating the effect of LD reference panel on heritability estimation of the lipid levels in the Finnish population
3. Partitioning the heritability of lipid levels in the Finnish population by functional annotations using stratified LD Score regression

The second aim relates to the Finnish population being one of the most studied genetic isolates. The gene pool of the Finns has been shaped by founder effects that occur when a new population is established by a small number of individuals leading to reduced genetic variation. This has led to increased LD compared to other European populations [Service et al., 2006], [Peltonen et al., 1999], [Exome Aggregation Consortium, 2016]. Therefore, LD estimates obtained from multi-ethnic Europeans might lead to biased estimates when applied to Finnish data.

1.2 Genome-wide association studies

A genome-wide association study (GWAS) examines whether there are any genetic variants associated with a trait or a disease (from now on "phenotype") by comparing the genomes of a large number of individuals with varying phenotypes. A GWAS aims to better understand the biology of disease, which could help to find better treatment or prevention of disease. The human genome consists of approximately three billion base pairs in a form of a linear sequence of four different bases, also called nucleotides, adenine (A), thymine (T), cytosine (C) and guanine (G). A *single nucleotide polymorphism* (SNP) is defined as a one nucleotide change in the genome sequence which is present within the population at least in the frequency of one percent. The GWAS tests whether an allele at a SNP appears more often than expected by chance in individuals with the disease compared to healthy controls, or whether the quantitative trait is distributed differently among the carriers of different alleles at a SNP. [Klug et al., 2012]

The two most commonly used study designs in GWAS are a case-control study and a cohort study [Pearson and Manolio, 2008]. In a case-control study, the allele frequencies

of the genomes of healthy individuals are compared to the genomes of the individuals with the disease. Case-control studies are usually easier and less expensive to conduct compared to cohort studies that involve collecting extensive baseline information. However, in cohort studies, participants are usually more representative of the population than in case-control studies, where cases are typically sampled from clinical sources which may lead to unrepresentation of the true variation of the disease because the mildest and/or the most fatal cases might be missed [Pearson and Manolio, 2008]. Depending on the trait of interest, most typical models used in the GWAS are logistic regression for binary traits and a simple linear regression for quantitative traits.

When carrying out hundreds of thousands or millions of tests of associations, as in GWAS, the number of false positives at traditional significance levels (such as 0.05) would be very high. A conventional way to deal with multiple testing problem and to reduce false-positive rate is to apply Bonferroni correction, where the significance level is divided by the number of tests performed. With GWAS, a threshold of 5×10^{-8} - which is equivalent of dividing significance level 0.05 by 10^6 - has become a standard genome-wide significance level regardless of the number of variants used in the study [Pe'er et al., 2008], [The International HapMap Consortium, 2005].

The 'common disease, common variant' hypothesis [Collins et al., 1997] states that the large number of variants with small effect together affect the disease phenotype which sets certain requirements for the study to reach enough statistical power to catch small effects. Sample sizes of the GWAS need to be large, and one way to increase sample size and power, is a meta-analysis. For example, one of the earliest GWAS of schizophrenia [The International Schizophrenia Consortium, 2009] that included only 3,322 cases and 3,587 controls was not able to detect any locus where association would have reached the genome-wide significance threshold. Years later after growing sample sizes, the meta-analysis study of 52 cohorts [Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014] including 36,989 cases and 113,075 controls was able to detect 108 genome-wide significant (GWS) schizophrenia-associated genetic loci, confirming that large sample sizes are needed to be able to detect the small genetic effects. Similar example is migraine, where the first GWAS [Anttila et al., 2010] including 2,731 cases and 10,747 controls detected one associated locus that reached the GWS threshold, whereas the most recent meta-analysis combining 22 GWAS [Gormley et al., 2016] with 59,674 cases and 316,078 controls increased the number of GWS migraine-associated locus to 38.

The results of GWAS are usually presented as a Manhattan plot, where x-axis displays the genomic coordinates in basepairs and y-axis displays the strength of the association as the negative logarithm of the association p-value for each SNP. Because of LD, variants

in LD with the causal variant will also show the same association, and true signals with the strongest associations will stand out as high peaks of stacked points forming a profile similar to view of skyscrapers in Manhattan. An example of a Manhattan plot of the GWAS results of LDL-C of the National FINRISK Study is presented in Figure 1.3.

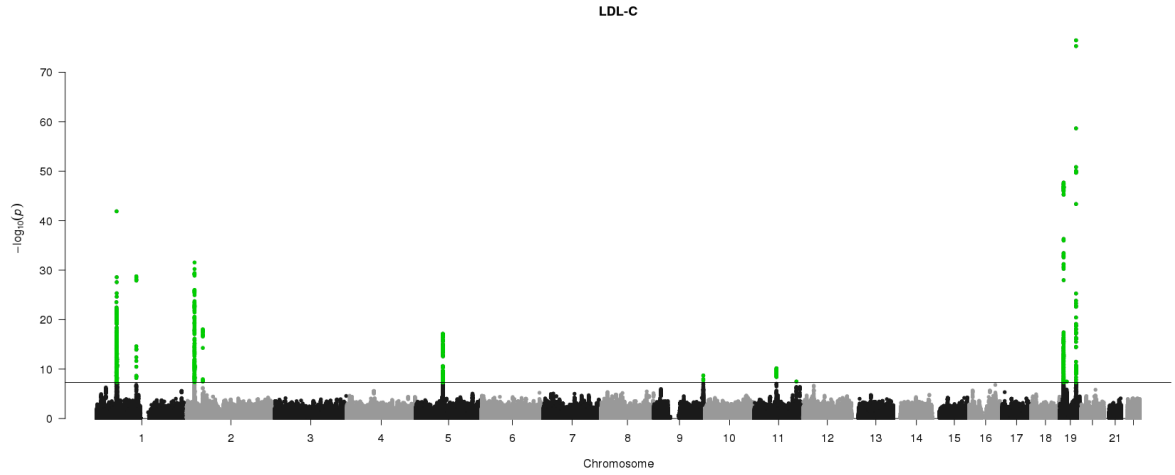


Figure 1.3: Manhattan plot of the genome-wide association study results of LDL-C, where x-axis displays chromosomal positions as basepairs and y-axis the strength of the association as the negative logarithm of the association p-value from linear regression. The horizontal solid line at p=5×10^{-8} is a genome wide significance (GWS) threshold and associations exceeding the GWS are highlighted as green dots. Loci with the strongest associations stand out as high peaks.

1.3 Linkage disequilibrium

Linkage disequilibrium (LD) is defined as a non-random association of alleles at different loci [Slatkin, 2008] that occurs due to the fragmented *recombination* in germ cells. Recombination is a process during the formation of gametes that leads to the formation of new allele combinations on chromosome. LD can be used to study different evolutionary and demographic events in a population history, such as natural selection, *genetic drift* - that is change in allele frequencies of a population over generations due to chance - and mutations, and also for the studies of genetic associations with quantitative traits and diseases. LD depends on the local recombination rates and of the genetic distance between genetic markers, measured in centimorgans (cM). The closer the markers are on

a chromosome, more likely they share similar genealogies, and less likely there is recombination between them. The amount of LD is usually higher between close relatives than between unrelated individuals because there have been less possible recombination events (less generations of gamete production).

Consider two loci in a genome with allele A at locus 1 at frequency p_A, allele B at locus 2 at frequency p_B, and AB *haplotype* - that is a set of alleles from the same chromosome inherited together as a unit - at frequency p_{AB} in a population. If the two loci are independent then the expected frequency of haplotype AB would be $p_A p_B$. If $p_{AB} - p_A p_B \neq 0$, then the two loci are in LD.

LD can be measured in several ways and one commonly used measure is a squared correlation coefficient (r^2) which is defined as

$$r^2 = \frac{D_{AB}^2}{p_A(1-p_A)p_B(1-p_B)}, \tag{1.1}$$

where $D_{AB} = p_{AB} - p_A p_B$. In practice, instead of population frequencies, sample frequencies are used in the estimation.

Human genome is hypothesized to contain several varying sized haplotype blocks which are defined as different non-overlapping sets of loci that are in a strong LD with each other [Slatkin, 2008]. Block sizes vary between few kilo bases to over hundred kilo bases. Figure 1.4 shows an example of haplotype block structure as pairwise correlations between 102 SNPs in PCSK9 gene in chromosome 1. LD and the haplotype block structure of the genome is utilized in GWAS, since one SNP of each block partially tags information of all the others SNPs on that block, and reduces the amount of SNPs that need to be tested for association. Therefore GWAS also gives similar results to a large number of variants that are in LD together [Vukcevic et al., 2011]. However, since the haplotype blocks vary between human populations, the population structure needs to be accounted. LD is affected by subpopulations, because the allele frequency differences create additional covariance and the measure used for estimating LD, D, is a measure of covariance between alleles at different loci. Also, LD in the population is not constant, instead it varies with the changes in population size and with migration rates between populations.

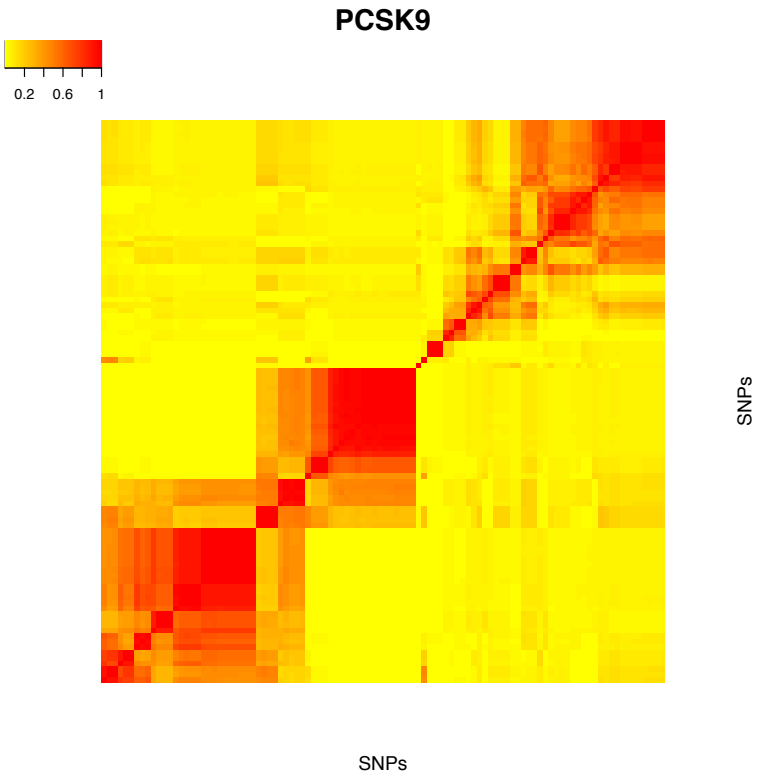


Figure 1.4: Pairwise correlations (r^2) between 102 SNPs from the National FINRISK Study in gene PCSK9 in chromosome 1 showing the patterns of haplotype blocks.

1.4 Inflation in GWAS results and genomic control

In a GWAS, it is essential to distinguish true *polygenicity* - that is many small genetic effects affecting the phenotype - from confounding biases, such as *population structure*, *cryptic relatedness* and genotyping errors. Population structure refers to differences in allele frequencies between subpopulations in a population, that could be due to for example different ancestry or non-random mating, and cryptic relatedness refers to the presence of close relatives in a sample of seemingly unrelated individuals. Especially, a case-control study of disease with genetic basis is susceptible to both population structure and cryptic relatedness [Devlin and Roeder, 1999]. In a case-control study, cryptic relatedness may occur because cases are often related whereas controls are more likely to be independent. In addition, cases are usually oversampled in contrast to controls and confounding may occur if the cases are members of an unobserved subpopulation. In a case-control design,

allele frequencies between the cases and controls are compared and tested whether they differ between the groups. Therefore, differences in allele frequencies due to population structure or cryptic relatedness can inflate the GWAS test statics.

However, if there is true polygenic inheritance, some genomic inflation is expected in the absence of population structure and other confounders. A genomic control (λ_{GC}) [Devlin and Roeder, 1999] is a conservative method for measuring and correcting the inflation of the GWAS and meta-analysis test statistics but it can not distinguish true polygenicity from confounding bias. Genomic control is based on the fact that even though a small fraction of the SNPs show true association with the disease or trait, most of the SNPs show no association, and thus, under the true null hypothesis, the test statistics of the SNPs should follow the distribution under the null hypothesis of no association between a SNP and the trait [Yang et al., 2011b]. In practice, inflation of the GWAS test statistics is corrected by dividing the χ^2 association statistics by λ_{GC}. Both the expected λ_{MEAN} and λ_{MEDIAN} can be used as genomic control. λ_{MEDIAN} is defined as a ratio between the median of the observed distribution of the test statistics and the expected median (=0.456). λ_{MEAN} is defined as a ratio between mean of the observed test statistics and the expected mean. Nonetheless, especially in large meta-analysis studies, adjusting by genomic control may be too conservative and decrease the power of test to detect true association with traits and diseases [Yang et al., 2011b]. Because the true polygenicity is to be distinguished from confounding bias - and LD Score regression [Bulik-Sullivan et al., 2015] is able to distinguish different sources of inflation in the GWAS test statistics - LDSC is preferred over genomic control.

1.5 Heritability

Heritability is a population parameter which estimates the proportion of phenotypic variation that can be explained by genetic factors [Visscher et al., 2008]. It is specific to a certain population in a particular environment and can change even without any changes at genetic level. An estimation of heritability is conducted by comparing phenotypic variation between differently related individuals in a specific population. There are two kinds of heritability estimates: a broad-sense heritability (H^2) is a measurement of the proportion of genetic variance from the total phenotypic variance, whereas a narrow-sense heritability (h^2) is a measurement of the additive genetic variance from the total phenotypic variance.

The broad-sense heritability is estimated as

$$H^2 = \frac{\sigma_G^2}{\sigma_P^2},$$

where σ_G^2 is genetic variance and σ_P^2 phenotypic variance. Phenotypic variance can be partitioned into genetic and environmental variance components and genotype-by-environment interaction variance component, $\sigma_P^2 = \sigma_G^2 + \sigma_E^2 + \sigma_{E\times G}^2$. In the studies of complex traits, it is often assumed that the interaction variance is very low and can be combined with the environmental variance [Klug et al., 2012]. In addition, the genetic variance can be partitioned into additive, dominance and epistatic variances; $\sigma_G^2 = \sigma_A^2 + \sigma_D^2 + \sigma_I^2$. Epistatic variance, σ_I^2, is assumed to be negligible so the genetic variance in the quantitative trait loci is usually estimated from the allelic effects that are either additive or dominant/recessive. Heritability varies between 0 and 1; low values indicating that environmental factors are mostly responsible for the phenotypic variation, and high values that genetic factors explain most of the phenotypic variation.

Narrow-sense heritability is estimated as

$$h^2 = \frac{\sigma_A^2}{\sigma_P^2}$$

Chip-heritability - also referred as SNP-heritability - is defined as the proportion of phenotypic variance explained by additive effects of genotyped variants, usually (SNPs). SNP-heritability is estimated as

$$h^2_{\text{SNP}} = \frac{\sigma^2_{\text{SNP}\in S}}{\sigma_P^2}, \tag{1.2}$$

where S is the set of SNPs used in the estimation. Usually $h^2_{\text{SNP}} \leq h^2 \leq H^2$, because GWAS genotyping arrays do not contain all variants in the genome, and h^2_{SNP} serves as a lower bound estimate of narrow-sense heritability. LD Score regression can be used for estimating SNP-heritability, and stratified LD Score regression for partitioning SNP-heritability by functional annotations.

Chapter 2

Materials

2.1 The National FINRISK Study

The National FINRISK Study [Borodulin et al., 2015] was a population-based health examination survey performed every five years from 1972 to 2012. FINRISK was coordinated by National Institute for Health and Welfare (THL) and studied risk factors of chronic and noncommunicable diseases, cardiovascular disease being one of the main interests. For each survey, independent random samples from different parts of Finland were drawn from the national population register from population aged from 25 to 74 years. In 2017, the National FINRISK Study was joined with Health 2000 Survey forming a new population study, the National FinHealth Study.

Genotypes

In this study, I used genotype data from the following four FINRISK cohorts: FR92, FR97, FR02 and FR07. All cohorts had been genotyped by Illumina genotyping chips and genotype imputation [Marchini and Howie, 2010] had been performed with a merged 1000 Genomes and Finnish whole genome sequencing reference panel. I used the genotypes for four separate genome-wide association studies of lipid levels. In addition, I used genotypes from a subset of individuals as an LD reference panel for the LD Score estimation. To control for possible chip effects, for the LD Score estimation I used a harmonised imputed genotype data including only individuals genotyped by Illumina HumanCoreExome chip

(n=10,659) which included three batches.

Phenotypes: lipid level concentrations

From the FINRISK study, I used four quantitative traits related to circulating lipid level concentrations as phenotypes: high-density lipoprotein cholesterol (HDL-C), low-density lipoprotein cholesterol (LDL-C), triglycerides (TG) and total cholesterol (TC). Lipids are fat-soluble compounds and an essential component of cell-membrane. They have several important biological functions, such as storing and releasing energy, and participating in cell signaling and hormone action. Elevated LDL-C and TC levels, as for example in dyslipidemia, are found to be associated with an increased risk for heart disease [Rana et al., 2010].

Main function of HDL-C, that is also called 'good' cholesterol, is to transport cholesterol from tissues to the liver, whereas main function of LDL-C, also called 'bad' cholesterol, is to transport cholesterol for the cells. Triglycerides - which are compounds of fatty acids and glycerides - are an important storage and source of energy.

2.2 The 1000 Genomes Project

The 1000 Genomes Project [The 1000 Genomes Project Consortium, 2015] aimed to characterize the variation of human genome by sequencing the genomes of individuals from different populations and to find the most genetic variants with at least one percent frequencies in each population. A public database (www.internationalgenome.org) was created for the use of scientists to study the relationship between genotype and phenotype, that is essential for example to understand the biology behind the diseases. The project was conducted in four phases between years 2008 and 2015. Besides a pilot phase, there were three phases of the main project for data production and technical method development. The first two main phases covered only bi-allelic sites, but the final phase was expanded to cover also multi-allelic sites, indels and structural variants. The project was completed in 2015 and the database now contains genomes of 2054 individuals from 26 populations covering over 88 million variants.

In this study, I used three different subsets from the 1000 Genomes Project Phase 3 data as an external LD reference panel: 99 Finnish individuals were used to form a Finnish-specific panel (1KG FIN), 504 European individuals including the Finnish were used to

form a multi-ethnic European panel (1KG all EUR) and 405 non-Finnish Europeans were used to form a panel that possibly represents a population mismatch (1KG non-Finnish EUR) with the Finnish GWAS statistics.

Chapter 3

Methods

3.1 Overview

In this chapter, the main focus is on LD Score regression (LDSC) that estimates heritability by regressing GWAS summary statistics on LD Scores. I will start the chapter by introducing LD Scores and how the LD Scores are estimated. Second, I will give a brief overview of linear regression and its parameter estimation. Next, I will first give an overview of the LDSC that is followed by a more detailed description of two LD Score regression models: first, an univariate LD Score regression model that is used to estimate the total SNP-heritability and second, a stratified LD Score regression model that is used to partition heritability by functional annotations. Finally, I conclude the chapter by describing how LDSC intercept can be used to correct confounding bias in GWAS summary statistics.

3.2 LD Scores

LD Score [Bulik-Sullivan et al., 2015] of a SNP j measures the amount of genetic variation tagged by j and is defined as the sum of squared Pearson correlation coefficients, r_{jk}^2, between j and all other SNPs k:

$$l_j := \sum_{k=1}^{M} r_{jk}^2, \tag{3.1}$$

where M is the number of SNPs.

A category specific LD Score [Finucane et al., 2015] of SNP j is defined as:

$$l(j, C) = \sum_{k \in C} r_{jk}^2, \tag{3.2}$$

where C is the set of SNPs belonging to the category of interest. Instead of summing over all SNPs, the category specific LD Score is the sum of squared Pearson correlation coefficients over all SNPs in the category.

3.2.1 Estimating LD Scores

The square of the standard estimator of correlation between SNPs j and k has approximately $\mathrm{E}[\hat{r}_{jk}^2] \approx r_{jk}^2 + \frac{(1-r_{jk}^2)}{N}$ [Bulik-Sullivan et al., 2015], and an expected LD Score of SNP j is:

$$
\begin{aligned}
\mathrm{E}\left[\sum_{k=1}^{M} \hat{r}_{jk}^2\right] &= \sum_{k=1}^{M} E[\hat{r}_{jk}^2] \\
&\approx \sum_{k=1}^{M} \left(r_{jk}^2 + \frac{(1-r_{jk}^2)}{N}\right) \\
&= \sum_{k=1}^{M} r_{jk}^2 + \sum_{k=1}^{M} \frac{(1-r_{jk}^2)}{N} \\
&= l_j + \sum_{k=1}^{M} \frac{1}{N} - \sum_{k=1}^{M} \frac{r_{jk}^2}{N} \\
&= l_j + \frac{M}{N} - \frac{1}{N} \sum_{k=1}^{M} r_{jk}^2 \\
&= l_j + \frac{M}{N} - \frac{1}{N} l_j \\
&= l_j + \frac{M - l_j}{N}
\end{aligned}
\tag{3.3}
$$

Because the standard estimator for Pearson correlation coefficient is biased upward, LD Scores are estimated by an approximately unbiased estimator:

$$
\begin{aligned}
\hat{r}^2_{adj} &= \hat{r}^2 - \frac{1-\hat{r}^2}{N-2} \\
&= r^2_{jk} + \frac{(1-r^2_{jk})}{N} - \frac{1-\left(r^2_{jk} + \frac{(1-r^2_{jk})}{N}\right)}{N-2} \\
&= r^2_{jk} + \frac{(N-2)(1-r^2_{jk})}{N(N-2)} - \frac{N-N\left(r^2_{jk} + \frac{(1-r^2_{jk})}{N}\right)}{N(N-2)} \\
&= r^2_{jk} + \frac{N-Nr^2_{jk}-2+2r^2_{jk}}{N(N-2)} - \frac{N-Nr^2_{jk}+(1-r^2_{jk})}{N(N-2)} \\
&= r^2_{jk} + \frac{3r^2_{jk}-3}{N(N-2)} \\
&\overset{*}{\approx} r^2_{jk} \\
&* \left(\frac{3r^2_{jk}-3}{N(N-2)} \to 0, \text{when } N \to \infty\right),
\end{aligned}
\tag{3.4}
$$

where $\hat{r}^2$ is the standard estimator of the squared Pearson's correlation and N is sample size. Because r^2_{jk} is between $[0, 1]$ and N in a typical GWAS is in thousands, $| \frac{3r^2_{jk}-3}{N(N-2)} | \leq \frac{3}{N(N-2)} \approx 0$. Even though the estimator is not completely unbiased, I chose to use it because it was reported in the original publication [Bulik-Sullivan et al., 2015].

3.3 Linear regression

This section is based on the book Linear Regression Analysis by [Seber et al., 2003].

Linear regression models the relationship between a dependent variable and an explanatory variable by a linear equation. Simple linear model has only one explanatory variable and multiple linear regression model has at least two explanatory variables. The strength of the relationship is measured by the coefficient of determination, expressed usually either as r^2 in a simple linear regression or R^2 in a multiple linear regression.

3.3.1 Linear models and parameter estimation

A simple linear regression model with one explanatory variable can be written as:

$$y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i, \qquad i = 1, ..., n$$

where y_i is a dependent variable, x_{i1} is an explanatory variable, β_0 is an intercept term, β_1 is a regression coefficient and ε_i is an error term.

A multiple linear regression model is similar to the simple linear model, but now there are at least two explanatory variables:

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{i,p} + \varepsilon_i, \qquad i = 1, ..., n$$

In a matrix form, the model is presented as

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} x_{1,0} & x_{1,1} & \dots & x_{1,p} \\ x_{2,0} & x_{2,1} & \dots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,0} & x_{n,1} & \dots & x_{n,p} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where the first column of $\boldsymbol{X}$ is a constant 1 corresponding to intercept parameter β_0.

Assumptions

Linear regression model has several assumptions: the dependent variable should be measured at least on an ordinal scale, the sample must be representative of the population to which the inference will be made and relationship between dependent and explanatory variables has to be linear. For optimality of the typical parameter estimator (ordinary least squares method), the error terms are assumed to have mean zero and to be uncorrelated and homoscedastic meaning that they have an approximately constant variance. Furthermore, for statistical inference, it is often assumed that error terms are normally

distributed.

Ordinary least squares (OLS)

The most common method for estimating regression coefficients is ordinary least squares (OLS) method, which fits the regression line by minimizing the sum of squared deviations from the fitted line to the observed values, called residuals , see Figure 3.1.

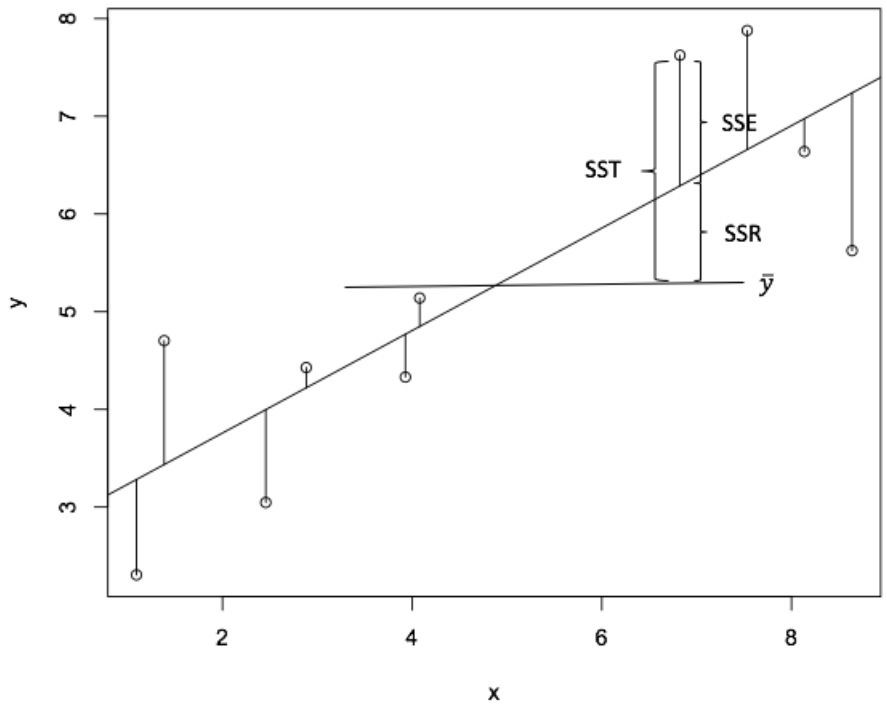


Figure 3.1: Ordinary least squares (OLS) method in a simple linear regression aims to minimize the sum of squared residuals (SSE). Total sum of squares (SST) is the sum of the sum of squares due to regression (SSR) and the sum of squared residuals (SSE): SST= SSR+SSE.

With a simple linear regression model, let's define squared sums as:

- Total sum of squares SST $= \sum_{i=1}^{n}(y_i - \bar{y})^2$
- The sum of squares due to regression SSR $= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$
- The sum of squared residuals SSE $= \sum_{i=1}^{n} \hat{\varepsilon_i}^2$

where $\hat{y}_i = \boldsymbol{x_i^T}\hat{\boldsymbol{\beta}}$ is the the fitted line and $\hat{\varepsilon}_i = y_i - \hat{y}_i$ are the residuals, and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. SST quantifies how much the observed data points vary around the mean, SSR quantifies how far the fitted regression line is from the mean and the SSE quantifies how much the observed data points vary around the fitted regression line.

The best fitted line is obtained by minimizing SSE $= \sum_{i=1}^{n} \hat{\varepsilon_i}^2 = \sum_{i=1}^{n}(y_i - \boldsymbol{x_i^T}\hat{\boldsymbol{\beta}})^2 = (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$ with respect to $\hat{\boldsymbol{\beta}}$. If we assume that the columns of $\boldsymbol{X}$ are linearly independent, then there is a unique solution to the minimization and the OLS estimate [Seber et al., 2003] for $\boldsymbol{\beta}$ is:

$$\hat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{(X^T X)^{-1} X^T y}.$$

If we also assume that ε has mean 0 and variance σ^2, then the estimator $\hat{\boldsymbol{\beta}}_{OLS}$ has following properties:

$$\mathrm{E}[\hat{\boldsymbol{\beta}}_{OLS}] = \boldsymbol{\beta}$$
$$\mathrm{Var}[\hat{\boldsymbol{\beta}}_{OLS}] = \sigma^2\boldsymbol{(X^T X)^{-1}}$$

Coefficient of determination, R^2

The sums of squares are related to the coefficient of determination, R^2, that measures the strength of the relationship between the dependent variable and the explanatory variables. R^2 is defined as one minus the ratio between the sum of squared residuals (SSE) and total sums of squares (SST). When there is an intercept term in the model SST = SSR + SSE [Seber et al., 2003] and then the coefficient of determination is

$$R^2 = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}} = \frac{\mathrm{SSR}}{\mathrm{SST}}.$$

Weighted least squares (WLS)

If the error terms do not have constant variance and therefore the model does not satisfy the assumption of homoscedasticity, instead of OLS, regression should be performed with weighted least squares (WLS) method which corrects the heteroscedasticity. Identity matrix $\boldsymbol{I}$ of the error terms is replaced with more general diagonal matrix $\boldsymbol{V}$,

$$\mathrm{Cov}(\boldsymbol{\varepsilon_i}) = \boldsymbol{V} = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_n^2 \end{bmatrix}.$$

If a reciprocal of each variance, σ_i^2, is defined to be a weight, $w_i = \frac{1}{\sigma_i^2}$, $w_i > 0$ and matrix $\boldsymbol{W}$ to be a diagonal matrix containing these weights such that $\boldsymbol{W} = \boldsymbol{V}^{-1}$, then the weighted least squares estimate for the $\boldsymbol{\beta}$ [Seber et al., 2003] is:

$$\hat{\boldsymbol{\beta}}_{WLS} = (\boldsymbol{X^T V^{-1} X})^{-1} \boldsymbol{X^T V^{-1} y} = (\boldsymbol{X^T W X})^{-1} \boldsymbol{X^T W y},$$

and

$$\mathrm{E}[\hat{\boldsymbol{\beta}}_{WLS}] = \boldsymbol{\beta}$$
$$\mathrm{Var}[\hat{\boldsymbol{\beta}}_{WLS}] = (\boldsymbol{X^T W X})^{-1}$$

3.4 LD Score regression

In this section, I will first give an overview of the LDSC method, that is followed by a detailed derivation of univariate LD Score regression model and stratified LD Score regression model. Then, I will give a description how LDSC conforms to the linear model assumptions. Lastly, I will conclude the section by shortly describing how LDSC intercept can be used to measure and correct confounding bias in GWAS summary statistics.

3.4.1 Overview

This section is based on [Bulik-Sullivan et al., 2015]

LD Score regression (LDSC) estimates heritability by regressing GWAS summary statistics on LD Scores. In a GWAS, test statistic distribution can be inflated by both polygenicity - meaning that many small genetic effects together affect the phenotype - and confounding bias such as cryptic relatedness and population stratification. To recognize the true genetic association, different inflation factors need to be distinguished. LD Score regression is able to quantify both contributions by fitting a linear regression model between GWAS summary test statistics and LD Scores. Test statistics of variants that are in LD with the causal variant increase in proportion to the squared correlation with the causal variant, whereas LD and inflation from confounding bias - due to cryptic relatedness and/or population stratification from genetic drift - will not correlate.

LDSC model is based on a simple linear regression and in a polygenic model the expected χ^2 statistics is defined as:

$$\mathrm{E}\left[\chi_j^2 | l_j\right] = 1 + Na + \frac{Nh_g^2}{M} l_j, \tag{3.5}$$

where the intercept term $(1 + Na)$ measures the amount of environmental variance, where a is inflation due to confounding bias such as cryptic relatedness and/or population stratification, and regression slope measures the polygenic effects. N is GWAS sample size, M is total number of SNPs, $\frac{h_g^2}{M}$ is the average heritability explained per SNP, and l_j is the LD Score of variant j. Derivation for the formula is provided in section 3.4.2.

SNP-heritability is estimated by regressing χ^2 statistics against LD Scores, see Figure 3.2 and rescaling the slope by the number of common SNPs (MAF > 0.05) used in the LD Score estimation (M) and by the sample size of GWAS study (N). The SNP-heritability which LDSC aims to estimate, is the proportion of phenotypic variance explained by the best linear predictor comprised of common SNPs in the LD Score reference panel.

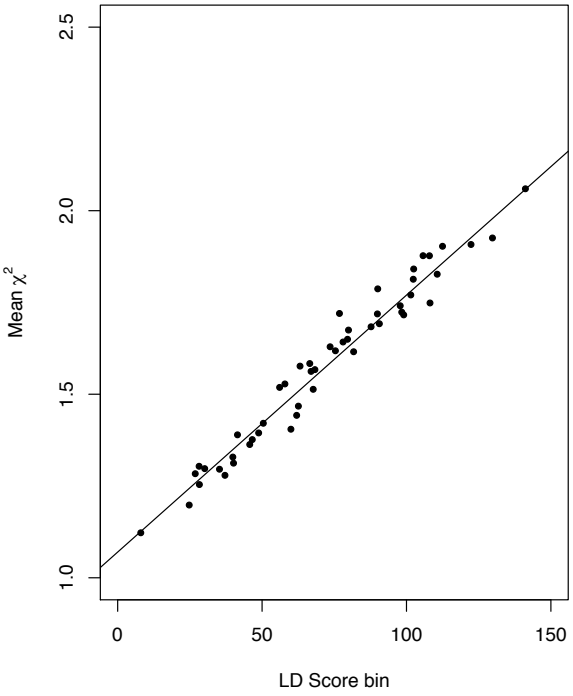


Figure 3.2: LD Score regression plot, where each point represents an LD Score quantile: x coordinate is the mean LD Score of variants in each quantile and y coordinate the mean χ^2 statistic of variants in the corresponding quantile. The black line is the LD Score regression line fitted by an equation 3.5. SNP-heritability is estimated by regressing χ^2 statistics from a GWAS on LD Scores.

As an input, LDSC needs only the summary statistics from a GWAS or from a meta-analysis, and LD Scores estimated from a reference panel - which can be external - that matches to studied population. In particular, LDSC does not require any individual level data, which is a big advantage in practice. LDSC can be applied for both quantitative and binary traits.

3.4.2 LD Score regression models

Univariate LD Score regression

LD Score regression assumes a standard additive polygenic model of a quantitative phenotype defined as

$$y_i = \sum_{j=1}^{M} \gamma_j x_{ij} + \varepsilon_i \tag{3.6}$$

where y_i is a standardized phenotype of an individual i, x_{ij} a standardized genotype of individual i at SNP j, γ_j an effect of SNP j and ε_i is an error term. In a GWAS, association for each SNP is tested separately

$$y_i = x_{ij}\beta_j + e_j, \tag{3.7}$$

where β_j is a marginal effect of SNP j. Estimate of the marginal effect $\hat{\beta}_j$ includes all coefficients from the true model weighted by correlation with all other SNPs:

$$\hat{\beta}_j = \sum_{k=1}^{M} \gamma_k r_{jk} + s_j + e_j, \tag{3.8}$$

where r_{jk} is a correlation between SNPs j and k, s_j is a bias from confounders such as population stratification or cryptic relatedness and e_j an estimation error.

In the polygenic model, that includes multiple SNPs with small effects, γ is assumed to be random with mean zero and variance of per-SNP heritability: $\mathrm{E}[\gamma_k] = 0, \mathrm{Var}[\gamma_k] = \frac{h^2_{\mathrm{SNP}}}{M}$. Also, $\mathrm{E}[s_j] = 0, \mathrm{Var}[s_j] := a$ and $\mathrm{E}[e_j] = 0, \mathrm{Var}[e_j] = \frac{\sigma^2_e}{N}$. Furthermore, genotypes, effect sizes and errors are assumed to be mutually independent meaning that:

$$\begin{aligned} &\mathrm{E}[\gamma_j \gamma_k] = 0, j \neq k \\ &\mathrm{E}[s_j \gamma_k] = 0 \\ &\mathrm{E}[e_j \gamma_k] = 0 \\ &\mathrm{E}[s_j e_j] = 0 \end{aligned} \tag{3.9}$$

χ^2 -statistics of SNP j from the GWAS is defined as

$$\begin{aligned}
\chi_j^2 &:= \left(\frac{\hat{\beta}_j}{\mathrm{se}(\hat{\beta}_j)}\right)^2 \\
&= \frac{(\hat{\beta}_j)^2}{\mathrm{Var}(e_j)} \\
&\approx \frac{(\hat{\beta}_j)^2}{\frac{1}{N}} \\
&= N(\hat{\beta}_j)^2 \\
&= N\left(\sum_{k=1}^{M} \gamma_k r_{jk} + s_j + e_j\right)^2
\end{aligned} \tag{3.10}$$

where the approximation follows from the polygenicity assumption that variance explained by a given SNP is small, and therefore an error variance is approximately the phenotypic variance: $\sigma_e^2 \approx 1$ leading to $e_j \sim N(0, \frac{1}{N})$.

An expected χ^2 -statistics of SNP j is

$$
\begin{aligned}
\mathrm{E}\left[\chi_j^2\right] &= \mathrm{E}\left[N\left(\sum_{k=1}^{M}\gamma_k r_{jk} + s_j + e_j\right)^2\right] \\
&= N\mathrm{E}\left[\left(\sum_{k=1}^{M}\gamma_k r_{jk} + s_j + e_j\right)^2\right] \\
&= N\mathrm{E}\left[\left(\sum_{k=1}^{M}\gamma_k r_{jk}\right)^2 + s_j\sum_{k=1}^{M}\gamma_k r_{jk} + e_j\sum_{k=1}^{M}\gamma_k r_{jk} + \right. \\
&\left. s_j\sum_{k=1}^{M}\gamma_k r_{jk} + s_j^2 + s_j e_j + e_j\sum_{k=1}^{M}\gamma_k r_{jk} + e_j s_j + e_j^2\right] \\
&\overset{*}{=} N\left(\mathrm{E}\left[\left(\sum_{k=1}^{M}\gamma_k r_{jk}\right)^2\right] + \mathrm{E}[s_j^2] + \mathrm{E}[e_j^2]\right) \\
&= N\left(\sum_{k=1}^{M}\mathrm{E}\left[\gamma_k^2\right] r_{jk}^2 + \mathrm{Var}(s_j) + \mathrm{Var}(e_j)\right) \\
&= N\left(\mathrm{Var}(\gamma_k)\sum_{k=1}^{M} r_{jk}^2 + a + \frac{1}{N}\right) \\
&= N\frac{h_{\mathrm{SNP}}^2}{M}\sum_{k=1}^{M} r_{jk}^2 + aN + 1 \\
&= 1 + Na + \frac{Nh_{\mathrm{SNP}}^2}{M} l_j \\
&* e_j \perp\!\!\!\perp \gamma_k, \gamma_k \perp\!\!\!\perp s_j, s_j \perp\!\!\!\perp e_j, \mathrm{E}[\gamma_k] = 0, \mathrm{E}[s_j] = 0, \mathrm{E}[e_j] = 0,
\end{aligned}
\tag{3.11}
$$

where χ_j^2 is the marginal summary association statistic from the GWAS that contains inflation from three different sources that now can be distinguished by a simple linear regression against the LD Score, l_j. The SNP-heritability is estimated by rescaling the regression slope, and the confounding bias due to the cryptic relatedness and population structure is included in the regression intercept and is obtained by subtracting 1 from the intercept.

In LDSC, standard errors (s.e.) are estimated by a block jackknife [Shao and Wu, 1989] - also called delete-d jackknife - over the block of SNPs, because the LD Scores are

correlated. The block jackknife is an iterative resampling method where the sample is divided into equal sized blocks, and new jackknife subsamples are formed by systematically leaving out each block at a time. At first, the parameter of interest is estimated from the whole sample. Next, each block at a time is omitted, and the parameter is estimated from the jackknife subsample. Then, *pseudovalues* are computed as the difference between the whole data estimate and the estimate of the subsample. Finally, the jackknife estimate of the parameter is obtained from the pseudovalues, and standard error is estimated from the standard deviation of the pseudovalues.

Stratified LD Score regression model

This section is based on [Finucane et al., 2015].

Stratified LD Score regression (S-LDSC) is an extension of the LD Score regression for partitioning heritability by functional categories. Instead of assuming a constant variance for effects sizes γ, variance can vary depending on the category:

$$\mathrm{Var}[\gamma_j] = \sum_{c:j\in C_c} \tau_c, \tag{3.12}$$

where C_c indexes the SNPs in c:th category. Also, instead of a single LD Score and a simple linear regression, S-LDSC model includes different LD Score for each category and estimates for each τ_c are obtained via multiple linear regression.

Heritability of a category C is defined as

$$h^2(C) = \sum_{j\in C} \gamma_j^2. \tag{3.13}$$

For a polygenic trait, Finucane et al. determine a category of SNPs to be enriched for heritability if SNPs with high LD to that category have higher χ^2 statistics compared to SNPs with low LD to that category. Furthermore, an enrichment of a category is defined as a proportion of the SNP-heritability in the category divided by the proportion of SNPs:

$$\text{Enrichment} = \frac{h^2_{\mathrm{SNP}}(C_c)/M(C_c)}{h^2_{\mathrm{SNP}}/M} \tag{3.14}$$

Model is same as in 3.6 and also χ^2 statistic of SNP j is same:

$$\begin{aligned}\chi_j^2 &:= \left(\frac{\hat{\beta}_j}{\mathrm{se}(\hat{\beta}_j)}\right)^2 \\ &= N\left(\sum_k \gamma_k r_{jk} + s_j + e_j\right)^2,\end{aligned} \tag{3.15}$$

but a category specific LD Score of SNP j is defined as

$$l(j,c) := \sum_{k\in C_c} r_{jk}^2. \tag{3.16}$$

Now the expected χ^2 -statistics of SNP j is

$$\begin{aligned}\mathrm{E}\left[\chi_j^2\right] &= \mathrm{E}\left[N\left(\sum_k \gamma_k r_{jk} + s_j + e_j\right)^2\right] \\ &= N\mathrm{E}\left[\left(\sum_k \gamma_k r_{jk} + s_j + e_j\right)^2\right] \\ &= N\left(\mathrm{E}\left[(\sum_k \gamma_k r_{jk})^2\right] + \mathrm{E}[s_j^2] + \mathrm{E}[e_j^2]\right) \\ &= N\left(\sum_k \mathrm{E}\left[\gamma_k^2\right] r_{jk}^2 + \mathrm{Var}(s_j) + \mathrm{Var}(e_j)\right) \\ &= N\left(\sum_c \tau_c \sum_{k\in C_c} r_{jk}^2 + a + \frac{1}{N}\right) \\ &= N\left(\sum_c \tau_c l(j,c) + a + \frac{1}{N}\right) \\ &= 1 + Na + N\sum_c \tau_c l(j,c)\end{aligned} \tag{3.17}$$

where in disjoint categories $\tau_c = h^2(C_c)/M(C_c)$ is the per-SNP heritability in category C_c and $M(C_c)$ is the number of SNPs in the category. In overlapping categories per-SNP heritability of the SNP j is $\sum_{c:j\in C} \tau_c$.

Interpretation of parameters τ_c and $h^2(C)$ differ: $h^2(C)$ is a more robust quantity and is defined as the sum of squared effects of SNPs in category C and should be independent of the categories chosen to be in the model. Because τ_c is a contribution of a category C_c after controlling for all other categories in the model, it is dependent on the choice of categories included in the model. [Finucane et al., 2015]

Three models to partition heritability by S-LDSC: a full baseline model, and models with specific cell types or specific cell type groups

To partition heritability by functional annotations, Finucane et al. constructed a full baseline model that includes 53 overlapping functional categories. Categories had been formed from 24 publicly available non-cell-type-specific functional annotations including coding and evolutionary conserved regions, regulatory elements and histone marks. A more detailed description of each functional annotation is provided in Appendix B. To avoid inflation of heritability in flanking regions, Finucane et al. added 500-bp windows around each functional category forming additional 24 categories and 100-bp windows around chromatin immunoprecipitation and sequencing (ChIP-seq) peaks of the following marks: DHS, H3K4me1, H3K4me3 and H3K9ac, forming four additional categories. Also, a category containing all SNPs was included in the model.

Annotations for the models were obtained from several public sources - mostly from ENCODE [The ENCODE Project Consortium, 2007] and ROADMAP [Roadmap Epigenomics Consortium, 2015] data sources - and are listed with their post-processing procedures in Table 3.1. The RefSeq gene models from the human genome browser at UCSC [Kent et al., 2002] were used as a source for coding, 3'-UTR, 5'-UTR, intron and promoter annotations, and ENCODE was used as a source for DGF and TFBS. All these annotations were then post-processed by [Gusev et al., 2014]. [Hoffman et al., 2013] was used as a source for combined chromHMM/Segway annotations for six cell lines and an union over these six cell lines was taken to form categories for CTCF, promoter flanking, transcribed, TSS, strong enhancer and weak enhancer. Instead, a repressed category was formed by taking an intersection of the six cell lines. DHS category was formed by taking a union of all cell-type-specific annotations that were either from ENCODE or ROADMAP data. Fetal DHS category was formed as a union of only fetal cell types. Post-processing for these was done by [Trynka et al., 2013]. Different histone mark categories were formed by taking a union over all cell types for each histone mark. Annotations of H3K4me1, H3K4me3, H3K9ac H3K27ac were formed from ROADMAP data, and other version of H3K27ac from the data of [Hnisz et al., 2013]. H3K4me1, H3K4me3, H3K9ac and DHS were post-processed by [Trynka et al., 2013] and H3K27ac by PGC2 2004 [Schizophrenia

Working Group of the Psychiatric Genomics Consortium, 2014]. Super enhancer category was formed and post-processed by [Hnisz et al., 2013] and FANTOM5 enhancer was formed and post-processed by [Andersson et al., 2014]. Conserved category was formed from [Lindblad-Toh et al., 2011] and post-processed by [Ward and Kellis, 2012].

Besides the full baseline model, Finucane et al. constructed two models for partitioning heritability either by specific cell types or by cell type groups. At first, they used four different histone marks - H3K4me1, H3K4me3, H3K9ac and H3K27ac - that were specific for each cell type to form a model with 220 different cell-type-specific annotations. Next, they grouped each cell-type-specific annotation into ten different groups by taking a union of the histone marks within the group forming cell type groups: adrenal or pancreas, cardiovascular, central nervous system (CNS), connective or bone, gastrointestinal (GI), immune or hematopoietic, kidney, liver, skeletal muscle and other.

When performing the enrichment analysis with specific cell types or cell type groups, it is important to control over possible effects from functional categories such as coding. However, overlap with other cell types or cell type group categories should be enabled. Therefore, each cell type or cell type group is added separately to the full baseline model as an additional category forming 220 separate cell-type-specific models with 54 annotations or 10 separate cell-type-group-specific models with 54 annotations [Finucane et al., 2015].

Table 3.1: Annotation sources and post-processing procedures Finucane et al. used to construct different categories for S-LDSC

Annotation	**Source**	**Post-processed**
Coding, 3'-UTR, 5'-UTR, promoter, intron	UCSC	Gusev et al. AJHG (2014)
Digital genomic footprint (DGF), transcription factor binding site (TFBS)	ENCODE	Gusev et al. AJHG (2014)
CTCF, promoter flanking, transcribed, transcription start site (TSS), strong enhancer, weak enhancer, repressed	Hoffman et al. Nucleic Acids Res. (2013)	Hoffman et al. Nucleic Acids Res. (2013)
DNase I hypersensitivity sites (DHSs)	combination of ENCODE and ROADMAP data	Trynka et al. AJHG (2015)
Cell-type-specific H3K4me1, H3K4me3, H3K9ac	ROADMAP	Trynka et al. AJHG (2015)
Cell-type-specific H3K27ac(1)	ROADMAP	Schizophrenia Working Group of the Psychiatric Genomics Consortium, Nature (2014)
Cell-type-specific H3K27ac(2), super enhancer	Hnisz et al. Cell (2013)	Hnisz et al. Cell (2013)
Conserved	Lindblad-Toh et al. Nature (2011)	Ward and Kellis, Science (2012)
FANTOM5 enhancer	Andersson et al. Nature (2014)	Andersson et al. Nature (2014)

3.4.3 Regression weights and attenuation bias

LD Score regression model fails to satisfy three basic assumptions of linear model. First, the χ^2 statistics are not independent because of LD. Therefore, to improve the regression estimator, correlation is corrected by down-weighting each SNP in proportion to its LD to other SNPs used in the regression. Second, variance of the χ^2 statistics is not constant, instead χ^2 statistics of SNPs with high LD Score have higher variance than χ^2 statistics of SNPs with low LD Score. Heteroskedasticity is corrected by weighting with the reciprocal of the conditional variance function $\mathrm{Var}[\chi^2|l_j]$ [Bulik-Sullivan et al., 2015]. Third, the explanatory variable, LD Score, is not measured without error. If there is a measurement error in an explanatory variable, the regression slope will be biased towards zero, which is called attenuation bias. However, if the variance of this error is known, the bias can be corrected by multiplying the regression slope by a disattenuation factor. In LD Score regression, the squared weighted Pearson correlation between the true value and noisy estimates of the LD Scores is used to correct the attenuation bias, and standard errors for the LD Score estimates are estimated by a delete-one jackknife method over block of individuals.

Over-counting weight for a SNP j is defined as:

$$w_{oc}(j) := \frac{1}{1 + l_j(S)} \tag{3.18}$$

where S denotes the set of SNPs used in the regression.

In LDSC, heteroskedasticity weight for SNP j is defined as

$$w_{h_{\text{LDSC}}}(j) := \frac{1}{\left(1 + \frac{Nh_g^2}{M} l_j\right)^2}, \tag{3.19}$$

and in S-LDSC as

$$w_{h_{\text{S-LDSC}}}(j) := \frac{1}{\left(1 + N\hat{\tau} \sum_C l(j, C)\right)^2}, \tag{3.20}$$

Regression is then weighted by the product of over-counting weight and heteroskedasticity weight: $w_j = w_{oc}(j)w_h(j)$.

3.4.4 LD Score regression intercept and attenuation ratio

A parameter a - included in the intercept term $Na + 1$ of LD Score regression - estimates the proportion of confounding bias in GWAS summary statistics inflation. Therefore, when estimating heritability by LDSC, the intercept protects from confounding bias due to population structure and cryptic relatedness. Because of this property, LDSC intercept can also be used as a correction factor for inflated GWAS test statistics, and has been shown to be more accurate and retaining more power than traditionally used genomic inflation factor (λ_{GC}) [Bulik-Sullivan et al., 2015]. In practice, the inflated GWAS test statistics can be corrected by multiplying standard errors by the LDSC intercept.

In the model, there is an assumption of no systematic correlation between a fixation index F_{st} and LD Score. F_{st} measures between-population variance in allele frequencies and Wright's fixation index F_{st} is defined as the correlation of randomly drawn gametes from same subpopulation, relative to the total population and quantifies genetic differentiation between subpopulations [Wright, 1949]. When there is a positive correlation between F_{st} and a LD Score, the inflation contribution from the confounding term is underestimated leading to a biased intercept estimate. However, during simulations with confounding factors the observed correlations were negligible [Bulik-Sullivan et al., 2015].

Another measure to estimate confounding is a ratio between intercept and mean χ^2 statistics, denoted as an attenuation ratio in LDSC. It aims to estimate the relative balance of confounding and genetic effects and is defined as:

$$\frac{\text{Intercept} - 1}{\bar{\chi}^2 - 1} \tag{3.21}$$

Values close to zero indicate that most of the inflation in the test statistics is due to polygenic effects, whereas high ratio indicates high proportion of other sources of inflation such as population stratification or model misspecification, for example due to mismatch between LD reference panel and GWAS sample.

Chapter 4

Results

4.1 GWAS summary statistics

In this study, I ran the GWAS for the four lipid levels from the National FINRISK Study by SNPTEST v2.5.2 [Marchini et al., 2007]. TG had been log-transformed and all phenotypes had been stratified by cohort and adjusted by sex, age, age^2 and first ten principal components (PCs) of genetic population structure. I conducted the GWAS using the residuals which had been inverse-normal transformed to the standard normal distribution. Original data contained 20,627 individuals and close relatives had been excluded from all phenotypes. In addition, I excluded individuals on lipid-lowering medication from LDL-C and TC analyses. After the exclusions there were 16,727 individuals left for the GWAS of HDL-C and TG, 15,474 individuals left for LDL-C and 15,689 individuals for TC.

As quality control (QC) at variant level, I used following thresholds: SNPs with minor allele frequency (MAF) < 0.05 or imputation info score (INFO) < 0.9, or SNPs that deviated from Hardy-Weinberg equilibrium with $p < 1 \times 10^{-6}$ or had missing genotype calls > 0.02 were excluded. INFO is a quality metric varying between 0 to 1 that is used to measure the confidence of the imputation. As results, for HDL-C there were 837 genome-wide significant (GWS) SNPs ($p < 5 \times 10^{-8}$), and for LDL-C 1404, for TC 1021 and for TG 1195 GWS SNPs. Figure 4.1 shows the GWAS results for all four lipid levels as a circular Manhattan plot, where x-axis displays the genomic coordinates in basepairs and y-axis displays the negative logarithm of the association p-value for each SNP.

Next, I used LDSC to reformat the GWAS summary statistics for the required format to estimate the SNP-heritability and to further partition the heritability. The reformatting procedure of LDSC included additional variant QC that excluded strand-ambiguous SNPs, outliers using a threshold of $z^2 > 80$ and SNPs that did not match with LD Scores.

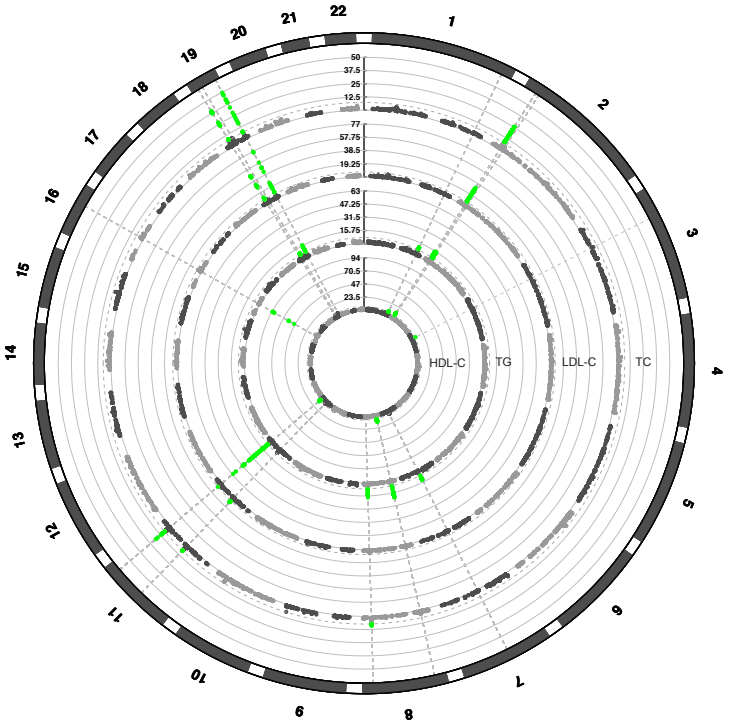


Figure 4.1: A circular Manhattan plot of the GWAS results of lipid levels in the National FINRISK Study, where x-axis displays chromosomal positions and y-axis displays $-log_{10}$(p-values) from adjusted linear regression. Associations exceeding the genome wide significance threshold at p= 5×10^{-8} are highlighted as green dots.

4.2 LD reference panels

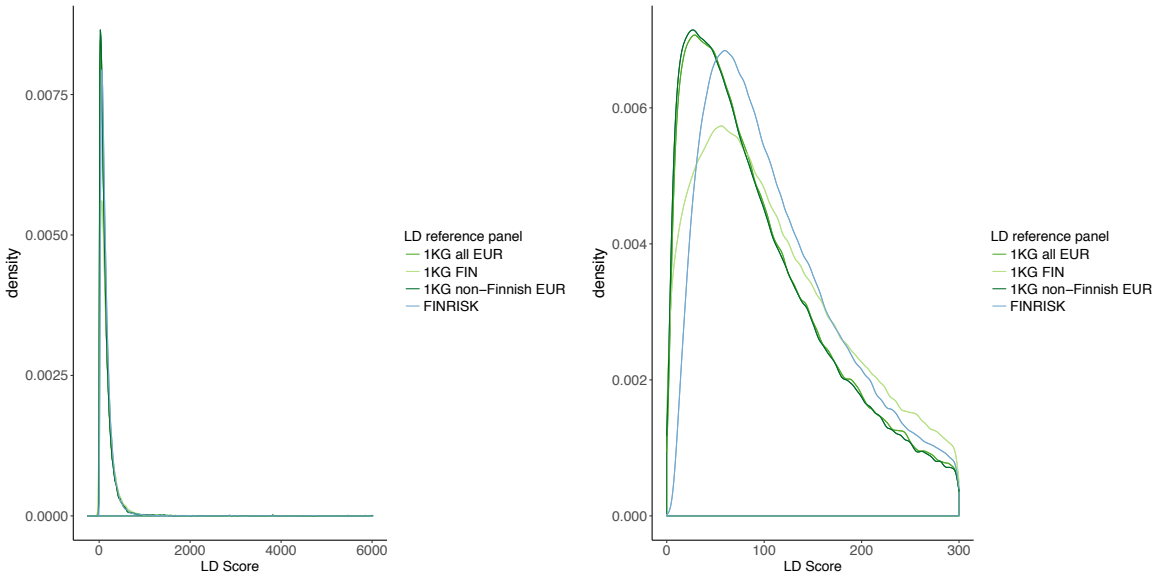
In this study, I used two different data sources as LD reference panels: the National FINRISK Study (FINRISK) data that included a subset of individuals used in the GWAS

representing an optimal LD structure, and publicly available 1000 Genomes Project (1KG) data as an external reference panel. From the National FINRISK study, I used a harmonised imputed genotype data of 10,659 individuals, and from the 1000 Genomes Project Phase 3 sequence data I used three different subsets: all Europeans (1KG all EUR) (n=504), Finns (1KG FIN) (n=99) and non-Finnish Europeans (1KG non-Finnish EUR) (n=405). I excluded indels, multiallelic variants and variants with MAF < 0.01. In addition, for the FINRISK panel I included only high imputation quality variants and excluded variants with INFO < 0.9. The accuracy of imputation of missing genotypes depends on LD patterns and frequency of variants [Marchini and Howie, 2010], and therefore using variants imputed with low-quality may bias the LD Score estimation. I estimated the LD Scores by LDSC v1.0.0 [Bulik-Sullivan et al., 2015] using one centimorgan (cM) window around an index variant, which has been shown to be a robust window size [Bulik-Sullivan et al., 2015].

Basic descriptive statistics of raw LD Scores for the different LD reference panels are shown in Table 4.1. There are some negative LD Scores, because of the bias correction in the LD Score estimator (equation 3.4) used in the LD Score estimation. Both Finnish specific panels have higher mean LD Score (FINRISK 163 and 1KG FIN 179) compared to the two multi-ethnic European panels (1KG all EUR 154 and 1KG non-Finnish EUR 152), which is consistent with previous studies [Service et al., 2006], [Exome Aggregation Consortium, 2016], [Bulik-Sullivan et al., 2015] about Finns having increased LD compared to rest of the Europeans due to recent genetic bottlenecks. All the distributions are highly right skewed, as shown in Figure 4.2a.

Table 4.1: Descriptive statistics of raw LD Scores for different LD reference panels

Panel	**#SNPs**	**Mean LD Score**	**Median LD Score**	**Min LD Score**	**Max LD Score**
FINRISK	7,948,601	162.84	113.93	0.466	4747.53
1KG FIN	10,247,012	178.76	116.36	-254.95	5487.56
1KG all EUR	9,755,791	153.66	94.73	-29.17	5898.57
1KG non-Finnish EUR	9,721,808	151.69	93.00	-43.21	6015.43



(a) Raw LD Scores

(b) Cleaned LD Scores

Figure 4.2: Density plots of LD Scores for each LD reference panels. (a) Densities of raw LD Scores. (b) Densities of cleaned LD Scores after excluding long range LD regions, MHC locus, ±3 cM around centromeres and outliers

Because LD Score regression methods are based on linear regression which is sensitive to the extreme observations, the outliers have been removed. For the LD Scores, a cutoff of 300 was used as being an outlier. Furthermore, Bulik-Sullivan et al. recommended to exclude at least following regions from the LD Score regression: long range LD regions listed in Supplementary Table S1, ± 3 cM regions around centromeres and a major histocompatibility complex (MHC) locus from chromosome six, which has an unusual LD structure. Basic descriptive statistics for the cleaned LD Scores after the exclusions for all four LD reference panels are presented in Table 4.2, and the distributions after exclusions are shown in Figure 4.2b. In addition, weights for the LD Scores were estimated including only variants used in the regression.

Table 4.2: Basic descriptive statistics for cleaned LD Scores for different LD reference panels after excluding long range LD regions, MHC locus, ±3 cM around centromeres and outliers

Panel	**#SNPs**	**Mean LD Score**	**Median LD Score**	**Min LD Score**	**Max LD Score**
FINRISK	6,469,765	113.23	98.50	0.63	300.00
1KG FIN	8,144,919	112.95	98.75	0.001	299.99
1KG all EUR	8,237,437	98.45	81.17	0.001	299.99
1KG non-Finnish EUR	8,216,754	97.09	79.76	0.001	299.99

Category specific LD Scores for the full baseline model and for cell-type-specific and cell-type-group-specific models

For the estimation of category specific LD Scores, I used readily available annotations from https://data.broadinstitute.org/alkesgroup/LDSCORE/ (1000G Phase 1). I estimated category specific LD Scores only for the FINRISK panel. I had to exclude one of the cell type annotations (spleen) from the analysis, because the original annotation file was corrupted. Therefore, in this study I used 219 cell types for the cell-type-specific analysis instead of the 220 cell types that were used in [Finucane et al., 2015]. Estimation procedure is similar to univariate LD Scores, besides the required annotation files that inform in which categories each variant belongs. I used same variant QC thresholds as with the univariate LD Scores, and excluded multiallelic variants, indels and variants with MAF < 0.01 or INFO < 0.9, and I estimated the category specific LD Scores by LDSC v.1.0.0 with a 1 cM window around an index variant.

4.3 SNP-heritability by univariate LD Score regression

I ran the univariate LD Score regression analysis for the lipid levels using the FINRISK LD reference panel, and LD Score regression plots of the results are presented in Figure 4.3. As results, I got following SNP-heritability estimates: HDL-C 0.074 (s.e. 0.039), LDL-C 0.193 (s.e. 0.048), total cholesterol 0.176 (s.e. 0.047) and triglycerides 0.14 (s.e. 0.042). I compared the estimates to previously published estimates in LD Hub database [Zheng et al., 2017], that is a centralised database of summary-level GWAS results and a web interface for LD Score regression, where registered users can upload GWAS summary statistics and browse previous LDSC results. Estimates are presented in Table 4.3. SNP-

Table 4.3: SNP-heritability estimates for four lipids levels: estimates from the FINRISK Study by LDSC and previously reported LDSC estimates from LD Hub.

Trait	h_g^2 **(LDSC)**	h_g^2 **(LD Hub)**
HDL-C	0.07 (0.04)	0.16 (0.02)
LDL-C	0.19 (0.05)	0.13 (0.02)
TC	0.18 (0.05)	0.15 (0.02)
TG	0.14 (0.04)	0.15 (0.02)

heritability estimates for HDL-C and TG obtained from the FINRISK were lower than the corresponding LD Hub estimates, and in contrast, were higher for LDL-C and TC than the corresponding LD Hub estimates. However, the differences were not statistically significant.

Estimates for the LDSC intercept - which measures the confounding bias in the GWAS summary statistics - for HDL-C, LDL-C, TC and TG were, respectively: 1.062 (s.e. 0.014), 1.020 (s.e. 0.016), 1.033 (s.e. 0.016) and 1.027 (s.e. 0.013). The highest intercept was for HDL-C, which also had unexpectedly low SNP-heritability estimate. Furthermore, HDL-C had a rather high attenuation ratio (0.67) compared to other lipids. Attenuation ratio for LDL-C was 0.22, and for both TC and TG 0.33. High attenuation ratio indicates that there might be some model misspecification or other confounding inflation in the GWAS test statistics.

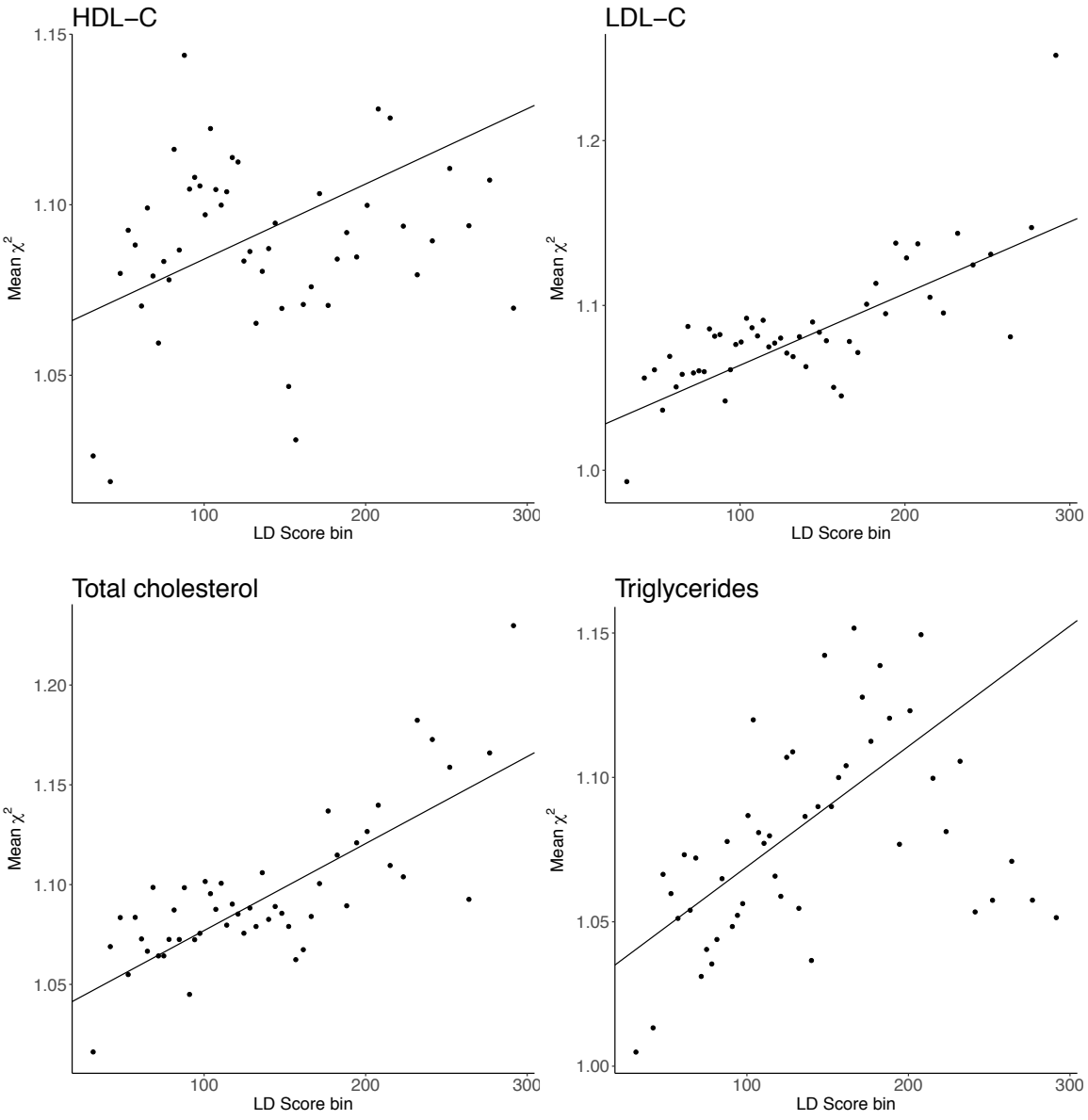


Figure 4.3: LD Score regression plots for four lipid levels, where each point represents an LD Score quantile. X-coordinate is the mean LD Score of variants in each quantile and y-coordinate the mean χ^2 statistic of variants in the corresponding quantile. The black line is the LD Score regression line fitted by an equation 3.5.

4.3.1 LD Score regression run by R

To verify that I had run LDSC correctly, I performed similar weighted linear regression analysis by R [R Development Core Team, 2008] for the four lipids. The R script that I used, is provided in the Appendix C. I performed an iteratively re-weighted least squares with a block jackknife in two steps, where the first step is used to obtain the intercept estimate, and the second step to obtain the heritability estimate. The two-step approach is used because linear regression performs poorly with outliers. The required data input are the LD Scores, GWAS summary association statistics as z-scores and the number of common variants (MAF > 0.05) used in the LD Score estimation.

At start, a crude heritability estimate for the initial weights is obtained from the formula: $\frac{M(\bar{\chi}^2-1)}{N\bar{l}_j}$, where $\bar{\chi}^2$ is the mean squared z-score and $\bar{l}_j$ is the mean LD Score. At the first step to estimate the LDSC intercept, all GWS SNPs - SNPs with a squared z-score over 30 - are removed. Weights are obtained by using a precision of four iterations, as is used in LDSC v.1.0.0, and both the LDSC intercept estimate and its standard error are obtained by a block jackknife linear regression using 200 blocks. At the second step, all SNPs are used and the block jackknife regression is performed with the intercept constrained to the step one intercept estimate. Heritability estimate is obtained by scaling the jackknife estimate by the GWAS sample size and by the number of common SNPs used in the LD Score estimation.

The results from the LD Score regression by R were consistent with the estimates from the LDSC v1.0.0 and are shown in Table 4.4.

Table 4.4: LD Score regression intercept and SNP-heritability estimates for four lipids from LDSC v.1.0.0 and from LD Score regression by R.

Trait	h_g^2 **(s.e.) (LDSC v1.0.0)**	h_g^2 **(s.e.) (R)**	**Intercept (s.e.) (LDSC v1.0.0)**	**Intercept (s.e.) (R)**
HDL-C	0.074 (0.039)	0.074 (0.033)	1.062 (0.014)	1.062 (0.014)
LDL-C	0.193 (0.048)	0.193 (0.054)	1.020 (0.016)	1.020 (0.016)
TC	0.176 (0.047)	0.176 (0.048)	1.033 (0.016)	1.033 (0.016)
TG	0.140 (0.042)	0.140 (0.040)	1.027 (0.013)	1.027 (0.013)

4.3.2 Effect of LD reference panel on SNP-heritability estimation

LDSC uses LD reference panel for the estimation of SNP-heritability. However, the reference panel population has to match to the target population of the GWAS sample. If there is a mismatch between the reference and target population, LDSC can lead to biased estimates. For example, if the reference population had equal LD Scores with the target population but included additional mean-zero noise, then the intercept estimate would be biased upward and the regression slope would be biased downward leading to underestimation of SNP-heritability [Bulik-Sullivan et al., 2015]. Also, if there was a directional bias in the average LD Score, then the intercept estimate would be biased either upward or downward depending on the direct of the bias [Bulik-Sullivan et al., 2015].

The Finnish population is a well-known genetic isolate within Europe and has gone through several genetic bottlenecks in recent history that had for example led to increased LD compared to other Europeans [Service et al., 2006], [Exome Aggregation Consortium, 2016]. Therefore, LD estimates obtained from other Europeans might lead biased estimation by LDSC methods when applied to the Finnish population. To evaluate the effect of LD reference panel on SNP-heritability estimation, I used four different LD reference panels: FINRISK panel from the underlying GWAS data representing an optimal LD, and three external panels from 1KG, of which one was a Finnish-specific panel (1KG FIN), and two were multi-ethnic European panels (1KG all EUR and 1KG non-Finnish EUR) that could possibly represent a mismatching LD.

LDSC was originally developed to distinguish confounding bias and polygenic effects from the GWAS summary statistics inflation [Bulik-Sullivan et al., 2015]. The intercept term measures the confounding inflation and should be close to one. Figure 4.4 shows a forest plot of the intercept estimate results from LDSC for the four lipid levels in different LD reference panels. Variation between the four panels is quite modest, but as could have been expected, intercept estimates from both multi-ethnic European panels are systematically a bit higher than from the two Finnish-specific panels. A downward directional bias in the mean LD Score in the reference panel in relation to the GWAS sample can lead to upward bias of the intercept estimate, as can be observed here. However, the differences are not statistically significant. The external 1KG FIN panel produced the lowest intercept estimated in all four traits. The high intercept value of HDL-C - indicating inflation due to some other factors than additive genetic effects - may be suggesting that there might be some model misspecification.

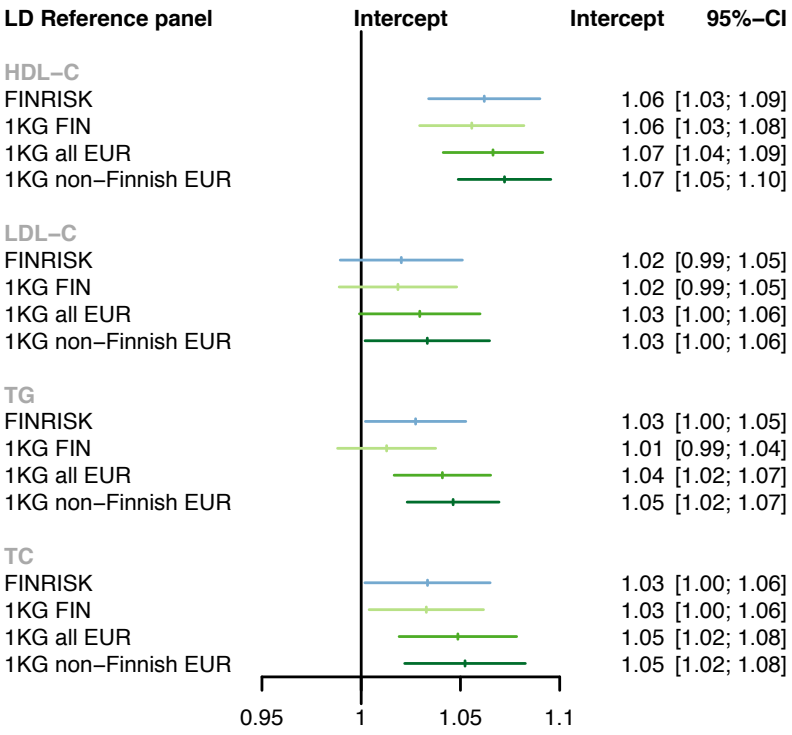


Figure 4.4: LDSC intercept estimates with 95%-confidence intervals for four lipid traits by different LD reference panels. LDSC intercept measures confounding inflation in the GWAS summary statistics.

LDSC attenuation ratio - a ratio between the intercept and mean χ^2 statistics which aims to estimate the relative balance of confounding and genetic effects - for all four lipid traits are shown in Figure 4.5. The attenuation ratio should be close to zero, which would mean that there is no inflation in the intercept term and inflation in χ^2 statistics is attributable to polygenic signal. Also, LDSC attenuation ratios are consistently higher in both multi-ethnic panels compared to the two Finnish specific panels for all four lipid levels.

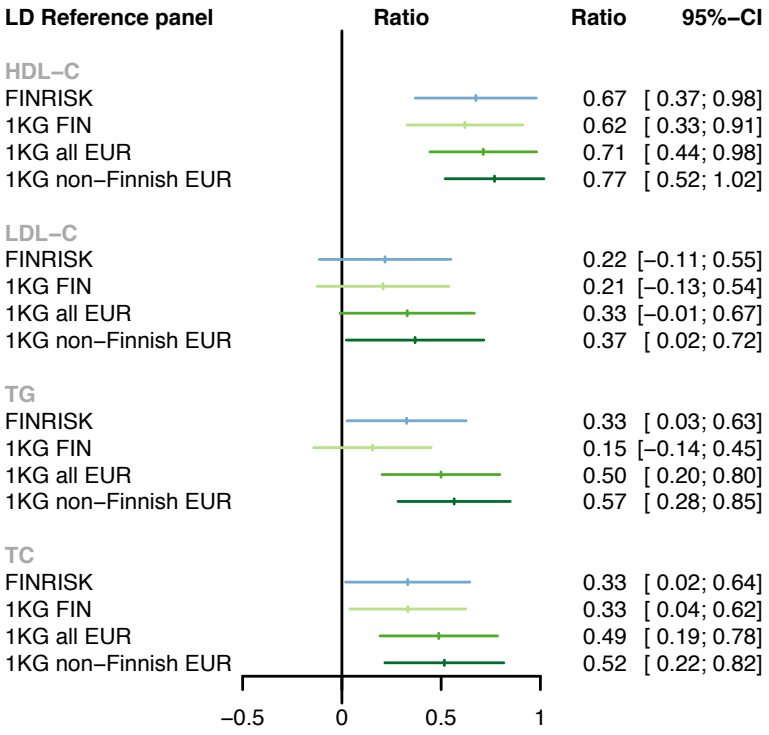


Figure 4.5: LDSC attenuation ratios for four lipid traits by different LD reference panels. LDSC attenuation ratio aims to estimate the relative balance of confounding and genetic effects.

SNP-heritability estimates with the 95% confidence intervals obtained by LDSC for all four lipid traits are presented in Figure 4.6, and are consistent between different LD

reference panels. The highest estimates were produced by an external 1KG FIN panel in all four lipid traits, but the variation between the panels is small. For TG, 1KG FIN panel point estimate is much higher compared to the estimates of other three reference panels, but the difference is not statistically significant.

In conclusion, LDSC was robust to the choice of LD reference panel when applied to the Finnish population, and SNP-heritability estimates were consistent between different panels regardless of the LD mismatch. Variation in the estimates between different LD reference panels was small. The highest heritability point estimates and the lowest LDSC intercept point estimates were produced by the Finnish specific panels, even though the differences were not statistically significant.

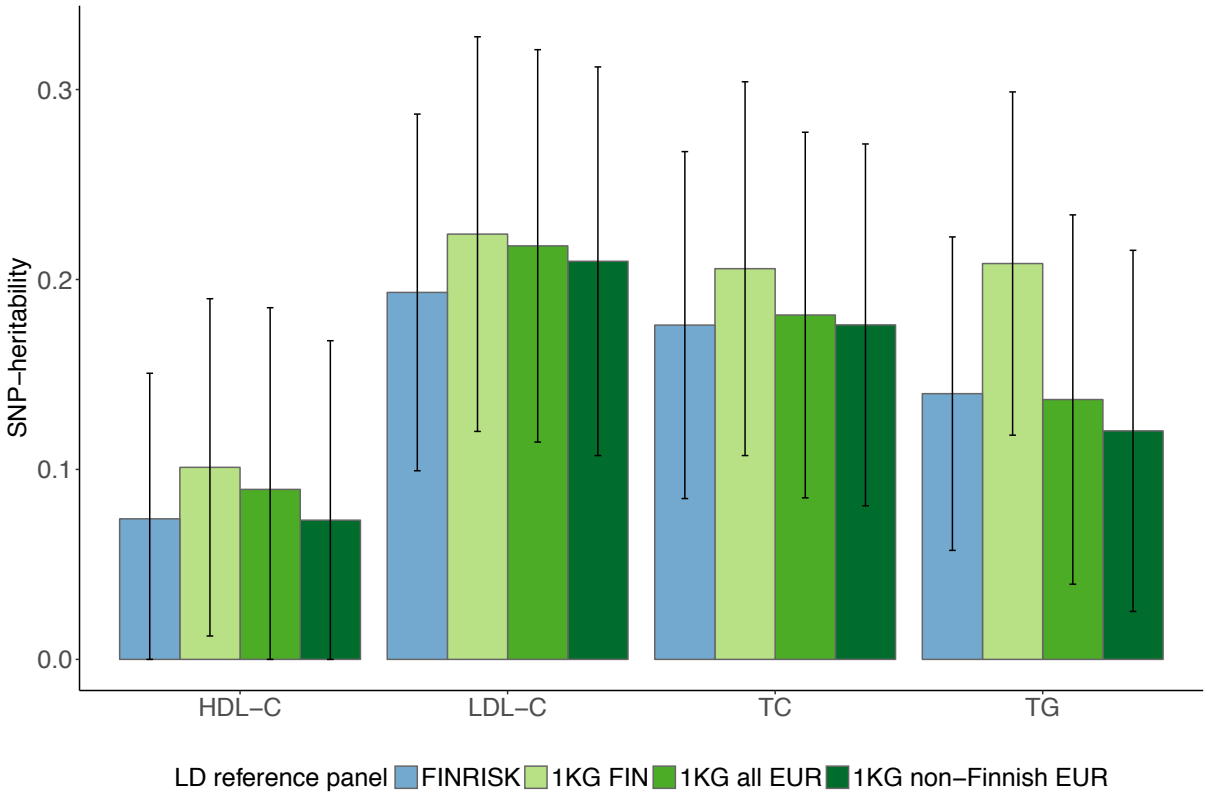


Figure 4.6: SNP-heritability estimates with 95%-confidence intervals for four lipid levels estimated by LDSC using four different LD reference panels.

4.4 Functional enrichment analysis using stratified LD Score regression

4.4.1 Analysis with the full baseline model

I applied S-LDSC with the full baseline model to all four lipid traits, and enrichment results for the 24 main independent functional annotations for HDL-C and LDL-C are presented in Figure 4.7, and for TC and TG in Figure 4.8. The enrichment p-value tests whether the category is enriched for heritability by testing whether the per-SNP heritability is greater in the category than out of the category [Finucane et al., 2015].

As a result, for HDL-C I observed statistically significant enrichment both at false discovery rate (FDR) < 0.05 and at p-value < 0.05 after Bonferroni correction for 24 hypotheses tested for three functional categories: for both of the acetylation of histone 3 at lycine 7 (for H3K27ac(Hnisz) and for H3K27ac(PGC2)) and for super enhancer. The estimates for the enrichment for H3K27ac(Hnisz) was 4.1× (s.e. 0.9) (p=1.8×10^{-5}), for H3K27ac(PGC2) was 6.7× (s.e. 2.1) (p=1.4×10^{-3}) and for super enhancer was 6.1× (s.e.1.5) (p=5.0×10^{-7}). Results are consistent with previously published results in [Finucane et al., 2015] that reported statistically significant enrichment after Bonferroni correction for 24 hypotheses tested also for H3K27ac(PGC2) and super enhancers. In addition, there were statistically significant enrichment for H3K4me1 and for H3K9ac that were not replicated in this study.

For LDL-C, I observed statistically significant enrichment for only one functional category at FDR < 0.05: the estimate for enrichment for acetylation of histone 3 at lycine 9 (H3K9ac) was 10.9× (s.e. 3.1) (p=0.001) which remained significant also after Bonferroni correction for 24 hypotheses tested. Besides H3K9ac, Finucane et al. observed enriched heritability for categories H3K27ac(PGC2), H3K4me1, super enhancer and repressed for LDL-C.

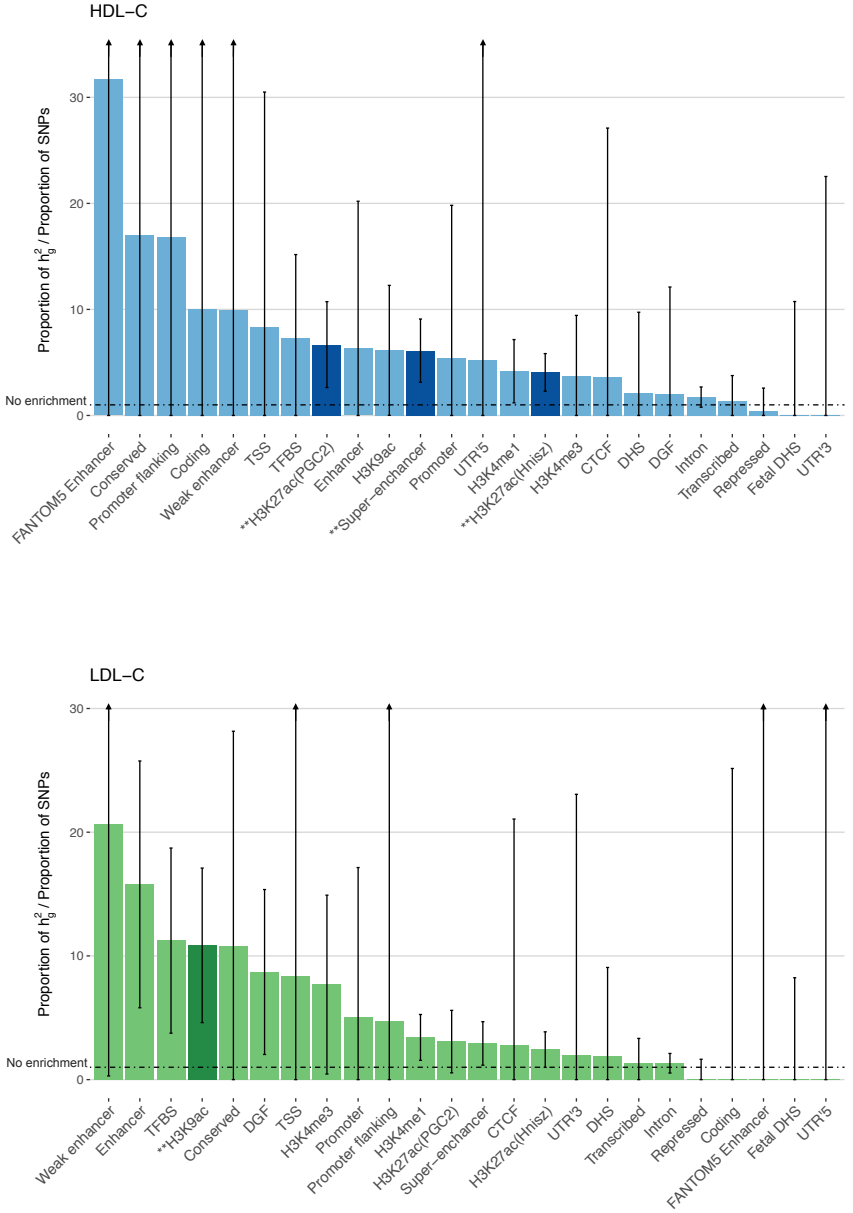


Figure 4.7: Enrichment estimates for the 24 main annotations for HDL-C and LDL-C with 95%-confidence intervals. The black dash-dotted line at 1 denotes no enrichment, and values above 1 mean enriched heritability for a given category and values below 1 depleted heritability for given category. Two asterisks and the darker shade of the bar colouring indicates significance at $P < 0.05$ after Bonferroni correction for the 24 hypotheses tested.

For TC, I observed statistically significant enrichment for six functional categories at FDR < 0.05: the estimates for the enrichment for enhancer was 15.1× (s.e. 4.7) (p=0.005), for H3K27ac(PGC2) was 4.5× (s.e. 1.3) (p=0.008), for H3K4me1 was 3.3× (s.e. 0.9) (p=0.01), for H2K9ac was 11.7× (s.e. 3.2) ($p=3.4 \times 10^{-4}$), for super enhancer was 3.5× (s.e. 0.9) (p=0.001) and for transcription factor binding site (TFBS) was 11.2× (s.e. 3.5) (p=0.005). After Bonferroni correction over 24 hypotheses tested, two of the categories remained statistically significant: the enrichment of H3K9ac and super enhancer. There were no results for TC reported in [Finucane et al., 2015].

For TG, I observed statistically significant enrichment for seven functional categories at FDR < 0.05: the estimates for enrichment for conserved was 23.9× (s.e. 10.2) (p=0.007), for H3K9ac was 7.9× (s.e. 3.0) (p=0.005), for super enhancer was 4.0× (s.e. 1.0) (p=0.008), for H3K27ac(Hnisz) was 3.0× (s.e. 0.7) (p=0.002), for H3K27ac(PGC2) was 8.4× (s.e. 2.2) (p=0.003), for promoter was 17.1× (s.e. 6.5) (p=0.01) and for repressed was $-1.5\times$ (s.e. 1.1) (p=0.01). However, after Bonferroni correction none of the enrichment reached statistical significance. Categories H3K9ac, H3K27ac(PGC2) and super enhancer were also observed to be enriched in Finucane et al. They reported also histone mark H3K4me3 to be enriched that was not significant in this study.

Full baseline model results including all 53 categories are provided in the supplementary materials: results for HDL-C are presented in Table S2, for LDL-C in Table S3, for TC in Table S4 and for TG in Table S5.

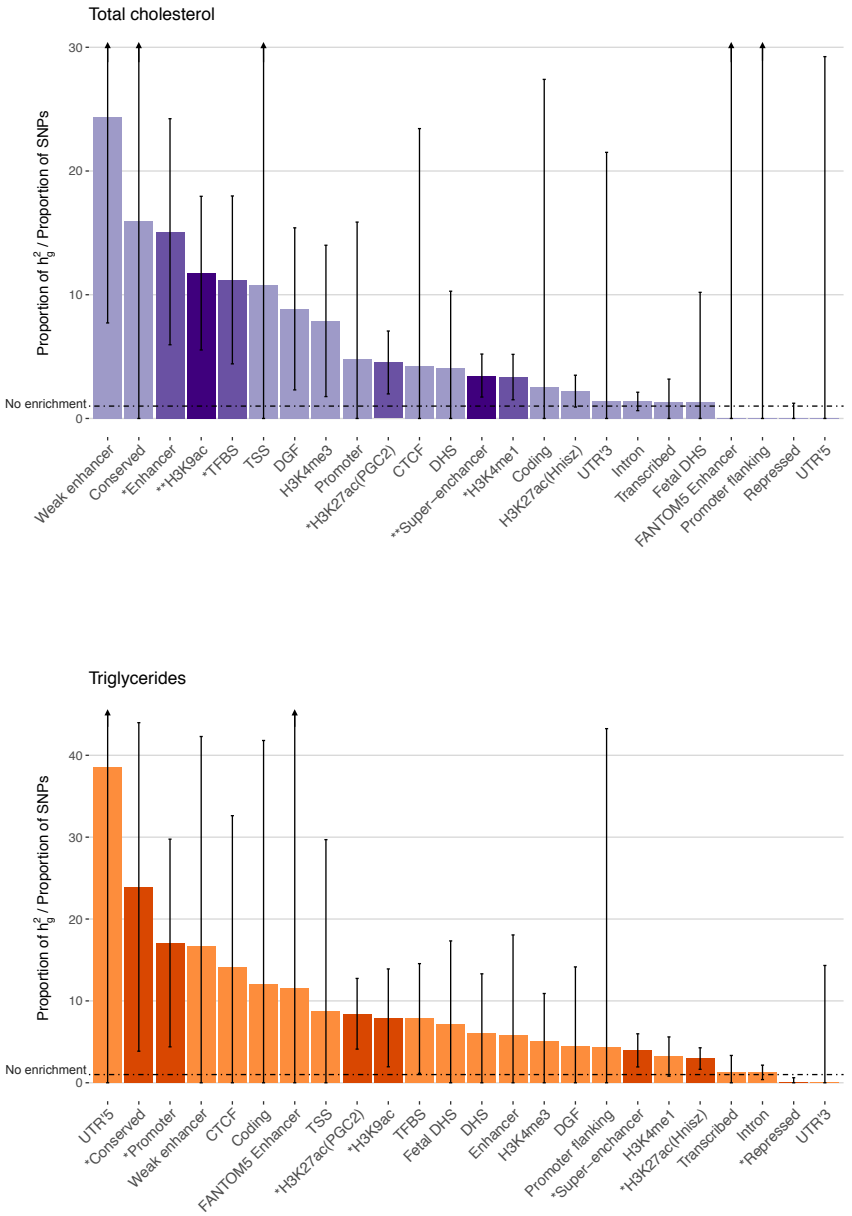


Figure 4.8: Enrichment estimates for the 24 main annotations for TC and TG with 95%-confidence intervals. The black dash-dotted line at 1 denotes no enrichment, and values above 1 mean enriched heritability for a given category and values below 1 depleted heritability for given category. One asterisk and middle shade of the bar colouring indicate significance at FDR < 0.05 and two asterisk and the darkest shade of the bar colouring indicates significance at P < 0.05 after Bonferroni correction for the 24 hypotheses tested.

4.4.2 Enrichment of specific cell types and specific cell type groups

Enrichment of cell type groups for all four lipid levels are presented in Figure 4.9 and in supplementary materials: HDL-C in Table S6, LDL-C in Table S7, TC in Table S8 and TG in Table S9. The results of the cell-type-specific and cell-type-group-specific analyses are ranked by the p-value of the coefficient τ_C instead of the enrichment p-value. P-value of τ_C tests whether the annotation contributes significantly to per-SNP heritability after controlling for the effects of the annotations in the full baseline model [Finucane et al., 2015].

Top cell type group for HDL-C, TC and TG was liver and for LDL-C adrenal or pancreas. However, only for HDL-C and TC the top cell type group passed the significance threshold at FDR < 0.05: the coefficient (τ_c) p-value for HDL-C was 3.5×10^{-4} and for TC 1.4×10^{-3}. After Bonferroni correction for 10 x 4 = 40 hypotheses tested, only the enrichment of liver cell type group for HDL-C remained significant. There were no other enriched cell type group that passed either of the significance thresholds. Results are in line with the top cell types reported in [Finucane et al., 2015]. In addition, Finucane et al. reported enrichment for liver for LDL-C, adrenal or pancreas and other for HDL-C, and immune or hematopoietic for TG that were not replicated in this study.

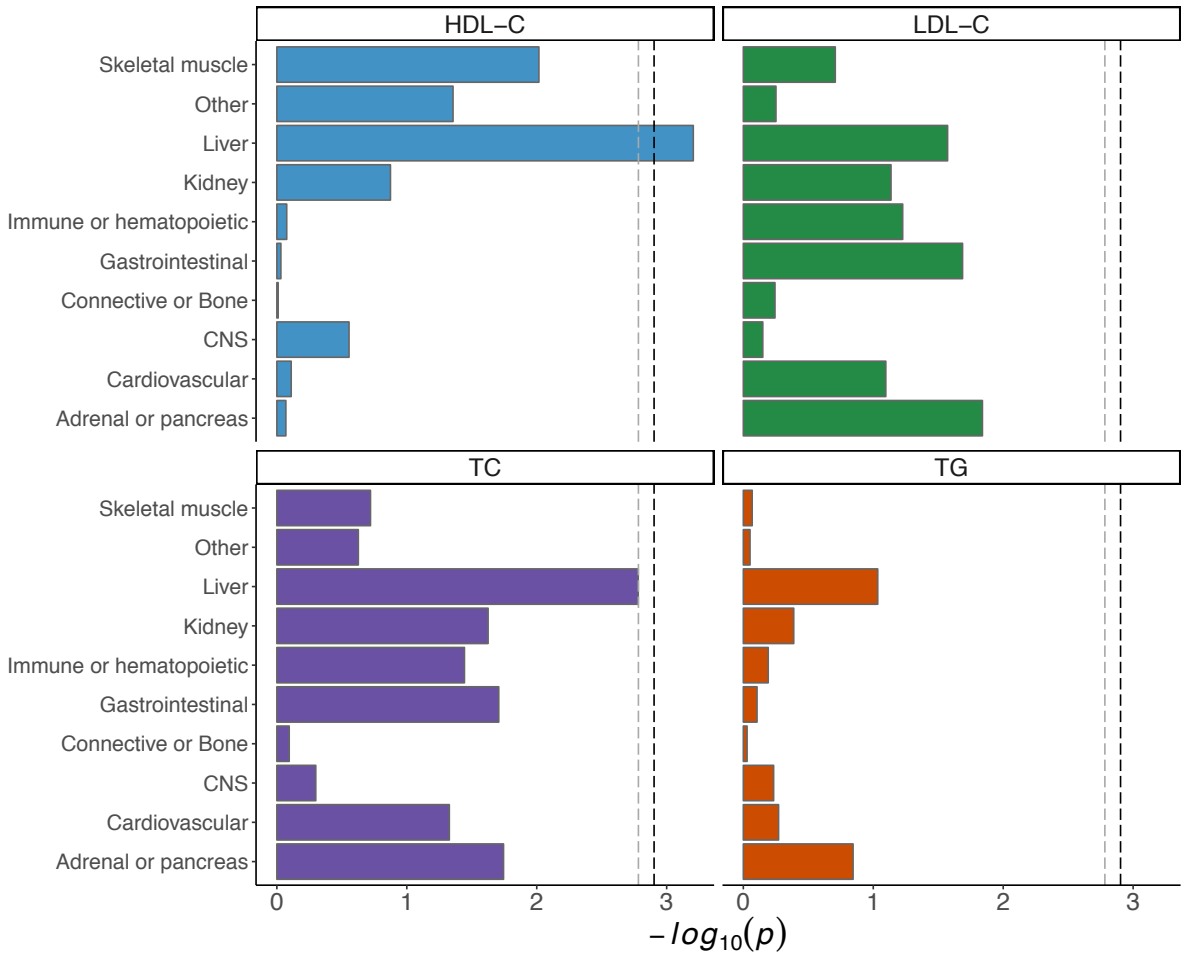


Figure 4.9: Enrichment of ten cell type groups for the lipid levels. The black dashed line at $-log_{10}(\mathrm{P}) = 2.9$ is the cutoff for Bonferroni significance and the grey dashed line at $-log_{10}(\mathrm{P}) = 2.78$ is the cutoff for FDR < 0.05.

Top cell type for HDL-C and TC was liver, for LDL-C fetal adrenal and for TG esophagus, and are presented in Table 4.5. However, only the liver cell type for TC passed significance threshold at the FDR < 0.05. After Bonferroni correction for 4x219=876 hypotheses tested, there were no statistically significant enrichment in any of the cell types. Results of all cell types for all traits are presented in Supplementary Table S10. Finucane et al. reported liver as a top cell type for all HDL-C, LDL-C and TG. In addition, they had observed statistically significant enrichment for the following cell types at FDR < 0.05 that were not replicated here: cell types adipose nuclei and CD14 primary for HDL-C; cell types fetal adrenal, CD14 primary and adipose nuclei for LDL-C; and cell types adipose nuclei, duodenum mucosa, colonic mucosa and rectal mucosa for TG.

Table 4.5: Enrichment of top cell types in lipid levels, $-log_{10}(P)$ of τ_c. An asterisk indicates significance at FDR < 0.05

Phenotype	**Cell type**	**Cell type group**	**Mark**	$-log_{10}(P)$
HDL-C	Liver	Liver	H3K27ac	2.84
LDL-C	Fetal adrenal	Adrenal or pancreas	H3K4me1	3.50
TC	Liver *	Liver	H3K4me1	3.86
TG	Esophagus	Gastrointestinal	H3K4me1	2.09

Chapter 5

Discussion

The first aim of this thesis was to explain the statistics behind LD Score regression that was done in the method section. The second aim was to evaluate how the choice of LD reference panel affects the SNP-heritability estimation of lipid traits by LDSC when applied to the Finnish population. The evaluation showed that the method is not as sensitive to the choice of the LD reference panel as would have been expected, at least with a quantitative trait with low or moderate heritability. The SNP-heritability estimates were consistent across all four panels, even with the non-Finnish multi-ethnic LD reference panels. The Finnish specific panels produced the highest SNP-heritability estimates and the lowest LDSC intercept point estimates. However, a caveat of the comparison was that the in-sample LD reference panel was imputed data and the external 1KG panels were sequence data. To further evaluate the role of the LD reference panel, comparison between both the in-sample LD reference panel and the external LD reference panel being the sequence data should be performed. Also, the role of the reference panel size should be evaluated.

The SNP-heritability estimates were otherwise consistent with previously reported estimates in LD Hub, except an unexpectedly low point estimate of HDL-C (0.07 (s.e. 0.04) compared to 0.16 (s.e. 0.02) in LD Hub). In addition, HDL-C showed high LDSC intercept value (1.06 (s.e. 0.014)) and high attenuation ratio (0.67), indicating confounding bias in the GWAS summary statistics, possibly due to model misspecification. Because LDSC assumes high polygenicity, in principle the low estimate could suggest that the genetic architecture of HDL-C differ from the other three lipids. However, since LDSC estimates in LD Hub for HDL-C are higher, and also the heritability estimates obtained by other methods have been higher [Vattikuti et al., 2012], [Kathiresan et al., 2007], the

unexpectedly low estimate in HDL-C may be a property of the data set used in this study.

The third aim of the study was to examine whether some functional categories are enriched for heritability of the lipid traits in the Finnish population by applying stratified LD Score regression. As results using the full baseline model, I observed statistically significant enrichment for heritability at FDR < 0.05 in many histone marks across all four lipid levels: acetylation of histone 3 at lycine 9 (H3Kac9) was enriched for LDL-C, TC and TG; acetylation of histone 3 at lycine 27 (H3K27ac(PCG2) or H3K27ac(Hnisz)) was enriched for HDL-C, TC and TG; and monomethylation of histone 3 at lycine 4 (H3K4me1) was enriched for TC. Histone marks are often enriched at active enhancers and promoters, and therefore can be used to study gene expression regulation. Furthermore, I observed statistically significant enrichment for super enhancer for HDL-C, TC and TG. Super enhancers are clusters of highly active enhancers in close proximity that have unusually high levels of activator binding or histone modifications, and are often located near genes with cell-type specific functions [Hnisz et al., 2013]. In addition, I observed enrichment for the categories of conserved, promoter and transcribed for TG, and enhancer for TC. These results replicated previously reported [Finucane et al., 2015] enrichment for super enhancer for HDL-C, LDL-C and TG, as well as enrichment for H3K27ac for HDL-C and TG, and enrichment for H3Kac9 for LDL-C and TG. I observed three new enriched categories for TG: conserved, promoter and repressed.

The enrichment analysis of specific cell type groups showed liver as a top cell type group for HDL-C and TC, that were significant at FDR < 0.05. In addition, the enrichment analysis of specific cell types showed liver as a top cell type for TC. However, there was not enough power in this study to replicate other previously reported top cell types [Finucane et al., 2015]. One caveat of S-LDSC is that in order to reach the optimal performance, the trait should be polygenic and have very high heritability and/or large enough sample size. Otherwise S-LDSC produces high standard errors. Finucane et al. proposed a heritability z-score method for evaluating whether the trait would be amenable for S-LDSC, and based on the power analysis, recommended to use a threshold of heritability z-score > 7. Because the obtained estimates for the total SNP-heritability for all four lipid levels were low or moderate (7%-19%), the sample sizes should have been larger to achieve required statistical power. Heritability z-scores in this study for the HDL-C, LDL-C, TC and TG were, respectively: 1.89, 4.03, 3.78 and 3.32, that are all below the recommended threshold of 7.

The advantage of LDSC and S-LDSC is that they do not require individual-level data, instead they use summary statistics from GWAS and LD reference panel matching to the GWAS target population. However, there are some caveats also in summary-level

based methods. In some applications, there is a loss of accuracy compared to individual-level based methods, for example REML-based methods [Yang et al., 2011a], [Loh et al., 2015] have been able to produce more accurate SNP-heritability estimates compared to LDSC [Speed et al., 2017]. In contrast, when partitioning heritability by multiple functional categories, S-LDSC performs better than methods based on REML, which become computationally intractable when sample sizes increase to tens of thousands or when number of categories increases [Finucane et al., 2015]. S-LDSC can also include overlapping categories. However, categories included in the model have to be large enough and small categories lead to large standard errors [Finucane et al., 2015]. In addition, the model is unable to catch the enrichment for categories that are not included in the model, and is therefore limited by the available functional data, and bias due to model misspecification cannot be ruled out [Finucane et al., 2015]. Also, because the method relies on LD between SNPs, and assumes that the per-SNP heritability is uniformly distributed across the whole category, it can be reliably applied only to common variants. Rare variants are usually not well-tagged by common variants [Spencer et al., 2009] and it is not well known how rare variants behave with respect to per-SNP heritability.

As a conclusion, I found LDSC easy to run on existing GWAS data. Especially, since the method is based on linear regression and did not require any individual-level genetic data, it was extremely fast to run. In addition, new functional annotations can be easily added to S-LDSC that can improve the accuracy when new functional data become available. Further, LDSC was robust to the choice of the LD reference panel. Therefore it is a useful tool for future analyses of large GWAS summary statistics.

Acknowledgments

I would like to thank my supervisor Matti Pirinen for providing me with the opportunity to write my thesis at FIMM. I would like to thank Matti for all the guidance and support during the process of writing this thesis, and I am grateful for all the valuable comments that helped me to develop and improve this work. I would also like to thank Sirkka-Liisa Varvio for introducing me to Matti and for reviewing this thesis. I would like to thank all the participants of the National FINRISK Study cohorts and all the researchers of the National FINRISK study for the permission to use these cohorts in this thesis, without them this thesis would not have been possible. Especially, I would like to thank the whole Human Genomics team at FIMM for providing an inspiring and supportive work environment. Lastly, I would like to thank my family, Kimmo, Pirkko, Juha, Henna and Sari, and all my friends for all the encouragement and for always believing in me.

Bibliography

[Andersson et al., 2014] Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., Ntini, E., Arner, E., Valen, E., Li, K., Schwarzfischer, L., Glatz, D., Raithel, J., Lilje, B., Rapin, N., Bagger, F. O., Jorgensen, M., Andersen, P. R., Bertin, N., Rackham, O., Burroughs, A. M., Baillie, J. K., Ishizu, Y., Shimizu, Y., Furuhata, E., Maeda, S., Negishi, Y., Mungall, C. J., Meehan, T. F., Lassmann, T., Itoh, M., Kawaji, H., Kondo, N., Kawai, J., Lennartsson, A., Daub, C. O., Heutink, P., Hume, D. A., Jensen, T. H., Suzuki, H., Hayashizaki, Y., Müller, F., Forrest, A. R. R., Carninci, P., Rehli, M., Sandelin, A., and Clevers, H. (2014). An atlas of active enhancers across human cell types and tissues. *Nature*, 507(7493):455–61.

[Anttila et al., 2010] Anttila, V., Stefansson, H., Kallela, M., Todt, U., Terwindt, G. M., Calafato, M. S., Nyholt, D. R., Dimas, A. S., Freilinger, T., Müller-Myhsok, B., Artto, V., Inouye, M., Alakurtti, K., Kaunisto, M. A., Hämäläinen, E., de Vries, B., Stam, A. H., Weller, C. M., Heinze, A., Heinze-Kuhn, K., Goebel, I., Borck, G., Göbel, H., Steinberg, S., Wolf, C., Björnsson, A., Gudmundsson, G., Kirchmann, M., Hauge, A., Werge, T., Schoenen, J., Eriksson, J. G., Hagen, K., Stovner, L., Wichmann, H.-E., Meitinger, T., Alexander, M., Moebus, S., Schreiber, S., Aulchenko, Y. S., Breteler, M. M. B., Uitterlinden, A. G., Hofman, A., van Duijn, C. M., Tikka-Kleemola, P., Vepsäläinen, S., Lucae, S., Tozzi, F., Muglia, P., Barrett, J., Kaprio, J., Färkkilä, M., Peltonen, L., Stefansson, K., Zwart, J.-A., Ferrari, M. D., Olesen, J., Daly, M., Wessman, M., van den Maagdenberg, A. M. J. M., Dichgans, M., Kubisch, C., Dermitzakis, E. T., Frants, R. R., and Palotie, A. (2010). Genome-wide association study of migraine implicates a common susceptibility variant on 8q22.1. *Nature Genetics*, 42(10):869–873.

[Borodulin et al., 2015] Borodulin, K., Vartiainen, E., Peltonen, M., Jousilahti, P., Juolevi, A., Laatikainen, T., Männistö, S., Salomaa, V., Sundvall, J., and Puska, P. (2015). Forty-year trends in cardiovascular risk factors in Finland. *European Journal of Public Health*, 25(3):539–546.

[Boyle et al., 2011] Boyle, A. P., Song, L., Lee, B.-K., London, D., Keefe, D., Birney, E., Iyer, V. R., Crawford, G. E., and Furey, T. S. (2011). High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. *Genome Research*, 21(3):456–464.

[Bulger and Groudine, 2010] Bulger, M. and Groudine, M. (2010). Enhancers: The abundance and function of regulatory sequences beyond promoters. *Developmental Biology*, 339(2):250–257.

[Bulik-Sullivan et al., 2015] Bulik-Sullivan, B. K., Loh, P.-R., Finucane, H. K., Ripke, S., Yang, J., Patterson, N., Daly, M. J., Price, A. L., and Neale, B. M. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics*, 47(3):291–295.

[Collins et al., 1997] Collins, F. S., Guyer, M. S., and Chakravarti, A. (1997). Variations on a theme: Cataloging human DNA sequence variation. *Science*, 278(5343):1580–1581.

[Devlin and Roeder, 1999] Devlin, B. and Roeder, K. (1999). Genomic Control for Association Studies. *Biometrics*, 55(4):997–1004.

[Exome Aggregation Consortium, 2016] Exome Aggregation Consortium (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616):285–4291.

[Finucane et al., 2015] Finucane, H. K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R., Anttila, V., Xu, H., Zang, C., Farh, K., Ripke, S., Day, F. R., Purcell, S., Stahl, E., Lindstrom, S., Perry, J. R., Okada, Y., Raychaudhuri, S., Daly, M. J., Patterson, N., Neale, B. M., and Price, A. L. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics*, 47(11):1228–1235.

[Gormley et al., 2016] Gormley, P., Anttila, V., Winsvold, B. S., Palta, P., Esko, T., Pers, T. H., Farh, K.-H., Cuenca-Leon, E., Muona, M., Furlotte, N. A., Kurth, T., Ingason, A., McMahon, G., Ligthart, L., Terwindt, G. M., Kallela, M., Freilinger, T. M., Ran, C., Gordon, S. G., Stam, A. H., Steinberg, S., Borck, G., Koiranen, M., Quaye, L., Adams, H. H. H., Lehtimäki, T., Sarin, A.-P., Wedenoja, J., Hinds, D. A., Buring, J. E., Schurks, M., Ridker, P. M., Hrafnsdottir, M. G., Stefansson, H., Ring, S. M., Hottenga, J.-J., Penninx, B. W. J. H., Färkkila, M., Artto, V., Kaunisto, M., Vepsäläinen, S., Malik, R., Heath, A. C., Madden, P. A. F., Martin, N. G., Montgomery, G. W., Kurki, M. I., Kals, M., Magi, R., Parn, K., Hämäläinen, E., Huang, H., Byrnes, A. E., Franke, L., Huang, J., Stergiakouli, E., Lee, P. H., Sandor, C., Webber, C., Cader, Z., Müller-Myhsok, B., Schreiber, S., Meitinger, T., Eriksson, J. G., Salomaa, V., Heikkilä, K.,

Loehrer, E., Uitterlinden, A. G., Hofman, A., van Duijn, C. M., Cherkas, L., Pedersen, L. M., Stubhaug, A., Nielsen, C. S., Männikkö, M., Mihailov, E., Milani, L., Gobel, H., Esserlind, A.-L., Christensen, A. F., Hansen, T. F., Werge, T., Kaprio, J., Aromaa, A. J., Raitakari, O., Ikram, M. A., Spector, T., Järvelin, M.-R., Metspalu, A., Kubisch, C., Strachan, D. P., Ferrari, M. D., Belin, A. C., Dichgans, M., Wessman, M., van den Maagdenberg, A. M. J. M., Zwart, J.-A., Boomsma, D. I., Smith, G. D., Stefansson, K., Eriksson, N., Daly, M. J., Neale, B. M., Olesen, J., Chasman, D. I., Nyholt, D. R., and Palotie, A. (2016). Meta-analysis of 375,000 individuals identifies 38 susceptibility loci for migraine. *Nature Genetics*, 48(8):856–866.

[Gusev et al., 2013] Gusev, A., Bhatia, G., Zaitlen, N., Vilhjálmsson, B. J., Diogo, D., Stahl, E. A., Gregersen, P. K., Worthington, J., Klareskog, L., Raychaudhuri, S., Plenge, R. M., Pasaniuc, B., and Price, A. L. (2013). Quantifying missing heritability at known GWAS loci. *PLoS Genetics*, 9(12):e1003993.

[Gusev et al., 2014] Gusev, A., Lee, S. H., Trynka, G., Finucane, H., Vilhjálmsson, B. J., Xu, H., Zang, C., Ripke, S., Bulik-Sullivan, B., Stahl, E., Kähler, A. K., Hultman, C. M., Purcell, S. M., McCarroll, S. A., Daly, M., Pasaniuc, B., Sullivan, P. F., Neale, B. M., Wray, N. R., Raychaudhuri, S., and Price, A. L. (2014). Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *American Journal of Human Genetics*, 95(5):535–552.

[Hnisz et al., 2013] Hnisz, D., Abraham, B. J., Lee, T. I., Ashley Lau, V. S.-A., Sigova, A. A., Hoke, H., and Young, R. A. (2013). Transcriptional super-enhancers connected to cell identity and disease. *Cell*, 155(4):934–947.

[Hoffman et al., 2013] Hoffman, M. M., Ernst, J., Wilder, S. P., Kundaje, A., Harris, R. S., Libbrecht, M., Giardine, B., Ellenbogen, P. M., Bilmes, J. A., Birney, E., Hardison, R. C., Dunham, I., Kellis, M., and Noble, W. S. (2013). Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Research*, 41(2):827–841.

[Kapranov, 2009] Kapranov, P. (2009). From transcription start site to cell biology. *Genome Biology*, 10(4):217.

[Kathiresan et al., 2007] Kathiresan, S., Manning, A. K., Demissie, S., D'Agostino, R. B., Surti, A., Guiducci, C., Gianniny, L., Burtt, N. P., Melander, O., Orho-Melander, M., Arnett, D. K., Peloso, G. M., Ordovas, J. M., and Cupples, L. A. (2007). A genome-wide association study for blood lipid phenotypes in the Framingham Heart Study. *BMC Medical Genetics*, 8(Suppl 1):S17.

[Kent et al., 2002] Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., Haussler, and David (2002). The Human Genome Browser at UCSC. *Genome Research*, 12(6):996–1006.

[Klug et al., 2012] Klug, W., Cummings, M., Palladino, M., and Spencer, C. (2012). *Concepts of Genetics*. Pearson Education, 10th edition.

[Lindblad-Toh et al., 2011] Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M. F., Parker, B. J., Washietl, S., Kheradpour, P., Ernst, J., Jordan, G., Mauceli, E., Ward, L. D., Lowe, C. B., Holloway, A. K., Clamp, M., Gnerre, S., Alföldi, J., Beal, K., Chang, J., Clawson, H., Cuff, J., Palma, F. D., Fitzgerald, S., Flicek, P., Guttman, M., Hubisz, M. J., Jaffe, D. B., Jungreis, I., Kent, W. J., Kostka, D., Lara, M., Martins, A. L., Massingham, T., Moltke, I., Raney, B. J., Rasmussen, M. D., Robinson, J., Stark, A., Vilella, A. J., Wen, J., Xie, X., Zody, M. C., Worley, K. C., Kovar, C. L., Muzny, D. M., Gibbs, R. A., Warren, W. C., Mardis, E. R., Weinstock, G. M., Wilson, R. K., Birney, E., Margulies, E. H., Herrero, J., Green, E. D., Haussler, D., Siepel, A., Goldman, N., Pollard, K. S., Pedersen, J. S., Lander, E. S., and Kellis, M. (2011). A high-resolution map of human evolutionary constraint using 29 mammals. *Nature*, 478(7370):476–482.

[Loh et al., 2015] Loh, P.-R., Bhatia, G., Gusev, A., Finucane, H. K., Bulik-Sullivan, B. K., Pollack, S. J., de Candia, T. R., Lee, S. H., Wray, N. R., Kendler, K. S., O'Donovan, M. C., Neale, B. M., Patterson, N., and Price, A. L. (2015). Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nature Genetics*, 47(12):1385–1392.

[Luco et al., 2010] Luco, R. F., Pan, Q., Tominaga, K., Blencowe, B. J., Pereira-Smith, O. M., and Misteli, T. (2010). Regulation of alternative splicing by histone modifications. *Science*, 327(5968):996–1000.

[Marchini and Howie, 2010] Marchini, J. and Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, 11(7):499–511.

[Marchini et al., 2007] Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics*, 39(7):906–913.

[Martin et al., 2011] Martin, D., Pantoja, C., Miñán, A. F., Valdes-Guezada, C., Moltó, E., Matesanz, F., Bogdanovic, O., Calle-Mustienes, E. D. L., Domíŋnguez, O., Taher, L., Furlan-Magaril, M., Alcina, A., Cañón, S., Fedetz, M., Blasco, M. A., Pereira, P. S., Ovcharenko, I., Recillas-Targa, F., Montoliu, L., Manzanares, M., Guigó, R., Serrano, M., Casares, F., and Gómez-Skarmeta, J. L. (2011). Genome-wide CTCF distribution

in vertebrates defines equivalent sites that aid the identification of disease-associated genes. *Nature Structural And Molecular Biology*, 18(9):708–714.

[Maurano et al., 2012] Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., Reynolds, A. P., Sandstrom, R., Qu, H., Brody, J., Shafer, A., Neri, F., Lee, K., Kutyavin, T., Stehling-Sun, S., and Johnson, A. K. (2012). Systematic localization of common disease-associated variation in regulatory DNA. *Science*, 337(6099):1190–1195.

[McVicker et al., 2013] McVicker, G., van de Geijn, B., Degner, J. F., Cain, C. E., Banovich, N. E., Raj, A., Lewellen, N., Myrthil, M., Gilad, Y., and Pritchard, J. K. (2013). Identification of genetic variants that affect histone modifications in human cells. *Science*, 342(6159):747–749.

[Mignone et al., 2002] Mignone, F., Gissi, C., Liuni, S., and Pesole, G. (2002). Untranslated regions of mRNAs. *Genome Biology*, 3(3):reviews0004.1.

[Miller and Kumar, 2001] Miller, M. P. and Kumar, S. (2001). Understanding human disease mutations through the use of interspecific genetic variation. *Human Molecular Genetics*, 10(21):2319–2328.

[Pearson and Manolio, 2008] Pearson, T. A. and Manolio, T. A. (2008). How to interpret a genome-wide association study. *JAMA*, 299(11):1335–1344.

[Pe'er et al., 2008] Pe'er, I., Yelensky, R., Altshuler, D., and Daly, M. J. (2008). Estimation of the multiple testing burden for genome-wide association studies of nearly all common variants. *Genetic Epidemiology*, 32(4):381–385.

[Peltonen et al., 1999] Peltonen, L., Jalanko, A., and Varilo, T. (1999). Molecular genetics of the Finnish disease heritage. *Human Molecular Genetics*, 8(10):1913–1923.

[Pott and Lieb, 2015] Pott, S. and Lieb, J. D. (2015). What are super-enhancers? *Nature Genetics*, 47(1):8–12.

[Price et al., 2008] Price, A. L., Weale, M. E., Patterson, N., Myers, S. R., Need, A. C., Shianna, K. V., Ge, D., Rotter, J. I., Torres, E., Taylor, K., Goldstein, D. B., and Reich, D. (2008). Long-range LD can confound genome scans in admixed populations. *The American Journal of Human Genetics*, 83(1):132 – 135.

[R Development Core Team, 2008] R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

[Rana et al., 2010] Rana, J. S., Visser, M. E., Arsenault, B. J., Després, J.-P., Stroes, E. S. G., Kastelein, J. J. P., Wareham, N. J., Boekholdt, S. M., and Khaw, K.-T. (2010). Metabolic dyslipidemia and risk of future coronary heart disease in apparently healthy men and women: The EPIC-Norfolk prospective population study. *International Journal of Cardiology*, 143(3):399–404.

[Roadmap Epigenomics Consortium, 2015] Roadmap Epigenomics Consortium (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, 7539(518):317–330.

[Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014] Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, 511(7510):421–427.

[Seber et al., 2003] Seber, G., Cummings, M., and Alan, J. (2003). *Linear Regression Analysis*. Wiley series in probability and statistics, 2nd edition.

[Service et al., 2006] Service, S., DeYoung, J., Galver, L., Peltonen, L., Monne, M., Varilo, T., Peddle, L., van Duijn, C., Ospina, J., Palha, J. A., Jarvelin, M.-R., Aulchenko, Y., Rahman, P., Roos, J. L., Bedoya, G., Heutink, P., , Macedo, A., Pretorious, H., Ruiz-Linares, A., Sabatti, C., Karayiorgou, M., Oostra, B., Murray, S., Collins, A., and Freimer, N. (2006). Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association studies. *Nature Genetics*, 38(5):556–560.

[Shao and Wu, 1989] Shao, J. and Wu, C. (1989). A general theory for jackknife variance estimation. *The Annals of Statistics*, 17(3):1176–1197.

[Slatkin, 2008] Slatkin, M. (2008). Linkage disequilibrium - understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*, 9(6):477–485.

[Speed et al., 2017] Speed, D., Cai, N., Johnson, M. R., Nejentsev, S., and Balding, D. J. (2017). Re-evaluation of SNP heritability in complex human traits. *Nature Genetics*, 49(7):986–992.

[Spencer et al., 2009] Spencer, C. C. A., Su, Z., Donnelly, P., and Marchini, J. (2009). Designing genome-wide association studies: Sample size, power, imputation, and the choice of genotyping chip. *PLOS Genetics*, 5(5):1–13.

[Stamm, 2008] Stamm, S. (2008). Regulation of alternative splicing by reversible protein phosphorylation. *Journal of Biological Chemistry*, 283(3):1223–1227.

[The 1000 Genomes Project Consortium, 2015] The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature*, 526(7571):68–74.

[The ENCODE Project Consortium, 2007] The ENCODE Project Consortium (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447(7146):799–816.

[The ENCODE Project Consortium, 2012] The ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74.

[The International HapMap Consortium, 2005] The International HapMap Consortium (2005). A haplotype map of the human genome. *Nature*, 437(7063):1299–1320.

[The International Schizophrenia Consortium, 2009] The International Schizophrenia Consortium (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, 460(7256):748–752.

[Trynka et al., 2013] Trynka, G., Sandor, C., Han, B., Xu, H., Stranger, B. E., Liu, X. S., and Raychaudhuri, S. (2013). Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nature Genetics*, 45(2):124–130.

[Vattikuti et al., 2012] Vattikuti, S., Guo, J., and Chow, C. C. (2012). Heritability and genetic correlations explained by common SNPs for metabolic syndrome traits. *PLOS Genetics*, 8(3):1–8.

[Visscher et al., 2008] Visscher, P. M., Hill, W. G., and Wray, N. R. (2008). Heritability in the genomics era - concepts and misconceptions. *Nature Reviews Genetics*, 9(4):255–266.

[Vukcevic et al., 2011] Vukcevic, D., Hechter, E., Spencer, C., and Donnelly, P. (2011). Disease model distortion in association studies. *Genetic Epidemiology*, 35(4):278–290.

[Ward and Kellis, 2012] Ward, L. D. and Kellis, M. (2012). Evidence of abundant purifying selection in humans for recently acquired regulatory functions. *Science*, 337(6102):1675–1678.

[Wright, 1949] Wright, S. (1949). The genetical structure of populations. *Annals of Eugenics*, 15(1):323–354.

[Yang et al., 2010] Yang, J., Heath, A. C., Benyamin, B., Nyholt, D. R., Goddard, M. E., Gordon, S., Martin, N. G., Henders, A. K., Madden, P. A., Visscher, P. M., McEvoy, B. P., and Montgomery, G. W. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, 42(7):565–569.

[Yang et al., 2011a] Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M. (2011a). GCTA: A tool for Genome-wide Complex Trait Analysis. *The American Journal of Human Genetics*, 88(1):76–82.

[Yang et al., 2011b] Yang, J., Weedon, M. N., Purcell, S., Lettre, G., Estrada, K., Willer, C. J., Smith, A. V., Ingelsson, E., O'Connell, J. R., Mangino, M., Maegi, R., Madden, P. A., Heath, A. C., Nyholt, D. R., Martin, N. G., Montgomery, G. W., Frayling, T. M., Hirschhorn, J. N., McCarthy, M. I., Goddard, M. E., and Visscher, P. M. (2011b). Genomic inflation factors under polygenic inheritance. *European Journal of Human Genetics*, 19(7):807–812.

[Zhang and Pugh, 2011] Zhang, Z. and Pugh, B. F. (2011). High-resolution genome-wide mapping of the primary structure of chromatin. *Cell*, 144(2):175–186.

[Zheng et al., 2017] Zheng, J., Erzurumluoglu, A. M., Elsworth, B. L., Kemp, J. P., Howe, L., Haycock, P. C., Hemani, G., Tansey, K., Laurin, C., Genetics, E., Consortium, L. E. E. E., Pourcain, B. S., Warrington, N. M., Finucane, H. K., Price, A. L., Bulik-Sullivan, B. K., Anttila, V., Paternoster, L., Gaunt, T. R., Evans, D. M., and Neale, B. M. (2017). LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics*, 33(2):272–279.

[Zuk et al., 2012] Zuk, O., Hechter, E., Sunyaev, S. R., and Lander, E. S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences of the United States of America*, 109(4):1193–1198.

Appendix A

Supplementary tables

Table S1: Long range LD regions [Price et al., 2008] excluded from the LD Score regression

CHR	**Start (bp)**	**End (bp)**
1	48 000 000	52 000 000
2	86 000 000	10 050 000
2	134 500 000	138 000 000
2	183 000 000	190 000 000
3	47 500 000	52 000 000
3	83 500 000	87 000 000
3	89 000 000	97 500 000
5	44 500 000	50 500 000
5	98 000 000	100 500 000
5	129 000 000	132 000 000
5	135 500 000	138 500 000
6	25 500 000	33 500 000
6	57 000 000	64 000 000
6	140 000 000	142 500 000
7	55 000 000	66 000 000
8	8 000 000	12 000 000
8	43 000 000	50 000 000
8	112 000 000	115 000 000
10	37 000 000	43 000 000
11	46 000 000	57 000 000
11	87 500 000	90 500 000
12	33 000 000	40 000 000
12	109 500 000	112 000 000
20	32 000 000	34 500 000

Table S2: Proportion of SNP-heritability and enrichment of different functional categories in the full baseline model: HDL-C

Annotation	**Prop. SNPs**	**Prop. h_g^2 (s.e.)**	**Enrichment (s.e.)**	**Enrichment p-value**
Coding	0.012	0.123 (0.178)	10.023 (14.487)	0.535
Coding + 500bp	0.062	0.166 (0.244)	2.685 (3.948)	0.668
Conserved	0.026	0.442 (0.288)	17.039 (11.104)	0.125
Conserved + 500bp	0.337	0.538 (0.337)	1.600 (1.000)	0.541
CTCF	0.022	0.081 (0.266)	3.637 (11.972)	0.825
CTCF + 500bp	0.066	-0.182 (0.294)	-2.761 (4.454)	0.383
DGF	0.135	0.272 (0.693)	2.022 (5.149)	0.842
DGF + 500bp	0.531	1.024 (0.390)	1.926 (0.733)	0.207
DHS peaks	0.112	0.104 (0.637)	0.932 (5.683)	0.990
DHS	0.169	0.362 (0.655)	2.140 (3.875)	0.768
DHS + 500bp	0.496	0.899 (0.613)	1.811 (1.235)	0.499
FANTOM5 Enhancer	0.004	0.128 (0.154)	31.684 (38.238)	0.432
FANTOM5 Enhancer + 500bp	0.018	0.094 (0.196)	5.326 (11.084)	0.693
Enhancer	0.060	0.376 (0.422)	6.314 (7.085)	0.452
Enhancer + 500bp	0.144	0.558 (0.340)	3.873 (2.363)	0.217
Fetal DHS	0.084	-0.095 (0.510)	-1.127 (6.057)	0.718
Fetal DHS + 500bp	0.281	1.024 (0.468)	3.652 (1.668)	0.124
H3K27ac (Hnisz)	0.375	1.522 (0.339)	4.062 (0.904)	0.000
H3K27 + 500bp (Hnisz)	0.405	1.238 (0.317)	3.054 (0.781)	0.002
H3K27ac (PGC2)	0.258	1.724 (0.533)	6.683 (2.066)	0.001
H3K27ac + 500bp (PGC2)	0.322	1.333 (0.370)	4.140 (1.150)	0.009
H3K4me1 peaks	0.169	0.911 (0.605)	5.403 (3.588)	0.215
H3K4me1	0.422	1.761 (0.642)	4.174 (1.522)	0.040
H3K4me1 + 500bp	0.600	1.987 (0.409)	3.314 (0.682)	0.000
H3K4me3 peaks	0.039	-0.115 (0.373)	-2.913 (9.447)	0.669
H3K4me3	0.127	0.478 (0.370)	3.752 (2.900)	0.337
H3K4me3 + 500bp	0.245	1.116 (0.511)	4.561 (2.086)	0.049
H3K9ac peaks	0.036	-0.127 (0.448)	-3.491 (12.363)	0.706
H3K9ac	0.119	0.729 (0.371)	6.140 (3.126)	0.091
H3K9ac + 500bp	0.217	1.542 (0.469)	7.111 (2.162)	0.000
Intron	0.392	0.680 (0.191)	1.735 (0.486)	0.087
Intron + 500bp	0.400	0.686 (0.152)	1.718 (0.380)	0.032
PromoterFlanking	0.008	0.137 (0.176)	16.795 (21.540)	0.430
PromoterFlanking + 500bp	0.031	0.490 (0.258)	15.654 (8.253)	0.034
Promoter	0.028	0.154 (0.208)	5.416 (7.347)	0.548
Promoter + 500bp	0.035	0.147 (0.151)	4.163 (4.293)	0.453
Repressed	0.462	0.202 (0.505)	0.438 (1.093)	0.608
Repressed + 500bp	0.721	0.123 (0.238)	0.171 (0.330)	0.004
Super Enhancer	0.156	0.953 (0.237)	6.112 (1.521)	0.000
Super Enhancer + 500bp	0.159	0.826 (0.200)	5.196 (1.259)	0.000
TFBS	0.128	0.929 (0.514)	7.281 (4.026)	0.104
TFBS + 500bp	0.334	0.413 (0.469)	1.234 (1.401)	0.865
Transcribed	0.350	0.486 (0.425)	1.386 (1.213)	0.748
Transcribed + 500bp	0.768	1.366 (0.315)	1.778 (0.410)	0.040
TSS	0.016	0.135 (0.182)	8.347 (11.301)	0.512
TSS + 500bp	0.031	0.642 (0.272)	20.916 (8.848)	0.003
3-prime UTR	0.010	-0.046 (0.144)	-4.373 (13.729)	0.687
3-prime UTR + 500bp	0.025	-0.199 (0.163)	-7.949 (6.524)	0.111
5-prime UTR	0.005	0.025 (0.122)	5.242 (25.864)	0.869
5-prime UTR + 500bp	0.025	-0.080 (0.147)	-3.197 (5.886)	0.459
Weak Enhancer	0.020	0.197 (0.291)	9.964 (14.717)	0.550
Weak Enhancer + 500bp	0.083	0.470 (0.307)	5.673 (3.702)	0.194

Table S3: Proportion of SNP-heritability and enrichment of different functional categories in the full baseline model: LDL-C

Annotation	**Prop. SNPs**	**Prop. h_g^2 (s.e.)**	**Enrichment (s.e.)**	**Enrichment p-value**
Coding	0.012	-0.018 (0.167)	-1.454 (13.574)	0.853
Coding + 500bp	0.062	0.568 (0.183)	9.190 (2.966)	0.001
Conserved	0.026	0.279 (0.230)	10.780 (8.868)	0.249
Conserved + 500bp	0.337	1.304 (0.323)	3.874 (0.959)	0.001
CTCF	0.022	0.063 (0.207)	2.811 (9.312)	0.842
CTCF + 500bp	0.066	0.387 (0.266)	5.860 (4.023)	0.212
DGF	0.135	1.172 (0.458)	8.699 (3.401)	0.032
DGF + 500bp	0.531	0.534 (0.392)	1.006 (0.738)	0.994
DHS peaks	0.112	0.205 (0.475)	1.826 (4.240)	0.845
DHS	0.169	0.324 (0.616)	1.916 (3.648)	0.800
DHS + 500bp	0.496	0.507 (0.374)	1.021 (0.754)	0.977
FANTOM5 Enhancer	0.004	-0.006 (0.091)	-1.517 (22.561)	0.912
FANTOM5 Enhancer + 500bp	0.018	0.061 (0.130)	3.467 (7.335)	0.737
Enhancer	0.060	0.939 (0.303)	15.783 (5.089)	0.006
Enhancer + 500bp	0.144	0.491 (0.234)	3.409 (1.626)	0.156
Fetal DHS	0.084	-0.140 (0.425)	-1.661 (5.049)	0.577
Fetal DHS + 500bp	0.281	0.760 (0.358)	2.709 (1.278)	0.216
H3K27ac (Hnisz)	0.375	0.910 (0.275)	2.428 (0.733)	0.078
H3K27 + 500bp (Hnisz)	0.405	1.090 (0.269)	2.687 (0.662)	0.022
H3K27ac (PGC2)	0.258	0.793 (0.332)	3.073 (1.288)	0.112
H3K27ac + 500bp (PGC2)	0.322	0.891 (0.243)	2.768 (0.755)	0.025
H3K4me1 peaks	0.169	0.485 (0.615)	2.876 (3.645)	0.597
H3K4me1	0.422	1.439 (0.399)	3.411 (0.946)	0.015
H3K4me1 + 500bp	0.600	1.043 (0.263)	1.740 (0.439)	0.103
H3K4me3 peaks	0.039	0.113 (0.303)	2.874 (7.685)	0.803
H3K4me3	0.127	0.979 (0.470)	7.682 (3.688)	0.065
H3K4me3 + 500bp	0.245	1.255 (0.522)	5.128 (2.134)	0.031
H3K9ac peaks	0.036	0.226 (0.326)	6.242 (8.980)	0.534
H3K9ac	0.119	1.289 (0.379)	10.854 (3.186)	0.001
H3K9ac + 500bp	0.217	0.882 (0.397)	4.066 (1.833)	0.099
Intron	0.392	0.519 (0.158)	1.325 (0.404)	0.410
Intron + 500bp	0.400	0.584 (0.171)	1.463 (0.428)	0.217
PromoterFlanking	0.008	0.038 (0.137)	4.697 (16.819)	0.824
PromoterFlanking + 500bp	0.031	0.428 (0.220)	13.668 (7.015)	0.051
Promoter	0.028	0.143 (0.175)	5.046 (6.172)	0.487
Promoter + 500bp	0.035	0.239 (0.138)	6.776 (3.912)	0.082
Repressed	0.462	0.004 (0.386)	0.008 (0.836)	0.222
Repressed + 500bp	0.721	-0.016 (0.210)	-0.023 (0.291)	0.000
Super Enhancer	0.156	0.456 (0.140)	2.925 (0.898)	0.008
Super Enhancer + 500bp	0.159	0.447 (0.142)	2.811 (0.893)	0.017
TFBS	0.128	1.434 (0.487)	11.236 (3.818)	0.007
TFBS + 500bp	0.334	1.035 (0.382)	3.095 (1.141)	0.050
Transcribed	0.350	0.477 (0.352)	1.362 (1.004)	0.707
Transcribed + 500bp	0.768	1.109 (0.333)	1.444 (0.433)	0.298
TSS	0.016	0.135 (0.204)	8.402 (12.653)	0.519
TSS + 500bp	0.031	0.518 (0.183)	16.88 (5.969)	0.002
3-prime UTR	0.010	0.021 (0.113)	2.011 (10.740)	0.922
3-prime UTR + 500bp	0.025	0.140 (0.125)	5.579 (4.988)	0.348
5-prime UTR	0.005	-0.084 (0.109)	-7.218 (23.076)	0.719
5-prime UTR + 500bp	0.025	0.039 (0.126)	1.546 (5.052)	0.911
Weak Enhancer	0.020	0.409 (0.206)	20.671 (10.394)	0.091
Weak Enhancer + 500bp	0.083	0.243 (0.272)	2.940 (3.288)	0.556

Table S4: Proportion of SNP-heritability and enrichment of different functional categories in the full baseline model: TC

Annotation	**Prop. SNPs**	**Prop. h_g^2 (s.e.)**	**Enrichment (s.e.)**	**Enrichment p-value**
Coding	0.012	0.031 (0.156)	2.524 (12.694)	0.902
Coding + 500bp	0.062	0.644 (0.190)	10.428 (3.076)	0.000
Conserved	0.026	0.414 (0.229)	15.957 (8.850)	0.077
Conserved + 500bp	0.337	1.353 (0.326)	4.021 (0.969)	0.000
CTCF	0.022	0.094 (0.218)	4.212 (9.804)	0.742
CTCF + 500bp	0.066	0.421 (0.241)	6.383 (3.642)	0.149
DGF	0.135	1.193 (0.450)	8.859 (3.342)	0.020
DGF + 500bp	0.531	0.599 (0.362)	1.127 (0.681)	0.850
DHS peaks	0.112	0.256 (0.438)	2.281 (3.909)	0.745
DHS	0.169	0.686 (0.536)	4.063 (3.174)	0.352
DHS + 500bp	0.496	0.423 (0.377)	0.853 (0.760)	0.843
FANTOM5 Enhancer	0.004	-0.039 (0.083)	-9.676 (20.711)	0.611
FANTOM5 Enhancer + 500bp	0.018	0.087 (0.114)	4.941 (6.451)	0.539
Enhancer	0.060	0.898 (0.277)	15.092 (4.661)	0.005
Enhancer + 500bp	0.144	0.690 (0.218)	4.788 (1.510)	0.016
Fetal DHS	0.084	0.111 (0.381)	1.324 (4.525)	0.942
Fetal DHS + 500bp	0.281	0.746 (0.326)	2.659 (1.163)	0.186
H3K27ac (Hnisz)	0.375	0.828 (0.246)	2.209 (0.656)	0.088
H3K27 + 500bp (Hnisz)	0.405	1.096 (0.238)	2.703 (0.588)	0.007
H3K27ac (PGC2)	0.258	1.168 (0.334)	4.526 (1.296)	0.008
H3K27ac + 500bp (PGC2)	0.322	0.970 (0.230)	3.012 (0.715)	0.006
H3K4me1 peaks	0.169	0.544 (0.532)	3.228 (3.155)	0.475
H3K4me1	0.422	1.412 (0.395)	3.347 (0.936)	0.012
H3K4me1 + 500bp	0.600	1.256 (0.247)	2.095 (0.411)	0.011
H3K4me3 peaks	0.039	0.195 (0.293)	4.942 (7.421)	0.595
H3K4me3	0.127	1.005 (0.398)	7.884 (3.120)	0.020
H3K4me3 + 500bp	0.245	1.162 (0.419)	4.747 (1.713)	0.012
H3K9ac peaks	0.036	0.052 (0.280)	1.426 (7.715)	0.956
H3K9ac	0.119	1.395 (0.376)	11.745 (3.168)	0.000
H3K9ac + 500bp	0.217	1.029 (0.353)	4.743 (1.629)	0.024
Intron	0.392	0.539 (0.149)	1.375 (0.379)	0.304
Intron + 500bp	0.400	0.598 (0.154)	1.497 (0.385)	0.135
PromoterFlanking	0.008	-0.016 (0.134)	-2.004 (16.401)	0.852
PromoterFlanking + 500bp	0.031	0.374 (0.187)	11.931 (5.974)	0.053
Promoter	0.028	0.136 (0.160)	4.788 (5.654)	0.494
Promoter + 500bp	0.035	0.201 (0.122)	5.692 (3.465)	0.143
Repressed	0.462	-0.102 (0.342)	-0.221 (0.740)	0.090
Repressed + 500bp	0.721	-0.028 (0.193)	-0.039 (0.267)	0.000
Super Enhancer	0.156	0.541 (0.138)	3.468 (0.886)	0.001
Super Enhancer + 500bp	0.159	0.522 (0.132)	3.287 (0.832)	0.001
TFBS	0.128	1.429 (0.442)	11.198 (3.461)	0.005
TFBS + 500bp	0.334	0.922 (0.325)	2.756 (0.971)	0.062
Transcribed	0.350	0.479 (0.324)	1.367 (0.926)	0.681
Transcribed + 500bp	0.768	1.046 (0.287)	1.362 (0.374)	0.324
TSS	0.016	0.174 (0.178)	10.776 (11.056)	0.344
TSS + 500bp	0.031	0.444 (0.175)	14.474 (5.716)	0.011
3-prime UTR	0.010	0.015 (0.107)	1.443 (10.237)	0.965
3-prime UTR + 500bp	0.025	0.144 (0.121)	5.738 (4.838)	0.319
5-prime UTR	0.005	-0.082 (0.087)	-6.904 (18.443)	0.667
5-prime UTR + 500bp	0.025	0.132 (0.126)	5.289 (5.019)	0.356
Weak Enhancer	0.020	0.482 (0.168)	24.358 (8.488)	0.017
Weak Enhancer + 500bp	0.083	0.363 (0.227)	4.378 (2.742)	0.228

Table S5: Proportion of SNP-heritability and enrichment of different functional categories in the full baseline model: TG

Annotation	Prop. SNPs	Prop. h_g^2 (s.e.)	Enrichment (s.e.)	Enrichment p-value
Coding	0.012	0.149 (0.186)	12.083 (15.166)	0.413
Coding + 500bp	0.062	0.419 (0.183)	6.782 (2.972)	0.055
Conserved	0.026	0.620 (0.265)	23.921 (10.238)	0.007
Conserved + 500bp	0.337	0.995 (0.304)	2.957 (0.903)	0.055
CTCF	0.022	0.313 (0.210)	14.090 (9.456)	0.150
CTCF + 500bp	0.066	-0.050 (0.244)	-0.755 (3.697)	0.630
DGF	0.135	0.606 (0.663)	4.502 (4.923)	0.413
DGF + 500bp	0.531	0.648 (0.379)	1.219 (0.713)	0.757
DHS peaks	0.112	0.848 (0.578)	7.572 (5.157)	0.151
DHS	0.169	1.027 (0.623)	6.079 (3.688)	0.141
DHS + 500bp	0.496	1.108 (0.519)	2.233 (1.045)	0.230
FANTOM5 Enhancer	0.004	0.047 (0.102)	11.586 (25.433)	0.678
FANTOM5 Enhancer + 500bp	0.018	0.065 (0.153)	3.698 (8.638)	0.732
Enhancer	0.060	0.344 (0.373)	5.784 (6.262)	0.436
Enhancer + 500bp	0.144	0.727 (0.349)	5.045 (2.421)	0.087
Fetal DHS	0.084	0.597 (0.439)	7.094 (5.220)	0.235
Fetal DHS + 500bp	0.281	0.803 (0.416)	2.864 (1.485)	0.239
H3K27ac (Hnisz)	0.375	1.108 (0.251)	2.958 (0.671)	0.002
H3K27 + 500bp (Hnisz)	0.405	1.223 (0.250)	3.015 (0.618)	0.000
H3K27ac (PGC2)	0.258	2.174 (0.568)	8.427 (2.203)	0.003
H3K27ac + 500bp (PGC2)	0.322	1.199 (0.306)	3.725 (0.951)	0.013
H3K4me1 peaks	0.169	0.796 (0.512)	4.719 (3.034)	0.191
H3K4me1	0.422	1.352 (0.515)	3.206 (1.221)	0.050
H3K4me1 + 500bp	0.600	1.733 (0.315)	2.891 (0.525)	0.001
H3K4me3 peaks	0.039	0.260 (0.339)	6.584 (8.582)	0.526
H3K4me3	0.127	0.651 (0.377)	5.109 (2.956)	0.112
H3K4me3 + 500bp	0.245	1.543 (0.436)	6.303 (1.781)	0.001
H3K9ac peaks	0.036	0.908 (0.345)	25.043 (9.506)	0.007
H3K9ac	0.119	0.942 (0.362)	7.929 (3.049)	0.005
H3K9ac + 500bp	0.217	1.088 (0.389)	5.017 (1.792)	0.038
Intron	0.392	0.498 (0.176)	1.270 (0.450)	0.542
Intron + 500bp	0.400	0.676 (0.131)	1.691 (0.327)	0.022
PromoterFlanking	0.008	0.035 (0.162)	4.347 (19.851)	0.858
PromoterFlanking + 500bp	0.031	0.193 (0.195)	6.149 (6.238)	0.356
Promoter	0.028	0.484 (0.184)	17.072 (6.471)	0.007
Promoter + 500bp	0.035	0.247 (0.143)	7.006 (4.047)	0.116
Repressed	0.462	-0.702 (0.503)	-1.520 (1.090)	0.010
Repressed + 500bp	0.721	0.094 (0.213)	0.130 (0.295)	0.000
Super Enhancer	0.156	0.618 (0.161)	3.959 (1.032)	0.008
Super Enhancer + 500bp	0.159	0.504 (0.122)	3.171 (0.767)	0.008
TFBS	0.128	1.002 (0.436)	7.853 (3.419)	0.035
TFBS + 500bp	0.334	0.480 (0.376)	1.435 (1.124)	0.690
Transcribed	0.350	0.456 (0.364)	1.301 (1.038)	0.771
Transcribed + 500bp	0.768	0.883 (0.281)	1.150 (0.365)	0.669
TSS	0.016	0.142 (0.172)	8.795 (10.665)	0.475
TSS + 500bp	0.031	0.391 (0.186)	12.73 (6.073)	0.022
3-prime UTR	0.01	-0.061 (0.108)	-5.800 (10.269)	0.509
3-prime UTR + 500bp	0.025	-0.059 (0.114)	-2.379 (4.556)	0.424
5-prime UTR	0.005	0.181 (0.094)	38.540 (20.075)	0.040
5-prime UTR + 500bp	0.025	0.259 (0.148)	10.352 (5.916)	0.079
Weak Enhancer	0.020	0.330 (0.259)	16.668 (13.075)	0.250
Weak Enhancer + 500bp	0.083	0.539 (0.303)	6.503 (3.660)	0.150

Table S6: Cell type group enrichment: HDL-C, $-log_{10}(P)$ of τ_c

Category	**Prop. SNPs**	**Prop.** h_g^2	**Coefficient z-score**	$-log_{10}(P)$
Adrenal or pancreas	0.09	0.56	0.36	0.07
Cardiovascular	0.11	0.52	0.09	0.11
CNS	0.14	0.83	0.65	0.56
Connective or bone	0.11	0.62	-0.10	0.01
GI	0.16	0.74	-0.45	0.03
Immune	0.22	1.20	0.66	0.08
Kidney	0.04	0.45	1.44	0.87
Liver	0.07	1.03	3.57	3.20
Other	0.20	1.26	2.01	1.36
Skeletal muscle	0.10	0.92	2.84	2.02

Table S7: Cell type group enrichment: LDL-C, $-log_{10}(P)$ of τ_c

Category	**Prop. SNPs**	**Prop.** h_g^2	**Coefficient z-score**	$-log_{10}(P)$
Adrenal or pancreas	0.09	0.84	2.54	1.84
Cardiovascular	0.11	0.38	-1.50	1.10
CNS	0.14	0.62	-0.25	0.15
Connective or bone	0.11	0.72	0.46	0.24
GI	0.16	1.13	2.59	1.69
Immune	0.22	0.99	1.84	1.23
Kidney	0.04	0.48	1.81	1.14
Liver	0.07	0.81	2.32	1.57
Other	0.20	1.09	0.78	0.25
Skeletal muscle	0.10	0.69	1.14	0.71

Table S8: Cell type group enrichment: TC, $-log_{10}(P)$ of τ_c

Category	**Prop. SNPs**	**Prop.** h_g^2	**Coefficient z-score**	$-log_{10}(P)$
Adrenal or pancreas	0.09	0.89	1.46	0.84
Cardiovascular	0.11	0.72	0.62	0.27
CNS	0.14	1.09	0.55	0.23
Connective or bone	0.11	0.79	0.08	0.03
GI	0.16	1.07	0.27	0.11
Immune	0.22	1.37	0.46	0.19
Kidney	0.04	0.52	0.83	0.39
Liver	0.07	0.80	1.68	1.03
Other	0.20	0.93	-0.14	0.05
Skeletal muscle	0.10	0.70	0.18	0.07

Table S9: Cell type group enrichment: TG, $-log_{10}(P)$ of τ_c

Category	**Prop. SNPs**	**Prop.** h_g^2	**Coefficient z-score**	$-log_{10}(P)$
Adrenal or pancreas	0.09	0.89	1.46	0.84
Cardiovascular	0.11	0.72	0.62	0.27
CNS	0.14	1.09	0.55	0.23
Connective or bone	0.11	0.79	0.08	0.03
GI	0.16	1.07	0.27	0.11
Immune	0.22	1.37	0.46	0.19
Kidney	0.04	0.52	0.83	0.39
Liver	0.07	0.80	1.68	1.03
Other	0.20	0.93	-0.14	0.05
Skeletal muscle	0.10	0.70	0.18	0.07

Table S10: Cell type specific enrichment: $-log_{10}(P)$ of τ_c for four blood lipid levels in the National FINRISK study

Cell type	**cell type group**	**mark**	**HDL-C**	**LDL-C**	**TC**	**TG**
Fetal adrenal	Adrenal/Pancreas	H3K4me1	0.26	3.50	3.80	0.23
Fetal adrenal	Adrenal/Pancreas	H3K4me3	0.30	0.09	0.19	0.24
Pancreas	Adrenal/Pancreas	H3K4me1	0.59	1.37	1.27	0.55
Pancreas	Adrenal/Pancreas	H3K4me3	0.00	0.04	0.03	0.50
Pancreatic islets	Adrenal/Pancreas	H3K4me1	0.13	0.94	1.02	0.51
Pancreatic islets	Adrenal/Pancreas	H3K4me1	0.08	0.85	0.76	0.79
Pancreatic islets	Adrenal/Pancreas	H3K4me3	0.19	0.62	0.68	0.29
Pancreatic islets	Adrenal/Pancreas	H3K4me3	0.06	0.37	0.32	0.30
Pancreatic islets	Adrenal/Pancreas	H3K9ac	0.02	0.02	0.05	0.23
Pancreatic islets	Adrenal/Pancreas	H3K27ac	0.47	0.33	0.28	0.20
Aorta	Cardiovascular	H3K4me3	1.23	0.00	0.17	1.16
Fetal heart	Cardiovascular	H3K4me1	0.14	2.00	2.15	0.43
Fetal heart	Cardiovascular	H3K4me3	0.17	1.29	1.50	0.14
Fetal heart	Cardiovascular	H3K9ac	0.06	1.45	1.60	0.04
Fetal lung	Cardiovascular	H3K4me1	0.01	0.37	0.24	0.47
Fetal lung	Cardiovascular	H3K4me3	0.78	0.07	0.19	0.06
Fetal lung	Cardiovascular	H3K9ac	0.76	0.04	0.17	0.11
Left Ventricle	Cardiovascular	H3K4me1	0.92	0.87	0.78	0.20
Left Ventricle	Cardiovascular	H3K4me3	0.02	0.62	0.59	0.20
Lung	Cardiovascular	H3K4me1	0.50	1.02	1.00	0.46
Lung	Cardiovascular	H3K4me3	0.27	0.30	0.34	0.86
Right atrium	Cardiovascular	H3K4me1	0.83	1.42	0.95	0.49
Right atrium	Cardiovascular	H3K4me3	0.42	0.67	0.61	0.17
Right ventricle	Cardiovascular	H3K4me1	0.10	1.76	2.94	0.39
Right ventricle	Cardiovascular	H3K4me3	0.02	0.76	0.72	0.50
Angular gyrus	CNS	H3K4me1	0.42	1.39	1.17	0.28

Table S10 continued from previous page

Cell type	cell type group	mark	HDL-C	LDL-C	TC	TG
Angular gyrus	CNS	H3K4me3	0.64	0.29	0.01	0.66
Angular gyrus	CNS	H3K9ac	0.63	0.79	0.61	0.24
Angular gyrus	CNS	H3K27ac	0.22	0.15	0.25	0.25
Anterior caudate	CNS	H3K4me1	1.09	0.06	0.14	0.47
Anterior caudate	CNS	H3K4me3	0.55	0.48	0.70	0.18
Anterior caudate	CNS	H3K9ac	0.87	0.46	0.48	0.28
Anterior caudate	CNS	H3K27ac	1.27	0.02	0.08	0.61
Cingulate gyrus	CNS	H3K4me1	0.86	0.02	0.00	0.34
Cingulate gyrus	CNS	H3K4me3	1.29	0.09	0.12	0.89
Cingulate gyrus	CNS	H3K9ac	0.51	1.22	0.92	0.15
Cingulate gyrus	CNS	H3K27ac	1.04	0.07	0.10	0.48
Fetal brain	CNS	H3K4me1	0.64	0.03	0.27	0.37
Fetal brain	CNS	H3K4me3	0.22	0.01	0.47	1.69
Fetal brain	CNS	H3K4me3	0.04	0.37	0.49	0.04
Fetal brain	CNS	H3K9ac	0.05	0.99	0.80	0.16
Germinal matrix	CNS	H3K4me3	0.33	0.11	0.11	0.06
Hippocampus middle	CNS	H3K4me1	1.03	0.08	0.12	0.45
Hippocampus middle	CNS	H3K4me3	0.85	0.03	0.28	0.71
Hippocampus middle	CNS	H3K9ac	0.47	0.51	0.39	0.01
Hippocampus middle	CNS	H3K27ac	0.90	0.08	0.08	0.46
Inferior temporal lobe	CNS	H3K4me1	0.66	0.10	0.06	0.24
Inferior temporal lobe	CNS	H3K4me3	1.06	0.14	0.09	0.59
Inferior temporal lobe	CNS	H3K9ac	1.06	0.63	0.56	0.13
Inferior temporal lobe	CNS	H3K27ac	0.65	0.16	0.25	0.31
Mid frontal lobe	CNS	H3K4me1	0.01	1.69	1.71	0.08
Mid frontal lobe	CNS	H3K4me3	0.71	0.08	0.01	0.35
Mid frontal lobe	CNS	H3K9ac	0.30	0.36	0.40	0.62
Mid frontal lobe	CNS	H3K27ac	0.98	0.21	0.06	0.02

Table S10 continued from previous page

Cell type	cell type group	mark	HDL-C	LDL-C	TC	TG
Neurosphere	CNS	H3K27ac	0.12	1.01	1.02	0.30
Substantia nigra	CNS	H3K4me1	1.48	0.03	0.05	0.47
Substantia nigra	CNS	H3K4me3	1.12	0.12	0.27	0.79
Substantia nigra	CNS	H3K9ac	1.13	0.21	0.11	0.12
Substantia nigra	CNS	H3K27ac	1.79	0.09	0.11	0.72
Breast fibroblast primary	Connective/Bone	H3K4me1	0.06	0.77	0.74	0.04
Breast fibroblast primary	Connective/Bone	H3K4me3	0.05	0.97	0.63	0.04
Chondrogenic dif	Connective/Bone	H3K27ac	0.05	1.44	2.00	0.06
Osteoblast	Connective/Bone	H3K27ac	0.61	0.70	1.20	0.07
Penis foreskin fibroblast primary	Connective/Bone	H3K4me1	0.09	0.46	0.30	0.02
Penis foreskin fibroblast primary	Connective/Bone	H3K4me3	0.28	0.79	0.83	0.23
Colon smooth muscle	Gastrointestinal	H3K4me1	0.08	0.60	0.39	0.02
Colon smooth muscle	Gastrointestinal	H3K4me3	0.33	0.20	0.27	0.25
Colon smooth muscle	Gastrointestinal	H3K9ac	0.14	0.11	0.29	0.40
Colon smooth muscle	Gastrointestinal	H3K27ac	0.08	0.87	0.60	0.01
Colonic mucosa	Gastrointestinal	H3K4me1	0.56	1.77	2.07	0.05
Colonic mucosa	Gastrointestinal	H3K4me3	0.03	0.08	0.16	0.38
Colonic mucosa	Gastrointestinal	H3K9ac	0.33	1.13	1.64	0.46
Colonic mucosa	Gastrointestinal	H3K27ac	0.07	1.10	1.36	0.03
Duodenum Mucosa	Gastrointestinal	H3K4me1	0.13	1.56	2.12	1.03
Duodenum Mucosa	Gastrointestinal	H3K4me3	0.39	0.29	0.67	0.44
Duodenum Mucosa	Gastrointestinal	H3K9ac	0.26	1.22	1.95	0.61
Duodenum mucosa	Gastrointestinal	H3K27ac	0.23	1.56	1.70	0.79
Duodenum smooth muscle	Gastrointestinal	H3K4me1	0.69	0.60	0.79	0.16
Duodenum smooth muscle	Gastrointestinal	H3K4me3	0.75	0.17	0.09	0.23
Duodenum smooth muscle	Gastrointestinal	H3K27ac	0.43	0.29	0.70	0.76
Esophagus	Gastrointestinal	H3K4me1	0.99	1.71	2.62	2.09
Esophagus	Gastrointestinal	H3K4me3	0.24	0.22	0.49	0.25

Table S10 continued from previous page

Cell type	cell type group	mark	HDL-C	LDL-C	TC	TG
Fetal large intestine	Gastrointestinal	H3K4me1	0.26	1.20	2.24	0.97
Fetal large intestine	Gastrointestinal	H3K4me3	0.38	0.52	1.73	0.69
Fetal small intestine	Gastrointestinal	H3K4me1	0.43	1.08	1.96	0.90
Fetal small intestine	Gastrointestinal	H3K4me3	0.32	0.48	1.41	0.36
Fetal stomach	Gastrointestinal	H3K4me1	1.19	0.30	0.47	0.13
Fetal stomach	Gastrointestinal	H3K4me3	0.48	0.11	0.25	0.94
Gastric	Gastrointestinal	H3K4me1	0.77	1.50	2.13	0.33
Gastric	Gastrointestinal	H3K4me3	0.37	0.31	0.33	0.25
Rectal mucosa	Gastrointestinal	H3K4me1	0.68	1.74	1.86	0.21
Rectal mucosa	Gastrointestinal	H3K4me3	0.13	0.04	0.42	0.13
Rectal mucosa	Gastrointestinal	H3K9ac	0.04	0.68	1.01	0.39
Rectal mucosa	Gastrointestinal	H3K27ac	0.34	1.74	1.47	0.16
Rectal smooth muscle	Gastrointestinal	H3K4me1	0.68	0.89	0.61	0.29
Rectal smooth muscle	Gastrointestinal	H3K4me3	0.26	0.06	0.10	0.23
Rectal smooth muscle	Gastrointestinal	H3K9ac	0.17	0.73	0.79	0.21
Rectal smooth muscle	Gastrointestinal	H3K27ac	0.32	0.30	0.20	0.46
Sigmoid colon	Gastrointestinal	H3K4me1	0.38	0.38	0.35	0.26
Sigmoid colon	Gastrointestinal	H3K4me3	0.08	0.47	0.37	0.06
Small intestine	Gastrointestinal	H3K4me1	0.42	1.54	0.81	1.04
Small intestine	Gastrointestinal	H3K4me3	0.05	0.42	0.40	0.25
Stomach mucosa	Gastrointestinal	H3K4me1	0.04	1.60	2.51	0.70
Stomach mucosa	Gastrointestinal	H3K4me3	0.11	0.26	0.22	0.66
Stomach mucosa	Gastrointestinal	H3K9ac	0.03	1.01	2.11	0.01
Stomach smooth muscle	Gastrointestinal	H3K4me1	0.78	1.08	1.16	0.14
Stomach smooth muscle	Gastrointestinal	H3K4me3	0.30	0.18	0.28	0.11
Stomach smooth muscle	Gastrointestinal	H3K9ac	0.23	0.50	0.77	0.23
Stomach smooth muscle	Gastrointestinal	H3K27ac	0.17	0.41	0.51	0.14
CD14	Immune	H3K27ac	0.20	0.37	0.03	0.32

Table S10 continued from previous page

Cell type	cell type group	mark	HDL-C	LDL-C	TC	TG
CD14 primary	Immune	H3K4me1	0.98	0.03	0.07	0.84
CD14 primary	Immune	H3K4me3	0.27	0.46	0.14	0.02
CD15 primary	Immune	H3K4me1	0.07	0.29	0.19	0.03
CD15 primary	Immune	H3K4me3	0.15	0.35	0.32	0.20
CD19	Immune	H3K27ac	0.09	0.30	0.60	1.35
CD19 primary (BI)	Immune	H3K4me1	0.18	0.14	0.03	0.05
CD19 primary (BI)	Immune	H3K4me3	0.21	0.88	0.85	0.03
CD19 primary (UW)	Immune	H3K4me1	0.06	0.23	0.18	0.28
CD19 primary (UW)	Immune	H3K4me3	0.38	0.16	0.02	0.30
CD20	Immune	H3K27ac	0.01	0.71	0.93	0.95
CD25- CD45RA+ naive	Immune	H3K27ac	0.15	0.03	0.02	1.39
CD25- IL17- Th stim MACS	Immune	H3K27ac	1.09	0.62	0.44	1.21
CD25- IL17+ Th17 stim	Immune	H3K27ac	2.15	0.82	0.57	1.18
CD25+ CD127- Treg	Immune	H3K27ac	1.06	0.66	0.73	1.24
CD25int CD127+ Tmem	Immune	H3K27ac	0.34	0.21	0.56	0.11
CD3 primary	Immune	H3K27ac	0.85	0.22	0.21	0.67
CD3 primary (BI)	Immune	H3K4me1	0.04	0.49	0.44	0.22
CD3 primary (BI)	Immune	H3K4me3	0.13	0.96	1.01	0.72
CD3 primary (UW)	Immune	H3K4me1	0.48	0.08	0.04	0.83
CD3 primary (UW)	Immune	H3K4me3	0.66	0.22	0.20	0.36
CD34 primary	Immune	H3K4me1	0.03	0.02	0.04	0.53
CD34 primary	Immune	H3K4me3	0.06	0.37	0.14	0.10
CD4 memory primary	Immune	H3K4me1	0.57	0.30	0.29	0.09
CD4 memory primary	Immune	H3K4me3	0.69	0.79	0.53	0.59
CD4 naive primary	Immune	H3K4me1	0.20	0.34	0.35	0.60
CD4 naive primary	Immune	H3K4me3	0.15	0.40	0.44	0.63
CD4 primary	Immune	H3K4me3	0.67	0.58	0.43	0.93
CD4+ CD25- CD45R0+ memory primary	Immune	H3K4me1	0.38	0.69	0.59	0.09

Table S10 continued from previous page

Cell type	cell type group	mark	HDL-C	LDL-C	TC	TG
CD4+ CD25- CD45R0+ memory primary	Immune	H3K4me3	0.66	1.08	1.07	0.28
CD4+ CD25- CD45RA+ naive primary	Immune	H3K4me1	0.07	0.18	0.10	0.96
CD4+ CD25- CD45RA+ naive primary	Immune	H3K4me3	0.01	0.47	0.40	0.95
CD4+ CD25- IL17- PMA Ionomycin stim MACS Th sprimary	Immune	H3K4me1	0.25	0.32	0.26	0.10
CD4+ CD25- IL17- PMA Ionomycin stim MACS Th sprimary	Immune	H3K4me3	0.46	1.06	1.39	1.77
CD4+ CD25- IL17+ PMA Ionomycin stim Th17 primary	Immune	H3K4me1	0.93	0.82	0.65	0.11
CD4+ CD25- IL17+ PMA Ionomycin stim Th17 primary	Immune	H3K4me3	1.21	0.74	0.53	0.37
CD4+ CD25- Th primary	Immune	H3K4me1	0.24	0.46	0.53	0.47
CD4+ CD25- Th primary	Immune	H3K4me3	0.31	0.35	0.24	0.85
CD4+ CD25+ CD127- Treg primary	Immune	H3K4me1	0.53	0.56	0.57	0.99
CD4+ CD25+ CD127- Treg primary	Immune	H3K4me3	0.42	0.46	0.28	0.36
CD4+ CD25int CD127+ Tmem primary	Immune	H3K4me1	0.25	0.21	0.06	0.14
CD4+ CD25int CD127+ Tmem primary	Immune	H3K4me3	0.72	0.61	0.43	0.77
CD56 primary	Immune	H3K4me1	0.15	0.30	0.13	0.01
CD56 primary	Immune	H3K4me3	0.02	0.33	0.16	0.01
CD8 memory primary	Immune	H3K4me1	0.18	0.24	0.26	0.43
CD8 memory primary	Immune	H3K4me3	0.01	0.52	0.39	0.35
CD8 naive primary (BI)	Immune	H3K4me1	0.35	0.44	0.57	0.85
CD8 naive primary (BI)	Immune	H3K4me3	0.76	0.36	0.31	0.8
CD8 naive primary (UCSF-UBC)	Immune	H3K4me1	0.18	0.43	0.40	0.51
CD8 naive primary (UCSF-UBC)	Immune	H3K4me3	0.48	1.04	1.06	0.75
CD8 naive primary (UCSF-UBC)	Immune	H3K9ac	0.18	0.11	0.05	0.32
CD8 primary	Immune	H3K4me3	0.68	0.26	0.19	0.44
Fetal thymus	Immune	H3K4me1	0.28	0.50	0.62	0.04
Fetal thymus	Immune	H3K4me3	0.10	0.08	0.09	1.16
Mobilized CD34	Immune	H3K27ac	0.27	1.15	2.00	0.19
Mobilized CD34 primary	Immune	H3K4me1	0.42	0.65	0.75	0.61
Mobilized CD34 primary	Immune	H3K4me3	0.06	0.00	0.52	0.43

Table S10 continued from previous page

Cell type	cell type group	mark	HDL-C	LDL-C	TC	TG
Peripheralblood mononuclear primary	Immune	H3K4me1	0.04	0.26	0.07	0.06
Peripheralblood mononuclear primary	Immune	H3K4me3	0.24	0.50	0.28	0.01
Peripheralblood mononuclear primary	Immune	H3K9ac	0.74	0.03	0.05	0.39
Spleen	Immune	H3K4me1	0.34	1.68	1.78	0.15
Th0	Immune	H3K27ac	1.51	0.48	0.66	0.16
Th1	Immune	H3K27ac	2.49	0.79	1.03	0.35
Th2	Immune	H3K27ac	2.39	0.20	0.00	0.50
Thymus	Immune	H3K4me1	0.58	0.85	0.84	0.22
Treg primary	Immune	H3K4me3	0.54	1.32	1.57	0.84
Fetal kidney	Kidney	H3K9ac	0.30	0.36	0.08	0.04
Kidney	Kidney	H3K4me1	0.16	2.02	2.28	0.14
Kidney	Kidney	H3K4me3	0.11	0.05	0.25	0.12
Kidney	Kidney	H3K9ac	0.42	0.09	0.59	0.04
Kidney	Kidney	H3K27ac	0.86	0.99	1.40	0.52
Liver	Liver	H3K27ac	2.84	1.20	2.25	1.54
Liver (BI)	Liver	H3K4me1	2.63	2.30	3.86	1.25
Liver (BI)	Liver	H3K4me3	2.11	0.29	0.57	0.09
Liver (BI)	Liver	H3K9ac	1.79	0.84	1.69	0.63
Liver (UCSD)	Liver	H3K4me1	0.38	0.60	1.14	0.5
Liver (UCSD)	Liver	H3K4me3	1.48	1.14	2.15	0.38
Adipose nuclei	Other	H3K4me1	1.73	0.57	0.92	0.07
Adipose nuclei	Other	H3K4me3	1.36	0.13	0.13	0.18
Adipose nuclei	Other	H3K9ac	1.28	1.52	1.50	0.15
Adipose nuclei	Other	H3K27ac	2.03	0.61	1.09	0.33
Breast luminal epithelial	Other	H3K4me1	0.05	0.77	1.76	0.38
Breast myoepithelial	Other	H3K4me1	0.38	0.57	1.16	1.75
Breast myoepithelial	Other	H3K4me3	0.14	0.40	0.15	0.06
Breast myoepithelial	Other	H3K9ac	0.1	0.15	0.56	0.31

Table S10 continued from previous page

Cell type	cell type group	mark	HDL-C	LDL-C	TC	TG
Breast vHMEC	Other	H3K4me1	0.34	0.17	0.11	0.70
Breast vHMEC	Other	H3K4me3	0.00	0.18	0.10	0.09
Fetal placenta	Other	H3K4me1	0.25	1.39	1.69	0.20
Fetal placenta	Other	H3K4me3	0.02	0.23	0.03	0.02
Ovary	Other	H3K4me1	0.45	1.26	1.22	0.86
Ovary	Other	H3K4me3	0.05	0.09	0.11	0.82
Penis foreskin keratinocyte primary	Other	H3K4me1	0.16	0.35	0.40	1.23
Penis foreskin keratinocyte primary	Other	H3K4me3	0.23	0.40	0.37	0.35
Penis foreskin keratinocyte primary	Other	H3K9ac	0.62	1.02	1.21	0.76
Penis foreskin melanocyte primary	Other	H3K4me1	0.65	0.63	0.42	0.04
Penis foreskin melanocyte primary	Other	H3K4me3	0.43	0.69	0.81	0.22
Placenta amnion	Other	H3K4me1	0.26	0.12	0.06	0.29
Placenta amnion	Other	H3K4me3	0.08	0.49	0.24	0.06
Placenta chorion	Other	H3K4me1	0.09	0.53	0.72	0.35
Placenta chorion	Other	H3K4me3	0.38	0.26	0.02	0.27
Fetal leg muscle	Skeletal muscle	H3K4me1	0.41	1.21	1.10	0.16
Fetal leg muscle	Skeletal muscle	H3K4me3	0.18	0.78	0.60	0.11
Fetal trunk muscle	Skeletal muscle	H3K4me1	0.24	1.63	1.30	0.34
Fetal trunk muscle	Skeletal muscle	H3K4me3	0.53	0.51	0.29	0.11
Psoas muscle	Skeletal muscle	H3K4me1	0.57	0.87	1.04	0.69
Psoas muscle	Skeletal muscle	H3K4me3	0.20	0.52	0.19	0.00
Skeletal muscle	Skeletal muscle	H3K4me1	1.71	0.59	1.01	0.28
Skeletal muscle	Skeletal muscle	H3K4me3	0.84	0.24	0.06	0.25
Skeletal muscle	Skeletal muscle	H3K9ac	0.46	0.98	1.43	0.15
Skeletal muscle	Skeletal muscle	H3K27ac	1.09	0.05	0.13	0.05

Appendix B

Funtional annotations in the full baseline model

Coding annotation includes only regions that are transcribed into proteins, called exons. Broader definition of gene-coding regions include also introns, 5'UTR and 3'UTR regions. Only 11% of GWAS hits tags coding regions [McVicker et al., 2013].

Conserved annotation refers to those parts of the genome where the sequence has been maintained similar during evolution. It is believed that mutations in highly conserved regions lead more likely to harmful changes in the phenotype than mutations in non-conserved regions [Miller and Kumar, 2001]. Natural selection targets the conserved regions by eliminating harmful phenotypes, hence, it is also believed that conserved regions have an important role in gene function. Both coding and non-coding regions can be conserved. Approximately 5 % of the non-coding DNA sequences are highly conserved [The ENCODE Project Consortium, 2007] , for example TATA-promoter sequence in most eukaryotes.

CTCF refers to regions where transcriptional regulator CTCF binds and as an annotation, it is used to refer open chromatin sites that lacks signals of histone modifications and are enriched of CTCF binding sites [The ENCODE Project Consortium, 2012]. CTCF is a 11-zinc finger protein, which can bind either to silencer or to insulator. CTCF takes part in gene expression regulation, organization of the genome and V(D)J recombination. They are located usually next to transcription factor-coding genes [Martin et al., 2011].

CTCF sites are evolutionary conserved and found to be quite invariant in different cell types [The ENCODE Project Consortium, 2012]. In human cells, part of CTCF sites are constitutive, which means that an enzyme product is continuously produced regardless of the need of the gene product.

Digital genomic footprint (DGF) refers those regions of the DHS, that are protected from the cleavage of the DNase I enzyme. Because of being often transcription factor (TF) -occupied, DGFs can be used for identifying those cis-regulatory elements where specific TFs bind. DGFs have been found to have different evolutionary conservation patterns compared to surrounding DNA [Boyle et al., 2011].

DNase I hypersensitivity sites (DHS) refer to open chromatin regions in the genome that make DNA accessible for the transcription factor binding. They are sensitive to cleavage of the DNase I enzyme, which is an endonuclease that cleaves phosphodiester bonds within polynucleotide chain in DNA leaving a free hydroxyl group. DNase I enzyme cuts DNA non-specifically. DHSs are markers for different histone modifications, TSSs and transcription factor binding sites (TFBSs) which are used for identifying regulatory elements; promoters, enhancers, silencers and insulators [Zhang and Pugh, 2011]. Insulators are blocks between enhancer and promoter. 57% of noncoding GWAS hits are found within DHSs [Maurano et al., 2012].

Enhancer refers to regions that participate in gene expression regulation by enhancing a target gene activity. They control both time- and cell-type-specific gene expression and initiate RNA pol II transcription. Enhancers are cis-regulatory elements and can be grouped to proximal and distal enhancers. They can be located upstream, downstream or within the gene they affect. Chromatin marks, TFBSs and DHSs can be used to identify enhancers in the genome [Bulger and Groudine, 2010]. Enhancers produce noncoding enhancer RNA, eRNA, which can be divided into two subclasses by their transcription direction: to an unidirectional 1D eRNA or to a bidirectional 2D eRNA. 1D eRNAs are usually long and polyadenylated with lower H3K4me1/me3 ratio in their enhancers chromatin signature. 2D eRNas are shorter and non-polyadenylated, with higher ratio of H3K4me1/me3 in their enhancer chromatin signature. Levels of 2D eRNAs are strongly related to enhancer activity. eRNAs are actively degraded by exosomes.

FANTOM5 enhancer refers to enhancers that produce 2D eRNA, and which are

identified by using a cap analysis of gene expression (CAGE) in FANTOM5 panel of samples. They are cell-type-specific and bidirectional CAGE pairs are robust predictor for an activity of cell-type-specific enhancers [Bulger and Groudine, 2010], which make them useful for the recognition of cell-type-specific gene expression.

Fetal DHS refers to DHSs that are identified from the cells of a fetus.

Five prime untranslated region (5'UTR) is a region at the start of coding DNA which is not translated into the protein. It has crucial role in translation, because it contains a ribosome binding site where a translation initiator binds to. 5'UTR contains also several regulatory elements, such as enhancers and silencers, and can also contain introns. The average length of 5'UTR varies between 100-200 base pairs (bp) [Mignone et al., 2002].

Introns are located within the gene coding region between exons, but do not code for a protein. They include cis-regulatory elements; intronic enhancers and silencers, which participate in the regulation of alternative splicing. Alternative splicing enables the production of many different proteins from one gene.

Promoter is the region to which transcription factors, that initiate the transcription, bind. One way to classify promoters is to divide them into proximal and distal promoters by their location in relation to coding region. They can also be classified as TATA-box containing promoters and TATA-less promoters, which can be distinguished by their differing histone modification patterns. Focused promoters have one specific transcription starting site (TSS) and produce only a single type of transcripts. Dispersed promoters have several weaker TSSs and are able to produce different transcripts. Focused promoters usually associate with highly regulated genes and dispersed promoters with constitutive genes. Dispersed promoters are usually located in CpG-rich regions called CpG islands, which are usually heavily methylated and participate in epigenetic regulation of gene expression. Unmethylated promoters are usually active, and methylated promoters repressed [The ENCODE Project Consortium, 2012]. Promoters TF-binding sites (TFBSs) differ from their enrichment of methylation [Hoffman et al., 2013]. The minimal part of the promoter that are required for transcription initiation, is called core promoter. Core promoters include a TSS, a general TFBS and a RNA polymerase binding site.

Promoter flanking refers to region adjacent to TSS in the 5'end, which includes pro-

moter and can also include proximal enhancers and other protein binding sites (silencers, insulators and repressors).

Repressed refers to chromatin regions that are predicted to be either inactive or actively repressed. Actively repressed regions are H3K27me3 polycomb-enriched and also enriched for repressor proteins RE1-Silencing Transcription factors (REST), which are encoded from the gene REST. Regions without signal or with low observed signal in segmentation input assays are classified as an inactive, quiescent chromatin [The ENCODE Project Consortium, 2012].

Super enhancer refers to the cluster of many closely spaced highly active enhancers with unusually high level of activator binding or histone modifications. In ChIP-Seq measurements, they show enrichment for binding of mediator, which is a transcriptional coactivator. Super enhancers are often found in regions near genes with cell-type-specific functions [Hnisz et al., 2013] and probably point out enhancers which regulate tissue-type-specific transcription [Pott and Lieb, 2015]. The sequence composition is different between constituent enhancers of super enhancers and normal enhancers. Also, if the individual enhancers part of the super enhancers are taken outside of their context, they have more powerful activating capacity than normal enhancers [Pott and Lieb, 2015].

Three prime untranslated region (3'UTR) is a region at the end of coding DNA sequence and starts straight after termination sequence. It can include several different regulatory elements such as TSSs, enhancers and silencers. The average length of 3'UTR in humans is approximately 800 bp. [Mignone et al., 2002]

Transcribed refers to either regions of protein-coding or non-coding RNA or pseudogenes that are transcribed. Pseudogenes are copies of protein-coding genes that have lost their gene expression ability due to accumulation of inactivating mutations. Transcribed regions are enriched for polyadenylated RNA and for the signals of elongating polymerase Pol II in a phosphorylated form. Transcribed states are significantly cell-type-specific. [The ENCODE Project Consortium, 2012].

Transcription factor binding sites (TFBSs) are regions in which transcription factors bind and are important part of gene expression regulation. TFBSs include all cis-regulatory elements; enhancers, promoters, silencers and insulators, and can be used

to identify them.

Transcription start site (TSS) is a region to which a RNA polymerase binds and starts the transcription. As an annotation, TSS state refers to the predicted promoter region which includes TSS. They are often found within 100 bp from the 5'end, but can be found also within exons and 3'UTR regions. TSSs are usually H3K4me3-enriched and produce a high amount of short RNAs. Most of the histone modification signals are found around TSS states, where the nucleosome usage is unequal [The ENCODE Project Consortium, 2012]. TSSs are often identified by mapping the location of the nucleotide in RNA where the 5'cap is added and are used for identifying the regulation regions [Kapranov, 2009]. Strong correlation between TSS gene expression levels and the presence of distal functional elements such as enhancers in interacting loci pair has been observed, which indicate that there is interaction between distinct chromosomes [The ENCODE Project Consortium, 2012].

Weak enhancer refers to regions of predicted enhancers with low expression level in the close gene, or to some other cis-regulatory element of open chromatin with weak enrichment and signals [The ENCODE Project Consortium, 2012].

Histone marks

Histone modification patterns are cell-type-specific and can be used as markers when studying gene expression regulation. Histone tails provide potential targets along a chromatin fiber for a variety of chemical modifications; acetylation, methylation and phosphorylation. In acetylation, an acetyl group is added to the positively charged amino group on the side chain of lysine, which changes the net charge of the protein by neutralizing the charge. High levels of acetylation open up chromatin fiber. Acetylated histone marks increase in regions of active genes and decrease in inactive regions. In methylation, a methyl group is added either to arginine or lysine residues of histones. Methylation can have a positive or negative impact on gene activity and the degree of methylation in genes correlates with transcriptional activity. Methylation in promoters is often associated with repression but in some cases can refer also to active promoters. Phosphorylation has a role in the regulation of alternative splicing [Stamm, 2008].Histone marks participate in alternative splicing at chromatin level by influencing the recruitment of splicing regulators [Luco et al., 2010]

- **H3K27ac(Hnisz)** and **H3K27ac(PCG2)** two versions of acetylation of histone 3 at lysine 27 are enriched at active enhancers.
- **H3K9ac** acetylation of histone 3 at lysine 9 is enriched at promoters and enhancers.
- **H3K4me1** monomethylation of histone 3 at lysine 4 is enriched at active enhancers.
- **H3K4me3** trimethylation of histone 3 at lysine 4 is enriched at actively transcribed promoters.

Appendix C

LD Score regression by R

```
##############################################################################
## LD Score regression (LDSC)
## Iteratively re-weighted least squares
##  with block jackknife
## precision is set to 4 iteration (as used in the LDSC v 1.0.0)
## Two-step estimator
## step 1: Z^2 < 30, to obtain estimate for the ldsc intercept and se
## step 2: all, with constrained intercept from step_1,
## to obtain h2 estimate and se
##############################################################################

# regression weights
# hsq = h2 estimate, ld = ld scores (vector dim = #SNPs ), w_ld = weight ld scores (vector dim = #SNPs),
# N = GWAS sample size, M = number of common variants used in the ld score estimation, int = intercept
WEIGHTS <- function( hsq, ld, w_ld, N, M, int ){
  h <- max(hsq,0.0)
  h <- min(h,1.0)
  ld <- pmax(ld, 1.0)
  w_ld <- pmax(w_ld, 1.0)
  c <- h*N/M
  # heteroscedasticity weights
  het_w <- 1/(2*(int+c*ld)^2)
  # overcounting weights
  oc_w <- 1/w_ld
  w <- het_w*oc_w
  return( w )
}

# weight x by normalized weights
# w = weights ( vector dim= #SNPs) , x (matrix dim= #SNPs, p)
WEIGHT <- function( w, x ){
  w1 <- sqrt(w)
  w_norm <- w1/sum(w1)
  x_w <- x*w_norm
  return( x_w )
}

# estimate of h2,  regression slope scaled by M and N
HSQ_UP <- function( coef, m, n) ( m * coef / n)

##############################################################################
## DATA input
## LDS = LD Scores (at least columns SNP, L2 )
## M = number of common variants used in the LD Score estimation
## ( from ld score files -.l2.M_5_50)
## X2 = GWAS summary statistics, ( z- scores, at least columns SNP, Z, N )
##############################################################################
```

```
LDS <- fread( data1 ,  data.table=F )
X2 <- fread( data2 , data.table=F )
# Common variants used in LD Score estimation (maf > 0.05)
M <- sum(NUM_5_50)
# merge GWAS summary statistics and LD Scores
B2 <- merge(LDS, X2, by="SNP", sort=F)

# chi^2 -statistics ( z-score^2 )
B2$Z2 <- B2$Z^2
chi2 <- B2$Z2
# ld scores
ld <- B2$L2
# GWAS sample size
N <- B2[1,"N"]

################################################################################
## Crude estimate for initial weights
##  h2est = (M* (mean(z-score^2)-1)) / (mean(N* ldscores))
################################################################################

hsq1 <- (M*(mean(chi2)-1)) / (mean(N*ld))

#################################################################
## STEP 1
##  only variants with max z^2 30 to obtain intercept and se
##  intercept at start 1, updated at each iteration
#################################################################

# subset with z<30
B2_30<-subset(B2, Z2<30)
# ld scores
x1<-round(as.matrix(B2_30$L2),4)
# add intercept
x1_int<-cbind(x1, c=(rep(1,nrow(B2_30))))
# z-scores
y1<-round(as.matrix(B2_30$Z2),4)

## iterated weights
hsq_up<-hsq1
reg_int <- 1
precision <- 4

for(i in 1:precision){
  initial_w1<- as.vector(WEIGHTS( hsq_up, x1 , x1, N, M, reg_int))
  xw1 <- WEIGHT( initial_w1, x1_int)
  yw1 <- WEIGHT( initial_w1, y1)
  coefw<-summary(lm(yw1~0+xw1))$coef
  reg_coef<-coefw[1,1]
  reg_int<-coefw[2,1]
  hsq_up <- HSQ_UP( reg_coef, M, N)
}

## least squares with block jackknife
# number of snps with z2 < 30
n1<-nrow(B2_30)
## blocks values xTx and xTy
V<-(seq(0,n1, length.out = 201))
n_blocks = 200
xtx_block_values <- list()
for (i in 1:n_blocks) {
  xtx_block_values[[i]] <- matrix(0,2,2)
}
xty_block_values <- matrix(0,n_blocks,2)

for (i in 1:n_blocks){
  xtx_block_values[[i]]<-t(xw1[V[i]:(V[i+1]-1),])%*%xw1[V[i]:(V[i+1]-1),]
  xty_block_values[i,]<-t((xw1[V[i]:(V[i+1]-1),]))%*%(yw1[V[i]:(V[i+1]-1),])
}

## convert block values to estimate (solve xtx,xty )
xty <- colSums(xty_block_values)
xtx <- xtx_block_values[[1]]
for (i in 2: length(xtx_block_values)){
  xtx <- xtx+xtx_block_values[[i]]
}
```

```
est<-t(as.matrix(solve(xtx,xty)))

## block values to delete values
delete_values<-matrix(0, n_blocks, 2)
xty_tot<-colSums(xty_block_values)
xtx_tot <- xtx_block_values[[1]]
for (i in 2: length(xtx_block_values)){
  xtx_tot <-xtx_tot+xtx_block_values[[i]]
}
for (i in 1:n_blocks){
  delete_xty <- xty_tot-xty_block_values[i,]
  delete_xtx <- xtx_tot-xtx_block_values[[i]]
  delete_values[i,]<-solve(delete_xtx,delete_xty)
}
# STEP 1 delete_values
delete_values_step1 <- delete_values

## delete values to pseudovalues
# regression slope estimates
pseudovalues1<-n_blocks*est[1,1]-(n_blocks-1)*delete_values[,1]
# intercept
pseudovalues2<-n_blocks*est[1,2]-(n_blocks-1)*delete_values[,2]
# combine
pseudovalues_step1<-cbind(pseudovalues1,pseudovalues2)

## Jackknife estimates
jknife_cov<-cov(pseudovalues_step1)/n_blocks
jknife_var <- diag(jknife_cov)
jknife_se<- sqrt(jknife_var)
jknife_est<-t(as.matrix(colMeans(pseudovalues_step1)))

step1_int <- jknife_est[1,2]
step1_int_se <- jknife_se[2]

#####################################################
## STEP 2
## constrain intercept to step 1 intercept
##  - h2 estimate  and se
#####################################################

# Ld scores as a matrix
x<-round(as.matrix(ld),4)
# chi^2 as a matrix
y<-round(as.matrix(chi2),4)
# reduce intercept from y
yp<-y-step1_int

## iterated weights
hsq_up <- hsq1
reg_int <- 1

# for step 2, at start initial weights with crude h2 estimate and intercept 1,
# from second round intercept is step 1 intercept
for(i in 1: precision){
   initial_w2<- as.vector(WEIGHTS( hsq_up, x , x, N, M, reg_int))
   xw2 <- WEIGHT( initial_w2, x)
   yw2 <- WEIGHT( initial_w2, yp)
   coefw2<-summary(lm(yw2~0+xw2))$coef
   reg_coef2<-coefw2[1,1]
   reg_int <- step1_int
   hsq_up <- HSQ_UP( reg_coef2, M, N)
}

##  least squares with block jackknife
n1 <- nrow(x)
V<-(seq(0,n1, length.out = 201))
n_blocks = 200
xtx_block_values <- list()
for (i in 1:n_blocks) {
  xtx_block_values[[i]] <- matrix(0,1,1)
}
xty_block_values <- matrix(0,n_blocks,1)
for (i in 1:n_blocks){
```

```
  xtx_block_values[[i]]<-t(xw2[V[i]:(V[i+1]-1),])%*%xw2[V[i]:(V[i+1]-1),]
  xty_block_values[i,]<-t(t((xw2[V[i]:(V[i+1]-1),]))%*%(yw2[V[i]:(V[i+1]-1),]))
}

## block values to est
xty<-colSums(xty_block_values)
xtx <- xtx_block_values[[1]]
for (i in 2: length(xtx_block_values)){
  xtx <- xtx+xtx_block_values[[i]]
}
est2<-t(as.matrix(solve(xtx,xty)))

## block values to delete values
delete_values<-matrix(0, n_blocks, 1)
xty_tot<-colSums(xty_block_values)
xtx_tot <- xtx_block_values[[1]]
for (i in 2: length(xtx_block_values)){
  xtx_tot <-xtx_tot+xtx_block_values[[i]]
}
for (i in 1:n_blocks){
  delete_xty <- xty_tot-xty_block_values[i,]
  delete_xtx <- xtx_tot-xtx_block_values[[i]]
  delete_values[i,]<-solve(delete_xtx,delete_xty)
}

## delete values to pseudovalues
delete_values_step2 <- delete_values
pseudovalues1<-n_blocks*est2[1,1]-(n_blocks-1)*delete_values[,1]
pseudovalues_step2 <- as.matrix(pseudovalues1)

## Jackknife estimates
jknife_cov2<-cov(pseudovalues_step2)/n_blocks
jknife_var2 <- diag(jknife_cov2)
jknife_se2<- sqrt(jknife_var2)
jknife_est2<-t(as.matrix(colMeans(pseudovalues_step2)))

################################
## FINAL estimates
################################

# intercept and se (STEP 1)
int <-step1_int
int_se <- step1_int_se
# h2 estimate and se (STEP 2)
h2_est <- HSQ_UP(jknife_est2, M, N))
h2_est_se <- HSQ_UP(jknife_se2, M, N)
```

Appendix D

Abbreviations

- cM centimorgan
- CNS central nervous system
- DGF digital genomic footprint
- DHS DNase I hypersensitivity sites
- DZ dizygotic
- FDR false discovery rate
- GI gastrointestinal
- GC genomic control
- GRM genetic-relationship matrix
- GWAS genome-wide association study
- GWS genome-wide significant
- HDL-C high-density lipoprotein cholesterol
- LD linkage disequilibrium
- LDL-C low-density lipoprotein cholesterol
- LDSC LD Score regression

- LMM linear mixed model
- MAF minor allele frequency
- MHC major histocompatibility complex
- MZ monozygotic
- OLS ordinary least squares
- PC principal component
- REML restricted maximum-likelihood
- SNP single nucleotide polymorphism
- SSE sum of squared residuals
- SSR sum of squares due to regression
- SST total sum of squares
- S-LDSC stratified LD Score regression
- THL National Institute for Health and Welfare
- TC total cholesterol
- TFBS transcription factor binding site
- TG triglycerides
- TSS transcription start site
- WLS weighted least squares