ESTABLISHING THE GENETIC ETIOLOGY IN COMMON HUMAN PHENOTYPES

SIM XUELING

(BSc Hons, National University of Singapore)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF EPIDEMIOLOGY AND PUBLIC HEALTH

NATIONAL UNIVERSITY OF SINGAPORE

2012

ACKNOWLEDGEMENTS

This thesis and all the work over the last 6 years would not have been possible without the love and support of everyone who has stood behind me all the way. I would like to thank them here:

My parents and brother who showed unwavering support for my career choice, always making sure I have fruits for breakfast and hot meals when I get home. Small gestures in life that speak of boundless love.

Prof Chia Kee Seng. An Honors year project that led to six years of training and grooming. The work trips where I get to travel, work, learn (and play), all in one. Planning every step of my career, he is the superman boss whom I can always count on.

A/P Tai E Shyong and A/P Teo Yik Ying. My co-supervisors. I know them within months of each other. I had the luxury of learning from them when they were a lot less busy. YY would spend hours with me on MSN, explaining the concepts of GWAS to me via long distance. E Shyong would spend hours sitting with me, learning together and most importantly, making sure that I know what I am doing. E Shyong showed me the value of communicating with people and is never too busy to spare me a few minutes when I need it. YY, a superb teacher, whose patience I have seen nowhere. His drive to see projects to publications will be my motivation.

Prof Wong Tien Yin. E Shyong brought me into your world of ophthalmology and for the opportunities you have given me over the years, I really appreciate them. Working with you also led me to new-found friends.

Sharon, Gek Hsiang, Chuen Seng and Kaavya. My comrades in fun, laughter and gossips. I will always remember the time we had in GIS together. The fun, the laughter, the talking stick and the statistical pig (or hippo?). They made me realize the importance of moral support when working together and we click as well as ever, regardless of how long or how far apart we are. Thanks to Chuen Seng too, for proof-reading this thesis.

Rick, Adrian, Erwin and Jieming. These guys have never turned me away when I have problems with work. From them, I learned to live in the Linux world and the importance of programming.

Hazrin, who is always there with his IT support and taking care of the server (without it, none of this work can materialize) with me.

My colleagues in CME and everyone in EPH. All the academic staff who had provided guidance in lectures work, or even shared life lessons along the way. The non-academic staff who has helped me in one way or another, be it IT-related or administrative matters.

None of this work would have been possible without the participants of these studies and the people who run the recruitment, logistics and management of these studies.

To those whom I have missed out, my heartfelt thanks.

TABLE OF CONTENTS

SUMMARY

It has been increasingly valuable to look across populations of different ancestries, taking advantage of the allelic frequency and linkage disequilibrium differences that could shed more light on the genetic architecture of common diseases and complex traits. Singapore is a small country state at the tip of the Malaysia Peninsula, home to a population of 5 million. The unique demographic makeup of the three main ethnic groups, Chinese, Malays and Asian Indians, captures much of the genetic diversity across Asia. We first assembled a resource of 100 individuals from each of the three ethnic groups, with the aim of comparing their genetic diversity within ethnic groups and also with existing HapMap populations to determine if this genetic diversity might have implications for genetic association studies. The multi-ethnic demographic characteristic allowed us to investigate various aims: (i) to identify disease susceptibility genetic loci common to multiple ethnic groups; (ii) to assess the impact of allele frequencies differences and allelic heterogeneity on the transferability of European loci to non-Europeans; (iii) to identify population specific disease implicated loci in genetic association studies. In particular, we will describe findings from a Type 2 Diabetes genome-wide association study that highlight the transferability and consistency of established Type 2 Diabetes loci from European populations to Asian populations. Through meta-analysis with other South Asian populations, we report six new loci implicated in Type 2 Diabetes in South Asian Indians. Finally, using the same ethnic groups, we demonstrate that re-defining phenotype has an important role in improving existing knowledge of disease pathogenesis and complementing our physiological understanding of genetic susceptible variants.

LIST OF TABLES

LIST OF FIGURES

PUBLICATIONS

This thesis is based on the following publications:

1. Teo YY[*], **Sim X**[*], Ong RTH[*], Tan AKS, Chen JM, Tantoso E, Small KS, Ku CS, Lee EJD, Seielstad M and Chia KS. Singapore Genome Variation Project: A Haplotype map of three South-East Asian populations. Genome Res. 2009 Nov;19(11):2154-62. Epub 2009 Aug 21.
    a. Contributed to the analyses, manuscript writing and design of the website.
2. **Sim X**, Ong RT, Suo C, Tay WT, Liu J, Ng DP, Boehnke M, Chia KS, Wong TY, Seielstad M, Teo YY, Tai ES. Transferability of Type 2 Diabetes Implicated Loci in Multi-Ethnic Cohorts from Southeast Asia. PLoS Genet. 2011 Apr;7(4):e1001363. Epub 2011 Apr 7.
    a. Conducted the analyses and wrote the paper with Teo YY and Tai ES.
3. Kooner JS[*], Saleheen D[*], **Sim X**[*], Sehmi J[*], Zhang W[*], Frossard P[*], Been LF, Chia KS, Dimas AS, Hassanali N, Jafar T, Jowett JB, Li X, Radha V, Rees SD, Takeuchi F, Young R, Aung T, Basit A, Chidambaram M, Das D, Grunberg E, Hedman AK, Hydrie ZI, Islam M, Khor CC, Kowlessur S, Kristensen MM, Liju S, Lim WY, Matthews DR, Liu J, Morris AP, Nica AC, Pinidiyapathirage JM, Prokopenko I, Rasheed A, Samuel M, Shah N, Shera AS, Small KS, Suo C, Wickremasinghe AR, Wong TY, Yang M, Zhang F; DIAGRAM; MuTHER, Abecasis GR, Barnett AH, Caulfield M, Deloukas P, Frayling TM, Froguel P, Kato N, Katulanda P, Kelly MA, Liang J, Mohan V, Sanghera DK, Scott J, Seielstad M, Zimmet PZ, Elliott P[*], Teo YY[*], McCarthy MI[*], Danesh J[*], Tai ES[*], Chambers JC[*]. Genome-wide association study in individuals of South Asian ancestry identifies six new type 2 diabetes susceptibility loci. Nat Genet. 2011 Aug 28. doi: 10.1038/ng.921. [Epub ahead of print]
    a. Conducted the analyses for Singapore cohorts (discovery and replication cohorts), carried out meta-analysis in parallel with collaborators at Imperial College. Participated in the manuscript preparations and writing.

These papers also provided important background and relevant to the work of this thesis.

1. Teo YY, Fry AE, Bhattacharya K, Small KS, Kwiatkowski DP, Clark TG. Genome-wide comparisons of variation in linkage disequilibrium. Genome Res. 2009 Oct;19(10):1849-60. Epub 2009 Jun 18.
2. Teo YY, **Sim X**. Patterns of linkage disequilibrium in different populations: implications and opportunities for lipid-associated loci identified from genome-wide association studies. Curr Opin Lipidol. 2010 Apr;21(2):104-15.
3. Kato N[*], Takeuchi F[*], Tabara Y[*], Kelly TN[*], Go MJ[*], **Sim X**[*], Tay WT[*], Chen CH[*], Zhang Y[*], Yamamoto K[*], Katsuya T[*], Yokota M[*], Kim YJ, Ong RT, Nabika T, Gu D, Chang LC, Kokubo Y, Huang W, Ohnaka K, Yamori Y, Nakashima E, Jaquish CE, Lee JY, Seielstad M, Isono M, Hixson JE, Chen YT, Miki T, Zhou X, Sugiyama T, Jeon JP, Liu JJ, Takayanagi R, Kim SS, Aung T, Sung YJ, Zhang X, Wong TY, Han BG, Kobayashi S, Ogihara T[*], Zhu D[*], Iwai N[*], Wu JY[*], Teo YY[*], Tai ES[*], Cho YS[*], He J[*]. Meta-analysis of genome-wide association studies identifies common variants associated with blood pressure variation in east Asians. Nat Genet. 2011 Jun;43(6):531-8. Epub 2011 May 15.

[*] Joint first/last authors

CHAPTER 1 – INTRODUCTION

1.1. Mendelian Genetics and Inheritance

The evolution of modern genetics has seen the greatest change in the last decade. In 1865, Gregor Johann Mendel, the father of modern genetics, established Mendel's law of segregation (two copies of alleles separate during gamete formation such that each gamete only receives one copy. Offsprings then randomly inherit one gamete from each parent during transmission) and law of random assortment (two different genes randomly assort their alleles to be inherited independently). Mendelian inheritance models are typically characterized by single molecular defects (monogenic) segregating within families, such as cystic fibrosis which has an autosomal recessive inheritance pattern[1]. However, it soon became clear that there could be extensive phenotypic variation in these disorders, even in the presence of similar molecular patterns due to variable penetrance[2].

At the same time, the patterns of inheritance for common quantitative traits such as anthropometric measures and complex diseases like Type 2 Diabetes within families were not conforming to Mendelian laws but rather in a blending fashion from the parents. In 1918, R. A. Fisher demonstrated that individual differences observed at a particular trait could be attributable to genetic variations at more than one locus and that inter-individual differences are as a consequence of the collective effects from all contributing loci[3,4]. Traits of this nature were later termed as polygeneic, multifactorial or complex traits. The understanding of these models of inheritance shaped the development of methods for the discovery of common diseases or complex traits.

1.2. Candidate Gene Studies and Linkage Scans

Earlier studies of gene mapping to compare the inheritance patterns of complex traits were limited by our knowledge of the genome and the ease of detecting genetic variants. The candidate gene approach relied on prior biological knowledge to decide on the choice of target region, often based on specific hypothesis on the pathogenesis of disease. This type of study, limited by the lack of knowledge of the human genome to make informed selection of candidate regions and the small sample sizes of the experiments, often yielded irreproducible results. Despite these challenges, the candidate gene approach does have its success in Type 2 Diabetes. For example, the peroxisome proliferator-activated receptor gamma (*PPARG*)[5] and potassium inwardly-rectifying channel, subfamily J, member 11 (*KCNJ11*)[6] harbor common variants associated with Type 2 Diabetes in a highly reproducible manner. Both are drugs targets used to treat Type 2 Diabetes. They are implicated in rare monogenic syndromes characterized by severe metabolic disturbance of beta-cell function and insulin resistance[7,8].

Linkage studies leverage on the genetic markers segregating with disease alleles in affected families. Of note, the variant with the strongest effect on Type 2 Diabetes on chromosome 10 to date was discovered via linkage analysis[9] and a search for microsatellite association localized the variant to an intron within the transcription factor 7-like 2 gene (*TCF7L2*)[10,11]. The index variant replicated across multiple European populations and had an odds ratio of 1.40 (95% CI: 1.34 – 1.46)[12] in developing Type 2 Diabetes. Unfortunately, linkage has low power and resolution for variants with modest effects. In 1996, Risch and Merikangas suggested that for a disease risk of 1.5 and risk allele frequency of 0.10, the number of families required for 80% power using affected siblings design was close to 70,000[13]. On the contrary, for the same disease risk and risk allele frequency, the number of sibling pairs required for association analysis was a little under 1,000. Association studies, by design, compare the frequencies of alleles or genotypes of variants

between disease cases and controls in its simplest form, thus providing a simpler and more practical way of identifying disease implicated variants in complex traits.

1.3. Genome-Wide Association Study (GWAS)

The genomes of any two individuals are about 99.9% identical. The remaining 0.1% of genetic differences can be largely attributable to: (i) single nucleotide polymorphism (SNP), which represent single base change between individuals; and (ii) structural variants comprising of genomic alterations such as copy number polymorphisms, insertions, deletions and duplications[14]. While a comprehensive direct search for genetic determinants of disease would involve examining all genetic differences in substantially large number of affected and unaffected individuals through whole genome sequencing, this is currently not feasible with the high cost of sequencing in large studies.

The genetic architecture of diseases involves understanding how many susceptible genetic variants are involved, the risk allele frequencies at these variants and the magnitudes of the effects these risk alleles have on diseases. There have been two major views on the allelic spectra of variants affecting multi-factorial diseases[15,16]. The first being the common disease common variant (CDCV) hypothesis, that common diseases are attributed to the joint action of common genetic variants (minor allele frequency MAF at least 5%) which individually are likely to contribute marginally to the disease. On the other hand, the rare variant hypothesis proposes that disease incidences might be due to less common variants (MAF of less than 0.01) that are distinct in different individuals.

Genome-wide association studies adopt a hypothesis-free approach to identify genetic variants associated with complex traits with the common disease common variant approach as the

underlying model of allelic spectrum of diseases. It is an indirect approach to screen the genome where a set of well chosen variants, specifically SNPs, could serve as genetic markers to detect association between regions of the genome and the phenotype of interest, by making use of the inherent correlation between genetic variants along a chromosome. The SNPs queried are believed to be rarely the causal variants (variants that are biological functional or responsible for expressing the phenotype of interest) but instead are sufficiently correlated with the causal variants to show an association with the trait.

The unbiased approach of surveying the genome for disease implicated loci has been made possible with several crucial developments, including deeper understanding of linkage disequilibrium across the genome, the catalog of common genetic variation across four populations by the International HapMap Project[14,17,18] and technological advancement in the genotyping field. Most genome-wide association studies rely on commercial genotyping arrays from two major companies, Affymetrix (Santa Clara, California, United States of America, http://www.affymetrix.com/estore/) and Illumina (San Diego, California, United States of America, http://www.illumina.com/). Since the first genome-wide scan published in 2005 that discovered an association between the complementary H polymorphism (*CFH*) in 96 age-related macular degeneration cases and 50 controls[19], there has been a plethora of genome-wide association studies on chronic diseases Type 2 Diabetes, inflammatory disorders, infectious diseases, cancers and quantitative traits such as height and body mass index[20,21]. These will be discussed in greater details in the following sections.

1.3.1. Linkage disequilibrium and recombination in the human genome

Linkage disequilibrium (LD) reflects the shared ancestry of genetic variation in populations[22]. When new mutation arises, it is initially linked to the other alleles on the same chromosome. The

unique combination of alleles on a chromosome is called a haplotype and the non-random correlation of alleles on these haplotypes results in linkage disequilibrium.

Linkage disequilibrium is a balance between several population genetic forces including genetic drift, population structure, natural selection and recombination. Briefly, contrary to Mendelian law of independent assortment, genetic material close on the same chromosome are not passed down independently and thus correlation structures within populations tend to be more similar due to shared evolutionary history[23]. Genetic drift results in a change in the allele frequency due to random sampling as genetic materials are passed down from parents to offsprings. Natural selection is another evolutionary force favoring mutations that increase survival and reproduction (positive selection) while eliminating deleterious mutations that decrease survival and reproduction (negative selection). These population genetic forces influence the linkage disequilibrium within populations, generally inflating linkage disequilibrium. In the absence of recombination, genetic diversity arises solely through mutation. Recombination is the re-shuffling of genetic material between the paternal and maternal chromosomes at a specific location of the chromosome during meiosis. This process results in the unlinking of materials on the parental chromosomes and new chromosomes that are eventually transmitted contain new combinations of genetic materials from both parents. Genetic diversity is increased as this process allows genetic materials from all four grandparents to be passed down to the offsprings. The genetic materials that are passed down from the parents to offsprings will be different from what is passed down to the parents from the grandparents, thus breaking down linkage disequilibrium.

Linkage disequilibrium varies markedly across the genome and between populations of different ancestry. Using SNP data in 44 individuals from Utah from the Centre d'Etude du Polymorphisme Humain collection (CEPH) and 96 Yorubans from Nigeria in 19 regions of the

genome, Reich et al showed that linkage disequilibrium extends over longer distance compared to previous predictions from demographic models and decreases as a function of physical distance between SNPs[24]. Linkage disequilibrium patterns are closely related to recombination. Long stretches of linkage disequilibrium are often characterized by recombination hotspots (regions in the genome with elevated rates of recombination) at the ends, creating blocks of haplotypes where only a few common haplotypes are observed with little evidence of recombination within the block[25-28]. The presence of long stretches of linkage disequilibrium and haplotype blocks allows a small set of well-chosen SNPs to act as efficient tagging surrogates of other SNPs or haplotypes[29,30], thus reducing the number of SNPs to be queried and to provide a high degree of genome coverage. The selection of markers therefore depends on the strength of linkage disequilibrium between markers.

Several measures of linkage disequilibrium are commonly used, including the Lewontin's D'[31,32] and genetic correlation coefficient r^2 [33]. Consider two biallelic SNPs, with the alleles (A, a) on one locus and alleles (B, b) on the other locus. Let f_x denotes the frequency of the x allele and f_{xy} denotes the haplotype frequencies of the xy haplotype:

$$Lewontin\ D' = \begin{cases} \dfrac{f_{AB} - f_A f_B}{\min(f_A f_b,\ f_a f_B)} & if\ f_{AB} - f_A f_B > 0 \\ \dfrac{f_{AB} - f_A f_B}{\min(f_A f_B,\ f_a f_b)} & if\ f_{AB} - f_A f_B < 0 \end{cases} \quad and \quad r^2 = \frac{(f_{AB} - f_A f_B)^2}{f_A f_B f_a f_b}$$

From the numerator in D' and r^2, if there is no linkage disequilibrium (i.e. linkage equilibrium), then the observed haplotype frequency at the two SNPs should be equal to the expected haplotype frequency obtained from the product of allele frequencies at the two SNPs. D' can be interpreted as the number of differentiated haplotypes and is less than one if and only if all four haplotypes are observed. r^2 is a measure of how much information one SNP contains for a second SNP. An r^2

of one indicates that one variant is a perfect surrogate of the other while r^2 of zero means that the two variants provide no information about each other. Correlations between SNPs r^2 depends on the historical order and genealogy branches in which they arose while D' measures evidence of historical recombination. Thus knowledge of linkage disequilibrium in the genome (in the form of r^2) allows an efficient selection of informative tag SNPs, which act as proxies and provide information about unobserved SNPs, facilitating indirect genome-wide association studies[30].

1.3.2. The International HapMap Project (HapMap)

In order to efficiently select informative markers in the genome, it is important to understand the local linkage disequilibrium patterns in different populations. The International HapMap Consortium was first initiated in 2001 with the aim to catalogue common patterns of genetic variations in samples from populations of African, Asian and European ancestry[14], providing a guide to the design of genetic studies.

The project was carried out in a few phases. In the first phase, genotyping set out to capture at least one common SNP (defined as MAF at least 5%) in every 5 kilobases (kb) across the genome in individuals with African, Asian and European ancestries[17]. Specifically, the samples consisted of 30 Yoruba parent-offspring trios (90 individuals) from the Ibadan region of Nigeria (YRI) of African ancestry, 30 parent-offspring trios (90 individuals) in Utah from the Centre d'Etude du Polymorphisme Humain collection (CEU) of European ancestry, and 45 unrelated Han Chinese from Beijing (CHB) and 44 unrelated Japanese from Tokyo, Japan (JPT) of Asian ancestry[14,17]. This generated approximately one million SNPs that were polymorphic across the samples after stringent quality checks.

Phase II catalogued a further 3.1 million SNPs on the same individuals, capturing approximately 25 – 30% of the common variants in the assembled human genome[18]. At an r^2 threshold of at least 0.8 in common SNPs, only 520,111, 552,853 and 1,092,422 tag SNPs are required as proxies in CEU, JPT+CHB and YRI respectively to the 3.1 million common SNPs that are polymorphic in at least one of the three populations[18]. This provided an invaluable resource to commercial genotyping companies in the design of genome-wide genotyping arrays. Furthermore, the dense and high quality haplotype information from HapMap enabled new study samples to derive *in-silico* genotypes by virtue of haplotype similarity of the study samples with local haplotypic structure from HapMap through statistical imputation methods[18].

As commercial genotyping companies design their genotyping arrays using HapMap, it is essential to know how well the tag SNPs selected from populations of Asian, European and African ancestries capture genetic variations in other populations as it directly affects the power of genetic studies in these populations[34]. The Human Genome Diversity Project (HGDP) performed an initial evaluation of the portability of HapMap haplotypes to 927 unrelated individuals from 52 populations in 36 regions spanning 12Mb[35]. Results indicated substantial haplotype sharing in populations of similar ancestries to those included in HapMap, for instance, the Han and Japanese samples in HGDP had the highest haplotype sharing with HapMap Asians (CHB+JPT). Generally, the HapMap resource can be used to select tags for other populations that are not in HapMap[34]. However, SNP tagging performance varied across populations. Tagging performance is improved if (i) the tag SNPs panel was based on closest HapMap panel as determined by population structure analysis or (ii) the tag SNPs were selected from all four HapMap populations for those populations which are genetically more distinct compared to HapMap[35]. Overall, the transferability of tag SNPs across populations largely depends on the

strength of linkage disequilibrium with the Africans having the lowest portability due to their shorter linkage disequilibrium[24].

The third phase of HapMap extended the study to include additional individuals from the original four populations and seven additional populations to increase genetic diversity, (i) African ancestry in southwestern United States (ASW); (ii) Chinese in Metropolitan Denver, Colorado, United States (CHD); (iii) Gujarati Indians in Houston, Texas, United States (GIH); (iv) Luhya in Webuye, Kenya (LWK); (v) Maasai in Kinyawa, Kenya (MKK); (vi) Mexican ancestry in Los Angeles, California, United States (MXL) and (vii) Tuscans in Italy (Toscani in Italia, TSI)[36]. Genotyping was performed on two commercial genotyping arrays, Genome-Wide Human SNP Array 6.0[37] and Illumina 1M-single bead chip, with quality checks at the individual array level and post merging of the genotype calls from the two arrays.

1.3.3. Advances in genotyping technology and genotype calling

Improving technology and availability of public SNP databases such as the Single Nucleotide Polymorphism Database (dbSNP) and HapMap made it possible to survey up to a million variants for disease association on first generation commercial genotyping arrays from Affymetrix and Illumina, two key players in the industry.

Affymetrix introduced its first genome-wide array, GeneChip Mapping 10K 2.0 Array as part of their suite of robust DNA Analysis products in 2004[38]. Between 2004 and 2009, four more genome-wide SNP arrays were released, namely the Mapping 100K Set, Mapping 500K Array Set, Human SNP Array 5.0 and Genome-wide Human SNP Array 6.0 (http://www.affymetrix.com/estore/). Each SNP on the array is assayed by a number of probe cells containing unique oligonucleotides of defined sequences typically of length 25 bases or

more. These probing sequences will bind to the appropriate target sequences and emit fluorescence at the fluorescent end. The degree of fluorescence yields pixel intensity for each SNP which genotype calling is dependent on. Affymetrix selects probes evenly spaced across the genome[37] and retains redundancy when probes fail in the process of genotyping.

Illumina launched the Infinium Assay in mid 2005, which provided a way to intelligent SNP selection and unlimited access to the genome. The first Infinium product, Human-1 Genotyping BeadChip, assayed over 100,000 markers on a single BeadChip. Subsequently, Illumina introduced Infinium HumanHap300 BeadChip, HumanHap550 BeadChip, HumanHap610 BeadChip, HumanHap650Y, HumanHap660W and Human1M over the next two years (http://www.illumina.com/). These first generation genome-wide arrays generally contained tagged SNPs selected from the HapMap project (CEU). The Infinium workflow includes hybridization of unlabeled DNA fragment to 50-mer probe on the array and enzymatic single base extension with labeled nucleotide, giving rise to red and green intensities[39]. The latest genotyping family of microarrays, the Omni family, features contents from The 1000 Genomes Project (1KGP) which aim to characterize at least 95% of variants in the genome that is accessible to high-throughput sequencing and of allele frequency 1% and above in five major population groups (Europe, East Asia, West Asia, West Africa and the Americas)[40]. This family of next-generation genotyping array allows researchers progressive access to newly discovered variants and eventually aims to release five million marker set on a single BeadChip (Omni5 BeadChip)[41].

Generally, for both Affymetrix and Illumina, probes are designed to target specific regions of the genome. For each possible allele at the genomic position, hybridization of the probes with the samples will generate fluorescence intensities. Genotypes were previously manually determined by examining fluorescent intensities and assigning genotype calls. The scale of such genotyping

experiments involving at least hundred thousand of SNPs and thousands of samples make it impossible to perform genotype calling manually. Thus, there have been immense developments in unsupervised automated genotype calling algorithms for genotype assignments[42-49]. Genotyping calling algorithms evaluate the intensities (typically biallelic) and assign the most probable genotype call based on the highest posterior probabilities of the three genotype classes. The process of genotype assignment is highly dependent on the designated threshold, which is determined differently by each method, and there exists a tradeoff between SNP call rates (the number of samples with a valid call for a SNP) and the designated threshold. A more stringent threshold will likely reduce the number of SNPs with unusual clustering characteristics, resulting in lower call rates.

Ideally, genotype assignment should be visually assessed via clusterplots which are bivariate plots of intensities of the two alleles (Figure 1). As there are at least several hundreds of thousands of SNPs on these arrays, it is not possible to manually curate the continuous hybridization intensities to derive discrete genotype calls for association analyses. This implies that there would be inherent erroneous and missing genotype calls (i.e. the genotype of an individual is not called). Therefore a set of standard quality checks (QC) needs to be performed on the data to minimize false positive associations from these data artifacts in downstream analyses. The common strategy now is to visually assess clusterplots with suggestive signals of association to prevent spurious false positives caused by poor clustering of the intensities.

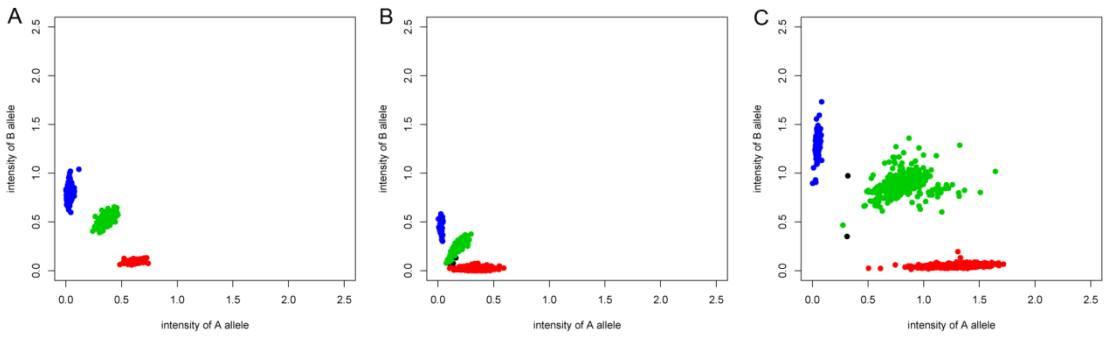


Figure 1. Clusterplots of biallelic hybridization intensities. The axes indicate the continuous hybridization intensities and the points are coloured (blue, green and red) based on their discrete genotype calls, with black indicating missing genotype call. A) A SNP with three distinct clusters, called with high confidence; B) A SNP with overlapping clusters and C) A SNP with a slight shift in the heterozygous cluster.

1.4. Potential for Non European Genome-wide Association Study

The majority of the first wave of genome-wide studies had been centered on populations of European descent[50]. Despite tremendous successes from European genome-wide association studies in identifying disease susceptibility loci, many questions remain to be answered. As the European populations only represent one aspect of human genetic variations, some of the most important questions relate to the relevance of current findings, mainly from populations of European descent, to other populations and the potential of non-European GWAS to detect novel susceptibility genetic variants that are either not present in the Europeans or are at considerably lower frequencies in European populations.

1.4.1. Patterns of LD in Asian ethnic groups

Early GWASs have primarily focused on populations of European descent. First generation genotyping arrays primarily make use of HapMap CEU for SNP selection which relied on the dbSNP database (mainly contained SNPs discovered and ascertained in populations of European descent) for SNPs to include in the genotyping. Thus commercial genotyping array favored

genetic association analyses in populations of European descent, resulting in the inclusion of some SNPs that are polymorphic in populations of European descent but are actually monomorphic in other populations. The availability of the HapMap CEU population of European descent meant that for association studies conducted in European, there is a sufficiently close reference population from which the tag SNPs could be selected from with similar underlying linkage disequilibrium structure.

The HapMap project has documented variations in linkage disequilibrium in global populations such as Africans, Europeans and Asians[17,18,36]. However, there exists substantial heterogeneity in genetic variation within each of these global populations, which is less well documented. For instance, within Asia, while South Asians from the India sub-continent are genetically more similar to the Europeans than Japanese or Chinese, they exhibit much more genetic diversity compared to that observed within Europe[51,52]. This genetic diversity poses challenges in performing association mapping in non-European populations, from limitations in SNP ascertainment of the genotyping array to downstream analyses such as imputation, meta-analysis and replication of association signals.

a. SNP ascertainment bias in first-generation GWAS arrays

SNP ascertainment bias is a phenomenon where there is systematic deviation from population theory due to sampling process in the population and variation in the size of the sampling effort[53]. As the initial efforts for SNP detection and subsequently the design of genotyping arrays were more focused on European populations, SNPs selected for genotyping arrays could have lower allele frequencies in non-European populations, thus compromising the tagging properties of these SNPs and the resultant coverage of the genome in non-European populations. Coverage here is determined by the linkage disequilibrium measure r^2, which translate to the percentage of

SNPs in a HapMap panel with a maximal r^2 of 0.8 with the SNPs on the genotyping array. Low frequency SNPs affect r^2, thus the same tagging SNP might not predict other SNPs as efficiently in non-European populations due to inter-population linkage disequilibrium differences, potentially affecting the ability to detect disease susceptibility locus in these populations.

b. Imputation, meta-analysis and replication

Current genome-wide association analyses typically utilize commercial genotyping arrays with different SNP contents. In order to maximize statistical power, evidences across multiple studies are combined through meta-analyses and any initial discovered variants will be validated in independent populations of the same ancestry and sometimes in different populations.

Imputation infers unobserved genotypes against a common reference panel for association mapping and thus enables meta-analysis to be carried out in multiple studies where different SNPs are assayed using different genotyping arrays by harmonizing the SNP content. It makes use of publicly available dense reference panels and statistical/population genetics methods to infer genotypes that have not been observed on genotyping arrays. The general framework of imputation compares the observed genotypes against a set of dense reference haplotypes (generally sharing a common ancestry and evolutionary history) and subsequently fills in the missing data from the most appropriate reference haplotype[54-58]. These imputation algorithms typically include quantification of the uncertainties in the imputed genotypes, allowing association analyses to properly account for imputation uncertainties.

The accuracy of the imputation method depends on several factors such as the strength of linkage disequilibrium in the population studied and the availability of a dense reference panel genetically similar to the population being imputed[50]. The extent of haplotype sharing is generally greater in

genomic regions with strong linkage disequilibrium, so the imputation can stretch across longer distances[59,60]. Using data from 52 populations around the world (Human Genome Diversity Project HGDP), Huang et. al. evaluated imputation accuracy using the HapMap populations as reference imputation panels[59]. They found imputing against a reference panel derived from a population that was geographically close generally produced higher imputation accuracy. In addition, population specific reference panels optimize imputation accuracy[59,61]. However, it might not be realistic to have sufficiently dense reference panels for all the genome-wide association studies in diverse populations. A mixture panel combining multiple reference panels has been recommended, with the advantage of increased haplotype diversity[59].

With imputation, data can be pooled together in an unbiased manner across the genome to combine evidences across multiple studies in order to boost the effective sample sizes especially in light of small effect sizes in genetic disease association. There are generally two commonly used meta-analysis methods, fixed and random effects modeling. In the context of fixed effect modeling, it is assumed that each individual study estimates a common population effect size. As meta-analysis is performed at individual SNP level, differential linkage disequilibrium patterns with the casual variants will result in different disease susceptibility variants, or index SNPs, emerging from the association analyses. Thus the same index SNP is likely to have different effect sizes across populations and combining evidence at the individual SNP level will mask any real association even though they share the same common causal variant. Multiple causal variants at each locus will also give rise to the same difficulty in detecting real association across populations. As meta-analysis leverages on imputation to augment the observed SNPs from genotyping arrays, imperfect imputation due to absence of appropriate reference panels is also likely to affect the validity of meta-analysis. The random effect model assumes that there is a distribution of population effect sizes around an overall population mean and each individual

study represents a draw from this distribution. Although the method accounts for additional variability between the studies, it is more conservative and tends to down-weigh studies with larger sample sizes, thus less commonly used in meta-analyses of genetic association studies.

Similarly, in replication studies, index SNPs from the discovery phase are often selected to be validated in other populations. This fundamentally assumes that the linkage disequilibrium patterns of the index SNP with causal variants across the discovery and replication populations are similar. Understanding the genetic diversity and inter-population linkage disequilibrium differences is thus vital for interpretation of genetic association studies and lay the foundation for inter-population studies.

1.4.2. Are findings from European studies relevant to other ethnic groups?

Recall that genome-wide association scans make use of indirect association leveraging on linkage disequilibrium. Thus the discovered variants are rarely the functional disease causing variants, but represent variants in sufficient correlation with the functional disease causing variants. Suppose that different populations share a common disease functional variant. The reproducibility of the European discovered implicated index SNPs in other populations depends on several factors: i) the linkage disequilibrium of the index SNPs with the same functional variants in the non-European populations; ii) the allele frequencies of the index SNPs across non-European populations; iii) the effect sizes of the index SNPs across the different populations due to differences in their genetic background or environmental exposures. Certainly, it is possible that there exist multiple causal variants across different populations, either at the same locus (allelic heterogeneity) or specific to particular populations. These factors have a direct impact on the sample sizes required and thus the power to detect the association across populations (Figure 2).

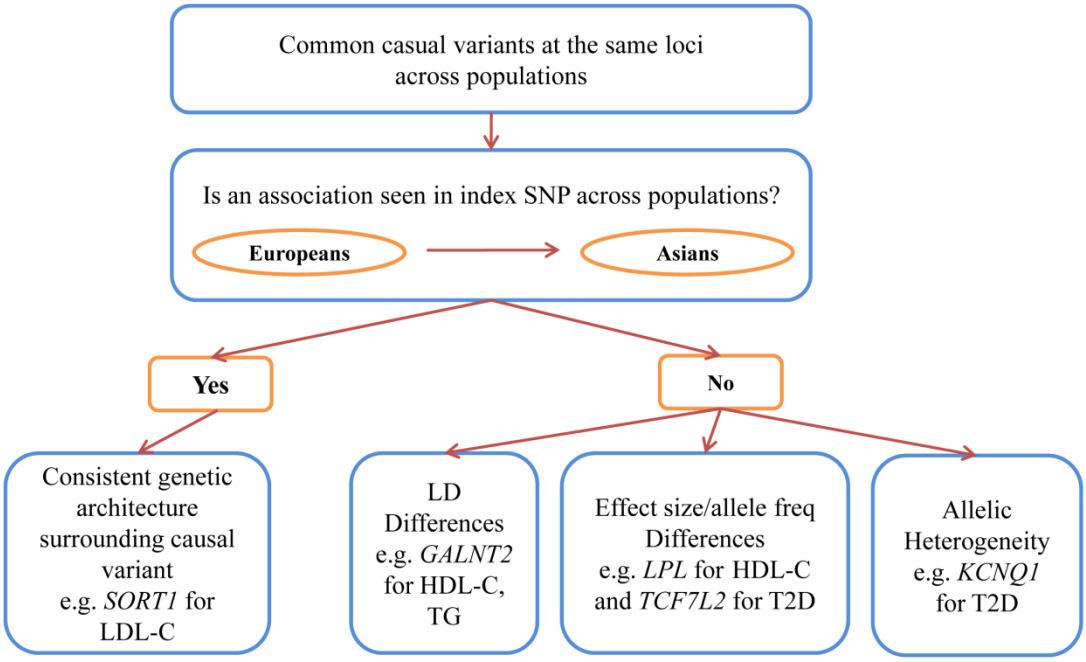


Figure 2. Schematic diagram describing the transferability of association signals across populations.

The consistent association of the sortilin 1 (*SORT1*) locus with low-density lipoprotein cholesterol (LDL-C) observed across different populations suggested common functional variants and/or similar linkage disequilibrium patterns with the functional variants[62,63]. In Kathiresan et. al., the discovery index SNP was rs646776 in European populations, with consistent evidence of association in Chinese (*P*-value ≤ 0.001), Malays (*P*-value = 4.00×10^{-3}) and Asian Indians (*P*-value = 3.00×10^{-3})[62]. There was no evidence of inter-population variation in linkage disequilibrium at this locus[64] and further meta-analysis in populations of European, East Asian, South Asian and African American ancestry further confirmed the association of this locus across multiple populations (*P*-value = 1.00×10^{-170} in 100,184 Europeans; *P*-value = 5.00×10^{-13} in 15,046 East Asians; *P*-value = 6.00×10^{-18} in 9,705 South Asians; *P*-value = 2.00×10^{-14} in 8,601 African Americans)[63].

Differences in effect sizes at implicated index variant or regional linkage disequilibrium patterns would affect the transferability of association signals across populations. In 2008, Kooner et. al. reported suggestive evidence of rs326 at the lipoprotein lipase (*LPL*) gene locus in 1,005 Europeans (*P*-value = 1.8×10^{-5}) with high-density lipoprotein cholesterol (HDL-C) but the same index SNP did not show any evidence of association in 1,006 Asian Indians (*P*-value = 0.14)[65]. The allele frequencies of the index SNP was comparable across the two populations, with a risk allele frequency of 0.71 in the Europeans and 0.76 in the Asian Indians, but the observed effect sizes were substantially different, with per allele change in log units of 0.025 in Europeans and 0.008 in Asian Indians. In one of the largest genome-wide meta-analyses of lipid traits, a different index SNP rs12678919 was found to be associated with HDL-C at the same *LPL* locus[63]. These two SNPs were correlated with $r^2 = 0.410$ using data from the CEU population in The 1000 Genomes Project[40,66], suggesting the presence of allelic heterogeneity. In 100,184 individuals of European descent, there was genome-wide significant association of HDL-C at the *LPL* locus (*P*-value = 1.00×10^{-97}) and suggestive evidence of association in 9,705 Asian Indians (*P*-value = 2.00×10^{-7})[63]. Thus it is possible that: (i) rs326 could be a poor surrogate of the functional variant; (ii) the heterogeneity in effect sizes were possibly modulated by differences in genetic background; or (iii) heterogeneous environmental exposures had an impact on the power to detect the association in Asian Indians.

Allelic frequency differences could determine the ease at which some disease implicated variants are more easily detected in particular populations. The *TCF7L2* locus is by far the locus associated with Type 2 Diabetes with the largest effect size. However, the risk allele frequencies of index SNP rs7903146 at this locus range from 0.026 in the HapMap Han Chinese CHB, 0.037 in HapMap Chinese in Metropolitan Denver CHD, Colorado, 0.035 in HapMap Japanese from Japan JPT and 0.279 in HapMap CEU. If the same locus is implicated in Type 2 Diabetes in these

East Asian populations, many more samples will be needed to detect the association due to low allele frequencies.

It is also possible that there exist different disease functional variants in the same locus across populations known as allelic heterogeneity. Alternatively, the particular disease causal variants are specific to certain populations. The potassium voltage-gated channel, KQT-like sub-family, member 1 (*KCNQ1*) was implicated in Type 2 Diabetes, and was first reported in Japanese populations and further replicated in a Danish population[67,68]. The Diabetes Genetics Replication And Meta-analysis (DIAGRAM+) Consortium reported a secondary signal at this locus in Europeans about 7.5Mb away from the previous reported finding. Conditional analysis by adjusting for previously reported variant in association analysis suggests that there might be more than one casual variant at this locus[12]. Within the Europeans, linkage disequilibrium between these two index SNPs was 0.01.

The protein coding gene UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase 2 (*GALNT2*) locus was found to be significantly associated with both HDL-C and triglycerides in European populations but no evidence was reported across populations of East Asian, South Asian and African American descent[62,63,69]. The index SNP could be a poor surrogate of the functional variants in non-European populations if indeed there are shared functional variants, or there could be allelic heterogeneity at the locus, or perhaps the risk implicated variant is specific to the Europeans only. Regional analysis of the linkage disequilibrium comparing HapMap CEU with HapMap Asian panel (CHB and JPT) and other reference populations in Singapore[70] indicates some evidence of variation in linkage disequilibrium patterns between populations at this locus[64]. Thus the ability to contrast local

regions of linkage disequilibrium between populations becomes vital to understand the transferability of such findings across different populations.

Linkage disequilibrium diversity at particular regions of the genome, differences in allele frequency or effect size, allelic heterogeneity at genetic loci and presence of different disease functional variants in diverse populations could all affect the transferability of association signals across populations, and affect our ability to use meta-analysis to increase statistical power or replication to confirm associations (Figure 2). Conducting genome-wide analyses in different populations thus has an important role in helping us understand the genetic architecture of diseases through the similarities and differences exhibited across populations and provide insights into the pathogenesis of these diseases.

1.4.3. Can we identify novel susceptibility loci by studying different ethnic groups?

Diseases prevalence varies across populations or the same disease could have heterogeneous pathogenesis resulting in differing genetic susceptibility in diverse populations. The prevalence of a particular disease in a population determines the population risk and ease of collecting diseased cases for such large scale genetic studies that generally allow us to detect variants of small effect sizes.

Genetic association studies have been extremely successful in populations of European descent, and these studies are increasingly being reported in other populations including East Asians, South Asians, Africans and Mexican Americans. Due to their evolutionary history, some disease implicated variants are more easily detected in some populations than others. *KCNQ1* was first shown to be associated with Type 2 Diabetes in 6,800 case control pairs from Japanese, Korean

and Chinese populations (odds ratio OR = 1.43, 95% CI = 1.34 – 1.52, *P*-value = 2.50 x 10^{-39})[67]. Of note, the allele frequency of the index SNP was 0.95 in the European replication population compared to 0.68 in the combined 6,800 Asian panel. In DIAGRAM+ Consortium, association at this index SNP was detected in 8,130 cases and 38,987 controls (OR = 1.14, 95% CI = 1.05 – 1.24, *P*-value of 2.70 x 10^{-3})[12]. Thus, there is still potential for other populations to detect novel susceptible locus that might have been harder to pin down in populations of European ancestry.

1.4.4. Importance of finer disease phenotyping

Fundamentally, the presentation of a disease is an interplay between genetic and environmental factors. Often, there are many subtypes within a disease and changes in the classification with time reflect our knowledge of the disease and its heterogeneity. Using diabetes mellitus as an example, there are predominantly two forms of diabetes: Type 1 Diabetes which could be seen as an autoimmune condition; and Type 2 Diabetes that is affected by insulin secretion and/or insulin action[71,72]. Given current knowledge on the disease pathogenesis, some of the implicated variants or genes can be linked to either of the two mechanisms: (i) defects in insulin secretion due to abnormalities in the beta-cells and/or function; and (ii) irregularities in the insulin action (Figure 3)[73,74]. Thus variants acting on glycemic traits and body mass index (BMI) could also be relevant to the pathogenesis of Type 2 Diabetes, as both pathways contribute towards the progression of Type 2 Diabetes[75]. In individuals with Type 2 Diabetes, either of these pathways may predominate. Analyzing individuals with different pathogeneses might dilute effects of genetic variants that affect specific pathways. Better phenotyping may improve the power to discriminate between genetic variants acting along different pathways.

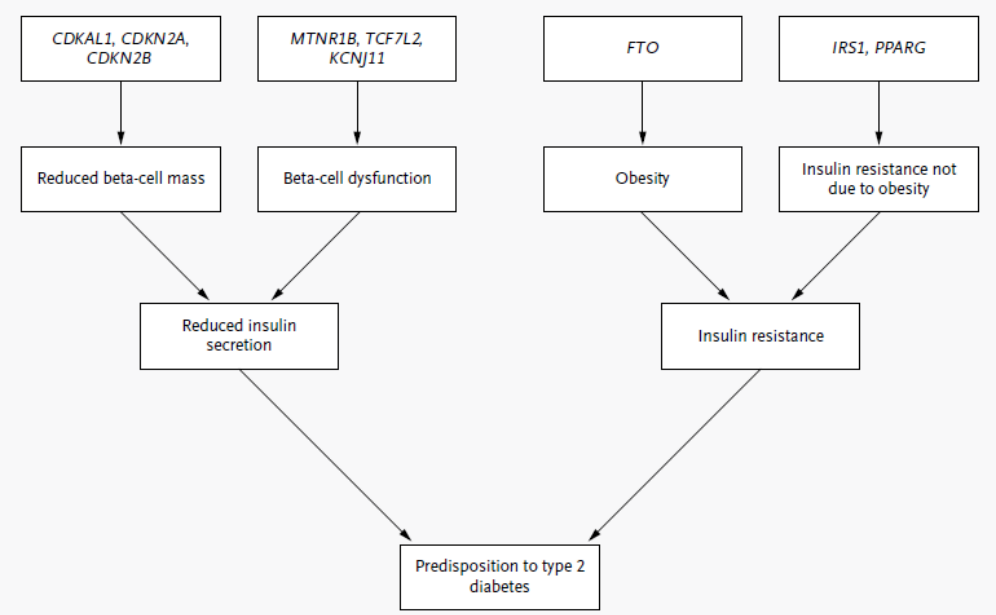


Figure 3. Pathways to Type 2 Diabetes Implicated by Identified Common Variant Associations.
Type 2 diabetes results when pancreatic beta cells are unable to secrete sufficient insulin to maintain normoglycemia, typically in the context of increasing peripheral insulin resistance. The beta-cell abnormalities fundamental to type 2 diabetes are thought to include both reduced beta-cell mass and disruptions of beta-cell function. Insulin resistance can be the consequence of obesity or of obesity-independent abnormalities in the responses of muscle, fat, or liver to insulin. Examples of susceptibility variants that, given current evidence, are likely to influence predisposition to type 2 diabetes by means of each of these mechanisms are shown.

Figure 3. Pathways to Type 2 Diabetes implicated by identified common variant associations (originally from reference 73).

Timpson et. al. performed a stratified analysis of Type 2 Diabetes, by defining non-obese cases below the median BMI and obese cases to be above the median BMI. The association between the *FTO* variant and Type 2 Diabetes was only present in the obese cases, consistent with the known effects of *FTO*. Careful selection of subjects in these studies could boost or dilute association signals. These search strategies for Type 2 Diabetes genetic susceptibility loci complement one another and provide more insights into the pathogenesis and heterogeneity of Type 2 Diabetes[12,76-78].

CHAPTER 2 – AIMS

2.1. Study 1 – Singapore Genome Variation Project (SGVP) – Chapter 4

Variation in linkage disequilibrium across populations of different ancestry has been previously documented. This study aimed to

i) Investigate the similarities and differences in linkage disequilibrium patterns across 100 Singapore Chinese, 100 Singapore Malays and 100 Singapore Asia Indians.

ii) Provide a sufficiently dense resource of at least 1.4 million SNPs to facilitate genetic association studies carried out in Singapore or populations with similar genetic background.

2.2. Study 2 – Transferability of established Type 2 Diabetes loci in three Asian populations – Chapter 5

As of 2010, there were more than 40 extensively replicated Type 2 Diabetes implicated loci, primarily discovered in populations of European ancestry. With the increasing prevalence of Type 2 Diabetes in China and India, the multi-ethnic demography of Singapore provided the genetic diversity to look at

i) Novel association signals with Type 2 Diabetes in Asian populations.

ii) Relevance of the established loci in Asian populations and their implications the genetic architecture of Type 2 Diabetes.

2.3. Study 3 – Meta-analysis of Type 2 Diabetes in populations of South Asian ancestry – Chapter 6

Large scale meta-analyses in populations of European descent have discovered Type 2 Diabetes implicated loci of small effect sizes. In one of the largest meta-analysis of Type 2 Diabetes in South Asians, we sought to

i) Discover novel genetic loci in the South Asians that might be better powered due to differences in allele frequency as a consequence of evolution or population specific effects due to differences in genetic and/or environmental background.

ii) Establish the relevance of European established loci in South Asians.

2.4. Study 4 – Heterogeneity of Type 2 Diabetes in subjects selected for extremes in BMI – Chapter 7

Type 2 Diabetes is a highly heterogeneous disease, with several pathways involved. Genetic and environmental risk factors interact. Refining cases and controls using risk factor BMI could provide insights into the mechanisms and pathogenesis of Type 2 Diabetes.

CHAPTER 3 – STUDY POPULATIONS AND METHODS

3.1. Genome-wide study populations and genotyping methods

3.1.1. Singapore Genome Variation Project (SGVP) – Study 1

Sampling from an inter-population study of healthy volunteers on the genetic variability to drug responses[79], 100 anonymised subjects from each of the three ethnic groups, Chinese, Malays and Indians, were randomly selected to participate in the Singapore Genome Variation Project. Gender and population membership information were available, and self-reported population membership to each of the three ethnic groups were further ascertained on the basis that all four grandparents belonged to the same ethnicity. Subjects were further required to declare a medical history free of cardiac condition at the time of recruitment. The use of volunteers from a drug response study might generate ascertainment bias, but the additional information of ethnic descent for two previous generations at recruitment was a more crucial condition for the purpose of this study. Ethical approval was granted by two independent Institutional Review Boards (IRBs), National University Hospital Singapore for the original drug response study and National University of Singapore for genome-wide genotyping of the selected subjects respectively.

Among the 300 subjects, a total of 292 unique subjects comprising of 99 Chinese, 98 Malays and 95 Indians with genomic DNA were successfully genotyped on two genome-wide commercial arrays, Affymetrix Genome-Wide Human SNP Array 6.0 and Illumina HumanHap1M-single. One subject from each ethnic group was genotyped twice for data quality purpose and an additional control subject was removed from the data after genotype calling, making the total number of subjects genotyped to be 295.

For the Illumina array, genotype calls for the 295 subjects were assigned by the proprietary calling algorithm GenCall[47,48] in Illumina's BeadStudio Suite using clusterfiles provided by

Illumina. A genotype calling (GC) score generated by the calling algorithm was implemented to determine the confidence of the assigned genotype. Any SNP with a GC score ≥ 0.15 was accepted while a SNP with GC score < 0.15 was assigned as a NULL genotype. Overall genotype call rate of 274 unique samples after genotyping quality control filters was 99.86%. Details of genotyping quality control are given in Section 3.3.1.

For Affymetrix, a preliminary calling on the 3,022 control probes on the array was performed using the Dynamic Modelling[42] (DM) algorithm. There were seven repeats due to failure to achieve the minimum DM call rate of 86% on the control probes on the array, of which one sample was eventually discarded when the second round of genotyping still failed to make the cut-off. CEL files containing intensity calculations of pixel information of 295 subjects were submitted for calling by the BirdSeed[46] calling algorithm developed by Broad and made available in Affymetrix Power Tools apt-1.8.6 (released March 4, 2008). Models files used were from version 2.6 and na24 of the Product files. Overall genotype call rate of 277 unique samples after genotyping quality control filters was 99.51% (see Section 3.3.1).

3.1.2. Singapore Diabetes Cohort Study (SDCS) – Studies 2 & 4

The Singapore Diabetes Cohort Study (SDCS) comprised of Chinese, Malay and Asian-Indian individuals with Type 2 Diabetes currently on follow-up in hospitals and polyclinics, namely the National Healthcare Group Polyclinics, National University Hospital Singapore and Tan Tock Seng Hospital since 2004[80]. The diagnostic criteria in Singapore primary health care providers follows international norm and physicians would use local clinical practice guidelines (CPG, http://www/moh.gov.sg/content/dam/moh_web/Publications/Guidelines/Withdraw20CPGs/cgp_Diabetes%20Mellitus-Jun%202006.pdf). Participants were not further tested for Type 2 Diabetes diagnosis. The primary aim of this initiative was to identify genetic and environmental risk

factors for diabetic complications such as diabetic nephropathy and to develop novel biomarkers for tracking disease progression. The participation response was excellent with a participation rate exceeding 90%. Questionnaire data as well as clinical data from case notes of consenting participants were obtained. The blood and urine specimens of these participants were collected and archived at -80°C.

Using a combination of Illumina HumanHap 610 Quad and HumanHap 1Mduov3 Beadchips on Illumina BeadStation, 2,202 unique Chinese subjects were genotyped for genome-wide analysis. Eight subjects were genotyped on both arrays for quality checks.

3.1.3. Singapore Prospective Study Program (SP2) – Studies 2 & 4

The Singapore Prospective Study Program (SP2) invited a total of 10,747 participants from four previous cross-sectional studies: Thyroid and Heart Study 1982–1984[81], National Health Survey 1992[82], National University of Singapore Heart Study 1993–1995[83] and National Health Survey 1998[84] to participate in a repeat examination in 2004 – 2007. By data linkage to the Registry of Births and Deaths in Singapore using each participant's National Registration Identity Card, 517 subjects who were deceased at the time of follow-up, six subjects who had migrated and 85 subjects with errors in their record and hence un-contactable were excluded. Of the remaining participants, 2,673 were not contactable and 30 refused to take part in the study. Among these participants 5,157 of them completed the questionnaire and provided their blood specimens. Informed consent was obtained from the participants and ethic approvals were obtained from two Institutional Review Boards (National University of Singapore and Singapore General Hospital)[85].

The questionnaires were interviewer-administered, collecting information on demographic and lifestyle factors such as smoking and alcohol consumption as well as medical history including

physician diagnosed diabetes mellitus, hypertension and hyperlipidemia. Participants were required to fast for ten hours overnight before the health examination in the following morning. Health examination included anthropometric measurements of weight, height and waist-hip-ratio. Two readings of blood pressure were also taken from the participants after five minutes of rest, seated, using an automated blood pressure monitor (Dinamap Pro100V2; Criticon, Norderstedt, Germany) by trained examiners. If the difference between two readings of either the systolic blood pressure was greater than 10mmHg or the diastolic blood pressure was greater than 5mmHg, a third reading is measured. The mean values of the closest two readings were then calculated. Venous blood was drawn, collected in plain and fluoride oxalate tubes to be stored at 4°C for a maximum of four hours prior to processing. All biochemical analyses of the blood specimens were carried out at the National University Hospital Referral Laboratory. Serum high density lipoprotein cholesterol, total cholesterol and triglycerides were measured using an automated autoanalyzer (ADVIA 2400; Bayer Diagnostics, Tarrytown, New York). Low density lipoprotein cholesterol level was calculated using the Friedewald formula. Plasma glucose was assayed with enzymatic methods (ADVIA 2400) from the blood collected. A random urine specimen (collected at subject's convenience without a pre-specified time or prior conditions) was collected and urinary creatinine was measured using a commercial assay (Immulite; Diagnostic Products Corporation, Gwynedd, United Kingdon for urinary albumin and Roche Diagnostics GmbH, Mannheim, Germany for creatinine).

In a case control design with SDCS, 2,483 Chinese DNA samples from SP2 were genotyped on a combination of Illumina HumanHap 610 Quad and HumanHap 1Mduov3 Beadchips. Similarly, eight subjects were genotyped on both arrays for quality checks.

3.1.4. Singapore Malay Eye Study (SiMES) – Studies 2 & 4

The Singapore Malay Eye Study (SiMES) was a population based cross sectional study of Singapore Malays (using the criteria set by the Singapore Census to define Malays[86]) living in Singapore[87]. Age-stratified random sampling of all Malay adults (provided by Ministry of Home Affairs) aged 40 to 80 years old residing in fifteen residential districts in the southwestern part of Singapore was used to obtain a list of 4,168 eligible participants. Of the eligible participants, 3,280 participated with a response rate of 78.7%. The study was designed to quantify the prevalence and risk factors for visual impairment and major eye diseases in an adult urban Malay population. Ethical approval was obtained from the Singapore Eye Research Institute Institutional Review Board and informed consent was obtained from the participants.

A detailed interviewer-administered questionnaire was administered to collect demographic data, lifestyle factors, eye symptoms, systemic medical history and current medications. Blood pressure was taken with the participant seated and after five minutes of rest using a digital automatic blood pressure monitor (Dinamap model Pro series DP110X-RW, 100V2, GE Medical Systems Information Technologies, Inc., United States of America). Each participant went through a series of eye photographs and imaging, including fundus photography to determine retinopathy and age-related maculopathy and retinal imaging. Participants were not required to fast overnight. To determine levels of serum lipids (high density lipoprotein cholesterol, low density lipoprotein cholesterol and total cholesterol), glycosylated haemoglobin A1c (HbA1c), creatinine and random glucose, 40 mL of non-fasting venous blood was collected from each participant and sent to the National University Hospital Reference Laboratory. Additional plasma was stored at -80°C and DNA extracted from serum was stored at the Singapore Tissue Network at -80°C. Samples of urine were collected to determine levels of microalbuminuria and creatinine at the Alexandra Hospital Laboratory.

In all, 3,072 Malay subjects were genotyped on Illumina HumanHap 610 Quad Beadchips. For the population-based Type 2 Diabetes genome-wide study, both cases and controls were selected from the population-based cross-sectional study where diabetic cases were defined as having either history of diabetes or had HbA1c level ≥ 6.5%[88] and controls had no history of diabetes and HbA1c level < 6%. Finally, for Malays, there were 794 diabetic cases and 1,240 controls available for analyses in the Type 2 Diabetes genome-wide study.

3.1.5. Singapore Indian Eye Study (SINDI) – Studies 2, 3 & 4

The Singapore Indian Eye Study (SINDI) is part of the Singapore Indian Chinese Cohort (SICC) Eye Study and comprised of the Indian arm of the cohort. Similar to the SiMES study, SINDI is a population-based cross-sectional study of Singapore Asian Indians (using the criteria set by the Singapore Census to define Indians) living in Singapore[89]. Age-stratified random sampling of all Asian Indian adults (provided by Ministry of Home Affairs) aged 40 to 80 years old residing in fifteen residential districts in the southwestern part of Singapore was used to obtain a list of 6,350 eligible participants. Of the eligible participants, 3,400 participated with a response rate of 53.5%. The study was designed to quantify the prevalence and risk of eye diseases in ethnic Indian residents of Singapore. Ethics approval was obtained from the Singapore Eye Research Institute Institutional Review Board and informed consent was obtained from the participants.

A detailed interviewer-administered questionnaire was administered to collect demographic data, lifestyle factors, eye symptoms, systemic medical history and current medications. The health examination procedures included measurements of height, weight, blood pressure and pulse rate, followed by a comprehensive ocular examination such as fundus photography and retinal imaging. Participants were not required to fast overnight. Non-fasting venous blood was collected to

determine levels of serum lipids (high density lipoprotein cholesterol, low density lipoprotein cholesterol and total cholesterol), glycated haemoglobin HbA1c, creatinine and random glucose. DNA was extracted from serum and stored at the Singapore Tissue Network at -80°C. Samples of urine were collected to determine levels of microalbuminuria and creatinine.

Finally, 2,953 Indian subjects were genotyped on Illumina HumanHap 610 Quad Beadchips. In SINDI, diabetes case control ascertainment was determined with HbA1c level[88]. Cases were defined as having either a history of diabetes or HbA1c $\geq$ 6.5%. Controls had no history of diabetes and HbA1c level < 6%. This yielded 977 diabetic cases with 1,169 controls for the Indian Type 2 Diabetes genome-wide study.

3.1.6. London Life Sciences Population (LOLIPOP) Study – Study 3

The LOLIPOP study is a population-based cohort of European white and South Asian Indian men and women aged 35 – 75 living in West London[90]. Ancestry was self reported and South Asians were only included in the study if all four grandparents were born in the India Subcontinent (countries of India, Pakistan, Sri Lanka or Bangladesh).

An interviewer-administered questionnaire conducted by trained research nurses collected information such as country of birth, language and religion of participants, parents and grandparents for assignment of ethnic subgroups. Data on medical history, family history, current prescribed medication, cardiovascular risk factors, alcohol intake and leisure-time physical activity were also obtained. The physical examinations included blood pressure (mean of 3 readings, taken with an Omron 705CP), height, weight, waist and hip circumference and 12 lead electrocardiography (ECG). Fasting blood (8 hours) was collected for plasma glucose, lipids, insulin and high sensitivity C-reactive protein.

Type 2 Diabetes was defined as physician diagnosis cases on treatment or fasting glucose ≥ 7.0mmol/L. Controls had no prior history of Type 2 Diabetes and had fasting glucose < 7.0mmol/L. The study was approved by the Local Research Ethic Committee and all participants gave written informed consent.

Subjects were genotyped on a combination of Illumina HumanHap317 and HumanHap610 arrays. Quality filters included call rate at least 97.5%, Hardy-Weinberg Equilibrium (HWE) *P*-value < 10^{-6}, MAF at least 1% and sample call rates of 95%. Duplicates and related individuals were also removed[91].

3.1.7. Pakistan Risk of Myocardial Infarction Study (PROMIS) – Study 3

PROMIS is an ongoing case-control study of acute myocardial infarction (MI) and other cardiometabolic traits in urban Pakistan which included about 7,500 case control pairs as at October 2010[92]. PROMIS has been approved by the research ethics committee of the Center for Non-Communicable Diseases (CNCD), Pakistan and research ethics committee of each of the institutions involved in participant recruitment. MI cases had typical ECG changes, positive troponin tests and MI symptoms within the previous 24 hours. Controls are frequency matched to cases by age (by 5 years age band) and gender from either: (i) visitors of patients attending the out-patient department, (ii) patients attending the out-patient department for routine non-cardiac complaints or (iii) non-blood related visitors of index MI cases. For each participant, non-fasting blood samples were collected. For MI cases, blood collection was done within 24 hours of symptom onset.

Type 2 Diabetes cases were defined based on physician diagnosis, prior use of oral hypoglycemic and/or HbA1c level > 6.5%. Controls had no history of Type 2 Diabetes and had HbA1c level < 6%.

Genotyping was performed using the Illumina HumanHap 670W array at the Sanger Institute, United Kingdom. Quality filters included call rate at least 98%, HWE *P*-value < 10^{-6}, MAF at least 1% and sample call rates of 95%. Duplicates and related individuals were also removed[93].

3.2. Replication study populations

3.2.1. The COBRA Study – Study 3

Cobra is a population-based study of adults recruited in a cluster of randomized trial of strategies to control hypertension ('Population Based Strategies for Effective Control of High Blood Pressure in Pakistan', trial registration number NCT00327574) in Karachi, Pakistan[94]. Within the largest metropolitan city in Pakistan, a multi-stage cluster random sampling design was used to randomly select twelve geographical clusters, of which a listing of all individuals from all households in the selected areas was made from the census. All individuals aged 40 and above and able to give consent were invited by trained community health workers to participate in the study. Ethical approval was obtained from the Ethics Review Committee at the Aga Khan University, Pakistan.

Physical examination included blood pressure with a calibrated automated device (Omron HEM-737 TM Blood Pressure Monitor) in the sitting position after 5 minutes of rest, and collecting anthropometry measurements such as height, weight, waist and hip circumferences. Laboratory tests included fasting plasma glucose (Synchron Cx-7/Delta, Beckman, US) and DNA extraction.

Type 2 Diabetes cases were defined as physician diagnosis on diabetic medications or fasting blood glucose ≥ 7 mmol/L. Controls had no history of Type 2 Diabetes and fasting glucose < 7mmol/L.

3.2.2. Chennai Urban Rural Epidemiology Study (CURES) – Study 3

CURES is an ongoing epidemiology study of a representative sample of 26,001 South Asians recruited using a random sample technique in Chennai, India. Written informed consent was obtained from all study participants and the research protocol was approved by the Institutional Ethics Committee of the Madras Diabetes Research Foundation[95]. This study was carried on in several phases. In the first phase, 26,001 individuals were recruited based on systematic random sample technique.

Type 2 Diabetes cases were defined as self-reported on drug treatment at Phase I. At Phase 3, every 10th subject without Type 2 Diabetes at Phase I were invited to undergo an oral glucose tolerance test (OGTT). Those with post-load glucose ≥ 11.1mmol/L were labeled as newly detected diabetic subjects. Controls had no history of Type 2 Diabetes, fasting glucose < 6.1mmol/L and post-load glucose < 7.8mmol/L.

3.2.3. Diabetes Genetics in Pakistan Study (DGP) – Study 3

Indigenous Pakistani subjects were recruited in collaboration with Baqai Institute of Diabetology and Endocrinology (BIDE), Karachi, Pakistan[96]. Informed consent was obtained from all study participants and the study was approved by the BIDE Institutional Review Board.

Type 2 Diabetes cases were recruited either from hospitals within Mirpur District or from specifically organised Diabetes Awareness camps. Controls were recruited from community

screening camps set up throughout Mirpur District. Fasting blood and post-load glucose tests were not available for these subjects. Controls were thus defined as random blood glucose < 7mmol/L.

3.2.4. Mauritius Cohort – Study 3

A population-based survey was undertaken in 1998 in the subtropical island Mauritius that included individuals who were 20 years and older, with a total of 6,291 individuals examined. Participants of self-reported South Asian ancestry (about 70% of the population) were included in the present study[97].

Participants not on any diabetes medication were subjected to 2-hour OGTT. Venous blood samples drawn at baseline fasting and 2 hours post ingestion of glucose were centrifuged and separated immediately. Plasma glucose was measured using the YSI glucose analyzer (Yellow Springs Instruments, OH, USA).

Using the World Health Organization (WHO) 1999 criteria, Type 2 Diabetes was diagnosed if subject reported a history of diabetes and was taking hypoglycaemic medication, or fasting plasma glucose level was ≥ 7.0mmol/L and/or the 2-hour post-load value was ≥ 11.1mmol/L. Normal glucose tolerance was assigned if the fasting plasma glucose level was < 6.1 mmol/L and the post-load value was < 7.8 mmol/L.

3.2.5. Ragama Health Study (RHS) – Study 3

The Ragama Health Study (RHS) is a population-based study of South Asian men and women aged 35-64yrs living in the Ragama Medical Officer of Health (MOH) administrative area, near Colombo, Sri Lanka[98]. Participants gave consent to their available health records and ethical

approval for the study was obtained from the Ethics Committees of the Faculty of Medicine, University of Kelaniya and the National Center for Global Health and Medicine.

Participants were subjected to a 12-hour fast and interviewed by trained personnel to obtain information on medical, socio-demographic and lifestyle variables. Blood pressure and anthropometric measurements were also obtained. For the purpose of fasting glucose and HbA1c quantification, 10-mL sample of venous blood was drawn from each participant.

Type 2 Diabetes cases were defined as either physician diagnosis on treatment or fasting glucose > 7.0mmol/L or HbA1c level > 6.5%. Controls had no history of Type 2 Diabetes, fasting glucose < 6.1mmol/L and HbA1c level < 6.0.

3.2.6. Sikh Diabetes Study (SDS) – Study 3

Participants of the Sikh Diabetes Study were recruited from Sikhs living in the Northern states of India, including Punjab, Haryana, Himachal Pradesh, Delhi, and Jammu and Kashmir[99]. All participants provided written informed consent for investigations and all protocols and consent documents were reviewed and approved by the University of Oklahoma and the University of Pittsburgh Institutional Review Boards as well as the Human Subject Protection Committees at the participating hospitals and institutes in India.

Type 2 Diabetes cases were defined as physician diagnosis on treatment, fasting plasma glucose level of ≥ 7.0mmol/L, or 2-hour post glucose load level ≥ 11.1mmol/L. Controls had no prior history of diabetes and had normal glucose tolerance given by fasting glucose < 6.0mmol/L and post glucose < 7.8mmol/L. Participants with impaired fasting glucose and/or impaired glucose tolerance were excluded from analysis.

3.2.7. Singapore Consortium of Cohort Studies (SCCS) – Study 3

The Singapore Consortium of Cohort Studies includes Type 2 Diabetes cases and population-based controls from Singapore. Type 2 Diabetic cases are recruited from hospitals and polyclinics, namely Alexandra and Changi General Hospitals and Ang Mo Kio, Jurong, Choa Chu Kang, Yishun and Pasir Ris Polyclinics while controls are recruited from the general population. The diagnostic criteria in Singapore primary health care providers follows international norm and physicians would use local clinical practice guidelines (CPG, http://www/moh.gov.sg/content/dam/moh_web/Publications/Guidelines/Withdraw20CPGs/cgp_Diabetes%20Mellitus-Jun%202006.pdf). Participants were not further tested for Type 2 Diabetes diagnosis. Participants gave broad consent for i) future biomedical research, ii) access to their medical records and iii) linkages to various registries.

All participants completed a structured questionnaire, providing information on demographics, socio-economic status and medical history (including history of diabetes) and had measurement of anthropometric measures and blood pressure. Fasting blood samples were collected for blood glucose and lipid measurements.

For the purpose of this study, only participants of self reported South Asian ancestry were included. Type 2 Diabetes cases were defined as physician diagnosis on treatment while controls had no prior history of diabetes and fasting glucose < 6.1mmol/L.

3.2.8. Sri Lankan Diabetes Studies – Study 3

The Sri Lankan Diabetes Cardiovascular Study (SLDCS) is a cross-sectional nationally–representative epidemiological investigation which recruited 4,388 subjects (40% male)[100]. DNA

collection was only initiated midway through the SLDCS collection, limiting the number of samples available for genotyping. The Sri Lankan Young Diabetes Study (SLYDS) recruited a total of 992 patients with early onset diabetes (aged between 16 and 40 and were ≤ 45 years of age when they first joined the study) from the three largest hospitals in Sri Lanka between 2005 and 2006.

Type 2 Diabetes cases were mainly from the SLYDS and included 176 diabetic cases from SLDCS. These cases included previous physician diagnosed Type 2 diabetics or newly diagnosed diabetics (fasting glucose ≥ 7.0mmol/L or post-load glucose ≥ 11.1mmol/L). After biochemistry and immunological testing (absence of anti-GAD antibodies) on the basis of clinical history (independence from insulin for at least 6 months after diagnosis), 890 subjects from SLYDS were ascertained to be diabetic. Across the two sets of diabetic cases, additional exclusion criteria were applied (GAD antibodies ≥ 14units/ml, age ≥ 80 years, and missing sex information) to generate a total of 1,066 cases available for genotyping at the Diabetes Research Laboratory, Oxford. Among recruited subjects from SLDCS, 3,372 had normal glucose tolerance based on the results of a 75g OGTT, interpreted using American Diabetes Association (ADA) and WHO criteria.

3.2.9. United Kingdom Asian Diabetes Study (UKADS) – Study 3

South Asians subjects residing in the United Kingdom with Type 2 Diabetes (physician diagnosed, on treatment, n = 892) were recruited to UKADS from Birmingham and Coventry, UK[101]. All subjects were of Punjabi ancestry, confirmed over three generations, and originated predominantly from the Mirpur region of Azad Kashmir, Pakistan. Ethnically-matched controls (n = 449) were recruited from the same geographical areas through community screening. Informed consent was obtained from each of the study participants and the study was approved by the Birmingham East, North and Solihull Research Ethics Committee. Genomic DNA was

extracted from venous blood using the Nucleon® protocol (Nucleon Biosciences, Coatbridge, UK).

Normal glucose tolerance was defined as either fasting plasma glucose < 6.1mmol/L and 2-hour plasma glucose < 7.8mmol/L on a 75g OGTT (where possible) or random blood glucose < 7.0mmol/L.

3.3. Methods for genome-wide data

3.3.1. *Genome-wide genotyping arrays*

Different genome-wide genotyping arrays were used by each of the different studies to survey the genome. Basic characteristics of these commercial arrays were given in Table 1.

Table 1. Basic characteristics of genome-wide genotyping arrays used in the different studies.

Arrays	Basic characteristics
Affymetrix Genome-Wide Human SNP Array 6.0[37]	Contains more than 906,600 SNPs and more than 946,000 non-polymorphic copy number probes. Approximately half of the SNPs were derived from previous generation Affymetrix arrays and the remaining from the HapMap project.
Illumina HumanHap1M (single/duo chips)[102]	Contains ~1.2 million SNPs, focusing on tag SNPs, SNPs in genes, and non-polymorphic markers in known and novel copy number regions.
Illumina HumanHap317[103]	Contains over 317,000 tagSNPs selected from HapMap Phase I[17]. In addition, approximately 7,300 non-synonymous SNPs and high density of tagSNPs on the Major Histocompatibility Complex (MHC) region was selected.
Illumina HumanHap610[39]	Contains over 600,000 tagSNPs.
Illumina HumanHap660[39]	Contains over 657,000 evenly spaced markers with approximately ~100,000 markers that target observed common copy number variants.

3.3.2. Quality control (QC)

Genotype calling is an automated unsupervised process to translate fluorescent intensities from hybridization experiments into discrete genotype calls. Across hundred thousands of SNPs, it is not realistic to inspect each SNP for the accuracy of their genotype calling assignments. Thus a set of quality filter is always implemented on genome-wide data to filter SNPs with potential erroneous calling, which will likely lead to spurious association results[45,104].

The SDCS and SP2 samples were genotyped as part of a Chinese diabetes case control design while SGVP, SiMES and SINDI were QCed as population-based studies. For each array in each

study population, genotype clustering was first performed with the proprietary cluster files from Illumina (GenCall)[47,48]. Samples achieving 99% sample call rate were then used to generate local clusterfiles (GenTrain) for a second round of calling. A threshold of 0.15 was implemented on the GenCall score to decide on the confidence of the final assigned genotypes for a valid genotype call.

A similar set of quality filters were implemented on the SGVP, SDCS, SP2, SiMES and SINDI populations. The genotype quality of the arrays was assessed independently and in three main phases according to the following sequence (Table 2):

a. Preliminary SNP QC on autosomal SNPs to obtain a pseudo-clean set of SNPs for sample QC.
b. Sample QC to identify duplicates, samples with high missing genotype calls, cryptic related samples and samples with discordant ethnic membership or gender across genetically inferred data and self-reported information from clinical data.
c. Final SNP QC after the exclusion of the samples from (b) on the full panel of SNPs to obtain the set of SNPs for downstream analyses.

The final set of samples and genotypes post-QC was used for subsequent analyses. For SDCS, SP2, SiMES, SINDI, LOLIPOP and PROMIS, the post-QC genotypes were used as imputation seed, to statistically infer unobserved genotypes that were present in the imputation reference panel.

In Study 3, genome-wide association scans comprised of the SINDI study population from Singapore, the LOLIPOP study from United Kingdom and the PROMIS study from Pakistan. Samples from LOLIPOP were genotyped on both Illumina610quad and IlluminaHap317 while samples from the PROMIS study were genotyped on the Illumina670Quad only.

Similar quality control criteria was implemented across studies conducted in Singapore, namely SGVP, SDCS/SP2, SiMES and SINDI, while LOLIPOP and PROMIS in Study 3 applied slightly different sets of quality filters. Due to the collaborative nature of Study 3, it was not practical at the meta-analysis stage to standardize quality filters across all participating discovery cohorts. Each participating cohort had applied a set of reasonable quality filters to their data. Details of the quality control filters for each genome-wide cohort were presented in Table 2 below.

Table 2. Description of the quality filters on the genome-wide populations.

	Study 1		**Studies 2 & 4**			**Study 3**			
Population name	**SGVP**		**Chinese (SDCS/SP2)**		**Malays (SiMES)**	**Indians (SINDI)**	**LOLIPOP**		**PROMIS**
Study design	Healthy individuals (Chinese, Malay, Indians)		Diabetes case control		Population-based cohort	Population-based cohort	Population-based cohort		Acute myocardial infarction case control
Array type	Affymetrix SNP6.0 (Affy6.0)	Illumina1M-single (Illu1M)	Illumina610-quad (Illu610)	Illumina1M-duov3 (Illu1M)	Illumina610-quad (Illu610)	Illumina610-quad (Illu610)	IlluminaHap317 (Illu317)	Illumina610-quad (Illu610)	Illumina660-quad (Illu660)
Genotype calling	Birdseed[46]	GenCall (Illumina clusterfiles; genotyping score GC > 0.15) [47,48]	1. GenCall with Illumina clusterfiles 2. GenTrain: Samples with 99% call rates used to train project specific clusters 3. GenCall using new clusterfiles from Step 2 [47,48]	1. GenCall with Illumina clusterfiles 2. GenTrain: Samples with 99% call rates used to train project specific clusters 3. GenCall using new clusterfiles from Step 2 [47,48]	1. GenCall with Illumina clusterfiles 2. GenTrain: Samples with 99% call rates used to train project specific clusters 3. GenCall using new clusterfiles from Step 2 [47,48]	1. GenCall with Illumina clusterfiles 2. GenTrain: Samples with 99% call rates used to train project specific clusters 3. GenCall using new clusterfiles from Step 2 [47,48]	GenCall[47,48]	GenCall[47,48]	Illuminus[49]
SNP QC exclusion	Missingness > 5%, HWE significance across all samples $P < 10^{-8}$, monomorphic across all samples, and more than 1 discordant genotypes across three pairs of duplicates.		Missingness > 5%, HWE significance across controls $P < 10^{-6}$, and monomorphic across all samples.		Missingness > 5%, HWE significance across all samples $P < 10^{-6}$, and monomorphic across all samples.		Missingness ≥ 3%, HWE significance across controls $P < 10^{-6}$, and MAF < 1%.		
	Annotation issues								
Sample QC exclusion	Missingness > 2%, excessive identity-by-state (IBS) genotypes (higher missingness of the pair), discordance of ethnic membership between		Missingness > 5%, excessive heterozygosity, excessive identity-by-state (IBS) genotypes (higher missingness of the pair), discordance of ethnic membership between self-reported ethnicity and genetically inferred population ascertainment by principle components analysis with HapMap and SGVP samples, and gender				Missingness > 5%, excessive heterozygosity, excessive identity-by-state (IBS) genotypes (pi_hat ≥ 0.5 in LOLIPOP and pi_hat ≥ 0.37 in PROMIS to allow for higher consanguinuity in Pakistanis), discordance of ethnic membership		

	self-reported ethnicity and genetically inferred population ascertainment by principle components analysis, and gender discordance.	discordance.			between self-reported ethnicity and genetically inferred population ascertainment by principle components analysis with HapMap samples and Indian reference samples by Reich and colleagues[52] (PLINK[105] was used to select a set of LD pruned SNPs).		
Final SNP QC exclusion	Population specific Missingness > 5%, HWE significance $P < 0.001$, and more than one discordant genotype across the three pair of duplicated samples.	Missingness > 5%, HWE significance across controls $P < 10^{-4}$, and monomorphic across all samples.			NA	NA	NA
Merging QC exclusion	Only samples passing QC in both arrays are retained. Missingness > 5% for unique and common SNPs across both arrays, and SNP with higher missingness for common SNPs on both arrays.	Only samples passing QC in both arrays are retained. Cryptic relatedness and gender discrepancies across the two arrays, allelic differences between common SNP across the two arrays as determined by departure from the QQ-plots.	NA	NA	NA	NA	NA
Imputation	NA	IMPUTEv0.5.0[54]					
Imputation reference panel	NA	HapMap II JPT+CHB panel on build 36 release 22	HapMap II CEU+JPT+CHB+YRI panels on build 36 release 22				
Data handling softwares	R[106], PLINK[105], Eigenstrat[107,108]	R[106], PLINK[105], Eigenstrat[107,108], IMPUTE[54], SNPTEST[45]					

3.3.3. Type 2 Diabetes case control ascertainment

For Type 2 Diabetes case control design in Study 2, Chinese cases included subjects from the SDCS with diagnosis of Type 2 Diabetes while Chinese controls were selected from SP2 with no prior history of diabetes and had fasting glucose level less than 6.0mmol/L. The Chinese cases and controls were checked across the two genotyping arrays, for duplicates and cryptic relatedness. This yielded a total of 2,010 cases from SDCS and 1,945 controls from SP2 over the two arrays. The sampling frame of the SP2 study was from a mixture of heart and thyroid studies and National Health Survey. These studies are cross-sectional population based with stratified sampling of the Singapore population using the census data, thus reducing the likelihood of ascertainment bias in the SP2. Using SP2 as a control set with SDCS could introduce misclassification bias, where a proportion of the controls could have the outcome of interest (thus meeting the criteria for being a case) and possibility of having undiagnosed cases in the control set. However, this misclassification would have modest impact on power unless the misclassification bias was substantial[45]. On the other hand, SP2 provided a convenient resource for the selection of controls, and fasting glucose levels were available to screen undiagnosed Type 2 Diabetes from the population.

Diagnostic criteria of Type 2 Diabetes were not consistently defined across the cohorts. Type 2 Diabetes can be diagnosed based on three different measures: (i) fasting plasma glucose greater than 7.0mmol/L (126mg/dl); (ii) after a 75g glucose load, venous post-load (2-hour) plasma glucose level greater than 11.1mmol/L (200mg/dl); and (iii) HbA1c level greater than 6.5%[88]. These three measures are highly correlated but not perfect. HbA1c is a chronic glycemic measure which captures the degree of glucose exposure over time. Biological pathways leading to Type 2 Diabetes may affect specifically one glycemic trait more than others. Heterogeneity in phenotype definition may mean that we are more likely to detect variants acting through at least one or more

of the pathways. Nonetheless, many of these common pathways which lead to elevated fasting glucose, 2hour-glucose and HbA1c are likely to have biological relevance to the pathogenesis of Type 2 Diabetes. Table 3 summarized the number of Type 2 Diabetes cases and controls (discovery populations) in Studies 1 to 4 and Table 4 provided a summary of the diagnostic criteria of Type 2 Diabetes in Studies 2 and 3, including replication cohorts.

Table 3. Final sample counts post-QC for the genome-wide populations.

N	Study 1		Studies 2 & 4					Study 3			
Array type Population	Affy6.0	Illu1M	Illu610 Chinese SDCS	Illu610 Chinese SP2	Illu1M Chinese SDCS	Illu1M Chinese SP2	Illu610 Malays SiMES	Illu610 Indians SINDI	Illu610 South Asian LOLIPOP**	Illu317 South Asian LOLIPOP**	Illu660 South Asian PROMIS**
Genotyped (duplicates)	295 (3)	295 (3)	1,195 (8)	1,467 (8)	1,015 (8)	1,016 (8)	3,072 (0)	2,953 (0)	NA	NA	NA
Post QC on individual array	277	274	1,115	1,251	930	960	2,542	2,538	1,783 4,773	440 1,699	2,361 6,817
Final set of samples after merging*	268		1,082	1,006	928	939	2,542	2,538	NA	NA	NA
Diabetes study Cases (Cas) Controls (Ctl)	NA		Cas:1,082 Ctl:1,006		Cas:928 Ctl:939		Cas:794 Ctl:1,240	Cas:977 Ctl:1,169	Cas:1,783 Ctl:4,773	Cas:440 Ctl:1,699	Cas:2,361 Ctl:6,817
Meta-analysis	NA		Cas:3,781 Ctl:4,354								
			NA					Cas:5,561 Ctl:14,458			

* Merging refers to cross array check on common SNPs for samples from SGVP and only the Chinese samples from SDCS and SP2.
** Detailed genotyping information not available for collaborating cohorts.

Table 4. Characteristics of participants in the Type 2 Diabetes discovery and replication cohorts (originally from reference 109).

Cohort	Location	Genotyping method	# Case # Control	Type 2 Diabetes ascertainment
Discovery (Genome-wide)				
Singapore Diabetic Cohort Study (SDCS)/ Singapore Prospective Study Program (SP2)	Singapore	Whole genome	2,010 1,945	Case: Physician diagnosis in primary care facilities. Control: No personal history and fasting glucose < 6.1mmol/L
Singapore Malay Eye Study (SiMES)	Singapore	Whole genome	794 1,240	Case: Personal history or HbA1c ≥ 6.5%. Control: No personal history and HbA1c < 6.0%.
Singapore Indian Eye Study (SINDI)	Singapore	Whole genome	977 1,169	Case: Personal history or HbA1c ≥ 6.5%. Control: No personal history and HbA1c < 6.0%.
London Life Sciences Population Study (LOLIPOP)	United Kingdom	Whole genome	2,223 6,472	Case: Physician diagnosis on treatment or fasting glucose ≥ 7.0mmol/L. Control: No personal history and fasting glucose < 7.0mmol/L.
Pakistan Risk of Myocardial Infraction Study (PROMIS)	Pakistan	Whole genome	2,361 6,817	Case: Physician diagnosis, prior use of oral hypoglycemic and/or HbA1c > 6.5%. Control: No personal history and HbA1c < 6.0%.
Replication cohorts in Study 3				
COBRA	Pakistan	Kaspar	465 1,580	Case: Physician diagnosis on diabetic medication or fasting blood glucose ≥ 7.0mmol/L. Control: No personal history and fasting blood glucose < 7.0mmol/L.
Chennai Urban Rural Epidemiology Study	India	Sequenom	1,316 1,265	Case: Self-reported on drug treatment or post-load glucose ≥ 11.1mmol/L. Control: No personal history and fasting glucose < 6.1 mmol/L and post-load glucose < 7.8mmol/L.
Diabetes Genetics in Pakistan Study	Pakistan	Kaspar	840 1,225	Case: Physician diagnosed Type 2 Diabetes in hospital or Diabetes Awareness camps. Control: Random blood glucose ≤ 7.0mmol/L.
London Life Sciences Population Study (LOLIPOP)	United Kingdom	Kaspar	1,132 7,652	Case: Physician diagnosis on treatment or fasting glucose ≥ 7.0mmol/L. Control: No personal history and fasting glucose < 7.0mmol/L.
Mauritius Study	Mauritius	Kaspar	780 1,536	Case: Personal history or taking hypoglycaemic medication or fasting glucose ≥ 7.0mmol/L and/or 2-hour glucose ≥ 11.1mmol/L. Control: Fasting glucose < 6.1mmol/L and 2-hour glucose < 7.8mmol/L.
Pakistan Risk of Myocardial Infraction Study (PROMIS)	Pakistan	Kaspar	3,128 5,277	Case: Physician diagnosis, prior use of oral hypoglycemic and/or HbA1c > 6.5%. Control: No personal history and HbA1c < 6.0%.
Ragama Health Study	Sri Lanka	TaqMan	776 1,981	Case: Physician diagnosis on treatment or fasting glucose > 7.0mmol/L or HbA1C > 6.5%. Control: No personal history, fasting glucose < 6.1mmol/L and HbA1c < 6.0%.
Sikh Diabetes Study	India	Sequenom	1,387 1,732	Case: physician diagnosis on treatment, fasting glucose > 7.0mmol/L, or post-load glucose > 11.1mmol/L. Control: No personal history and fasting glucose < 6.0mmol/L and post-load glucose < 7.8mmol/L.
Singapore Consortium Of Cohort Studies	Singapore	Sequenom	1,613 1,230	Case: Physician diagnosis on treatment. Control: No personal history and fasting glucose < 6.1mmol/L.
Sri Lankan Diabetes Studies	Sri Lanka	TaqMan	841 1,471	Case: Hospital-recruited early onset diabetics. Control: Normal post-load glucose tolerance (Fasting glucose < 6.1mmol/L and post-load glucose < 7.8mmol/L).
United Kingdom Asian Diabetes Study	United Kingdom	Kaspar	892 449	Case: Physician diagnosed on treatment Control: Fasting glucose < 6.1mmol/L and post-load glucose < 7.8mmol/L.

3.3.4. Type 2 Diabetes case controls, stratified by BMI status

There are several pathways leading to disease manifestation of Type 2 Diabetes, one of which includes defects of insulin action in fat, muscle and liver that is commonly linked to obesity. One of the most common measures of obesity is the anthropometric measure, body mass index (BMI), given by the weight of an individual in kilograms (kg) divided by the square of height in metres (m). WHO of the United Nations classifies individuals with BMI $\geq$ 25 kg/m^2 as overweight and BMI $\geq$ 30 kg/m^2 as obese, using BMI as a surrogate for adiposity. While this cut-off has been correlated with increased risk of Type 2 Diabetes, cardiovascular disease and mortality, the distribution of body fat (central adiposity) and percentage of body fat appear to correlate less well with BMI in Asian populations[110,111].

In Study 4, we explored potential differences in genetic susceptibility when cases and controls are selected by their obesity status. Following the international WHO classification, we denote individuals with BMI less than 25 kg/m^2 as non-obese and individuals with BMI greater than or equal to 25kg/m^2 as overweight. By stratifying cases and/or controls into overweight and non-obese strata, we compare the association signals in each of these subsets of case control studies. Figure 4 below illustrates the various combinations of dichotomous BMI status and case control status in six case control substudies. Looking vertically across Subset1, we looked at only non-obese cases where non-obese cases were compared to all controls (Substudy 1.1), overweight controls (Substudy 1.2) and non-obese controls (Substudy 1.3). The most interesting substudy here was comparing the non-obese cases with the overweight controls, under the most extreme phenotypic contrast. With non-obese cases, they were less likely to be affected by the insulin action due to obesity. On the other hand, the controls, being overweight, were still not showing signs of Type 2 Diabetes. In Substudy 1.1, all controls were considered for increased sample size. Subset 2 comprised of three substudies comparing overweight cases with all controls (Substudy

2.1), non-obese controls (Substudy 2.2) and overweight controls (Substudy 2.3). In Substudy 2.2, hypercontrols (non-obese controls) were used and all controls were considered in Substudy 2.1 for increased sample size. For completeness, we included Substudy 1.3 and 2.3 where both cases and controls were either non-obese or overweight.

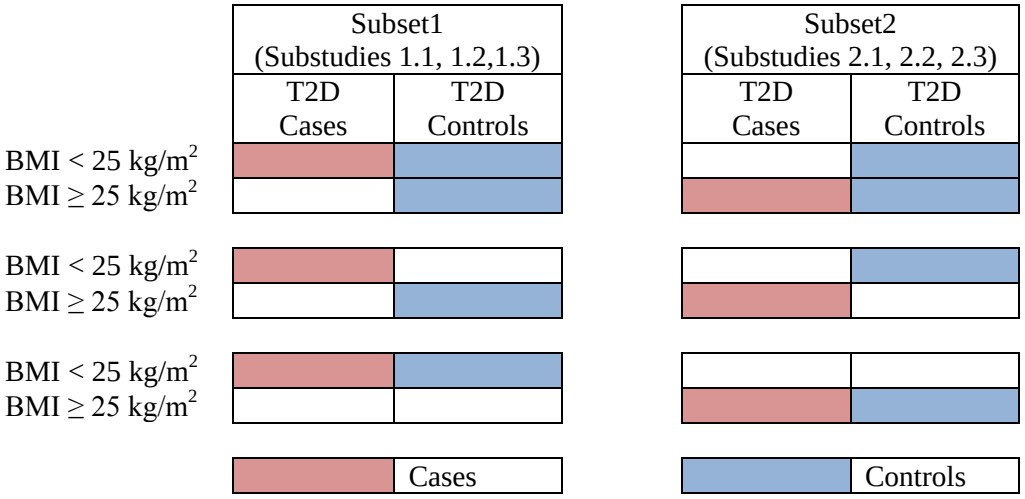


Figure 4. Schematic diagram for the study design of Study 4.

3.3.5. Population structure

Population structure arises when allelic frequencies variations within a genetic study are not related to the phenotype being studied, leading to false positive or false negative findings[112]. In the case-control setting, cases and controls could have inherent genetic differences and prevalence in diseases, generating spurious signals of association at loci where there are differences in the allelic frequencies. The large number of polymorphisms tested in genome-wide scans allowed an assessment of the impact of population structure on the association signals, using various statistical methods. In the following subsections, I will describe two commonly used statistical methods, genomic control and principal component analysis, to address population structure in genetic association studies.

a. Genomic control

The genomic control method (GC) makes use of a set of independent markers in the genome to estimate the inflation of the association test statistics as a result of underlying population structure[113,114]. Assuming that the majority of the SNPs tested are independent from the phenotype of interest, their statistical evidence is likely to resemble that obtained from the null hypothesis of no association, i.e. a Uniform distribution. The genomic control inflation factor estimates the inflation of the test statistics, and is defined as the ratio of the median of the empirically observed distribution of the test statistic to the expected median.

$$Genomic\ control\ inflation\ factor\ \lambda_{gc} = \frac{Median\ (observed\ \chi^2)}{Median\ \left(\chi_1^2\right)\ = 0.456}$$

The estimated inflation factor, λ_{gc}, is applied uniformly across the genome by inflating the standard error of the log odds. For each of the Type 2 Diabetes case control study, genomic control was applied at the individual study level to minimize any residual population structure and a final correction was further implemented at meta-analysis (double GC correction).

b. Principal components analysis (PCA)

Principal components analysis reduces the genotype data to continuous axes of genetic variation, that describe as much of the variability of the data as possible[107]. Using the covariance matrix between samples, orthogonal eigenvectors place individuals onto continuous axes of genomic variations. Plotting the eigenvectors against one another allow one to explore the presence of any population structure. Figure 5 below shows three PCA plots illustrating some examples of population structure. Figure 5A contained three well-separated clusters, indicating that samples from the three populations originated from genetically distinct subpopulations. The continuous cloud of points in Figure 5B suggested the presence of admixed populations where there was a spectrum of genetic variation within each of the two populations. A random scatter of points as in Figure 5C indicated the absence of population structure in the population.

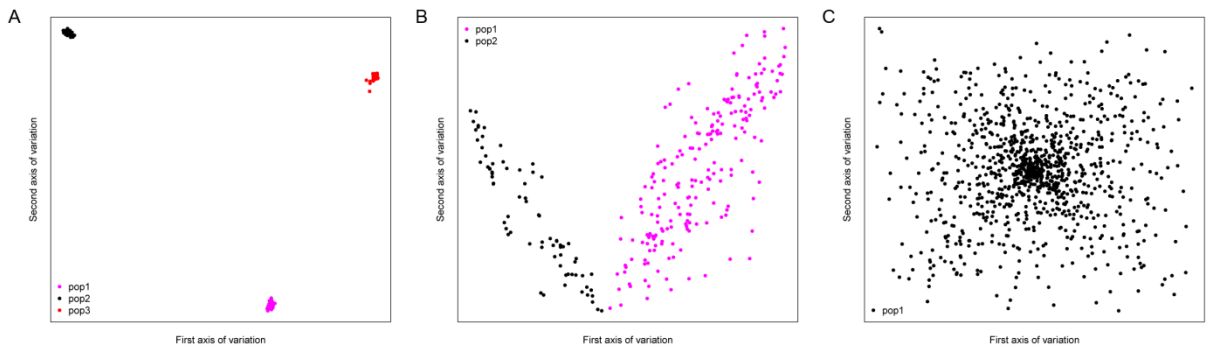


Figure 5. Principal components analysis plots of genetic variation. Points are colored in accordance to their self-reported ethnic membership. A) Well-separated clusters for three genetically distinct subpopulations; B) Two subpopulations showing some degree of admixture and C) Randomly scattered points indicating absence of population structure.

Principal components analysis was carried out in the genome-wide studies, against reference populations HapMap II[18], Indians from David Reich and colleagues[52] and SGVP[70]. In the context of the Singapore-based populations, a set of SNPs evenly thinned across the genome, reducing linkage disequilibrium, was used to infer the axes of variations. The PCA plots were showed in Figure 6. Figure 6A and Figure 6B compared the SGVP populations (Singapore Chinese: CHS; Singapore Malays: MAS and Singapore Indians: INS) with the HapMap II populations in the global scale over the first three components. The first axis of variation separated the Yoruba samples from Nigeria from the non-African populations. The second axis addressed the differences between the European CEU from the Asian populations. CHB, JPT and CHS were virtually indistinguishable from each other on the global scale. The SGVP populations were subsequently used to ascertain the ethnic membership of the samples in the GWAS. The Chinese GWAS samples cluster closely with the CHS samples from SGVP as shown in Figure 6C. The Malays and Indians on the other hand showed a continuous cloud suggesting greater degree of genetic diversity. In Figure 6D, the SiMES Malay samples exhibit continuous clouds over the first and second principal components while the SINDI Indian samples showed continuous spectrum of genetic diversity over the first three principal components. Hence for subsequent

association analyses, principal components were included as covariates to correct for population structure in these populations, namely first two principal components for Malays and first three principal components for Indians.

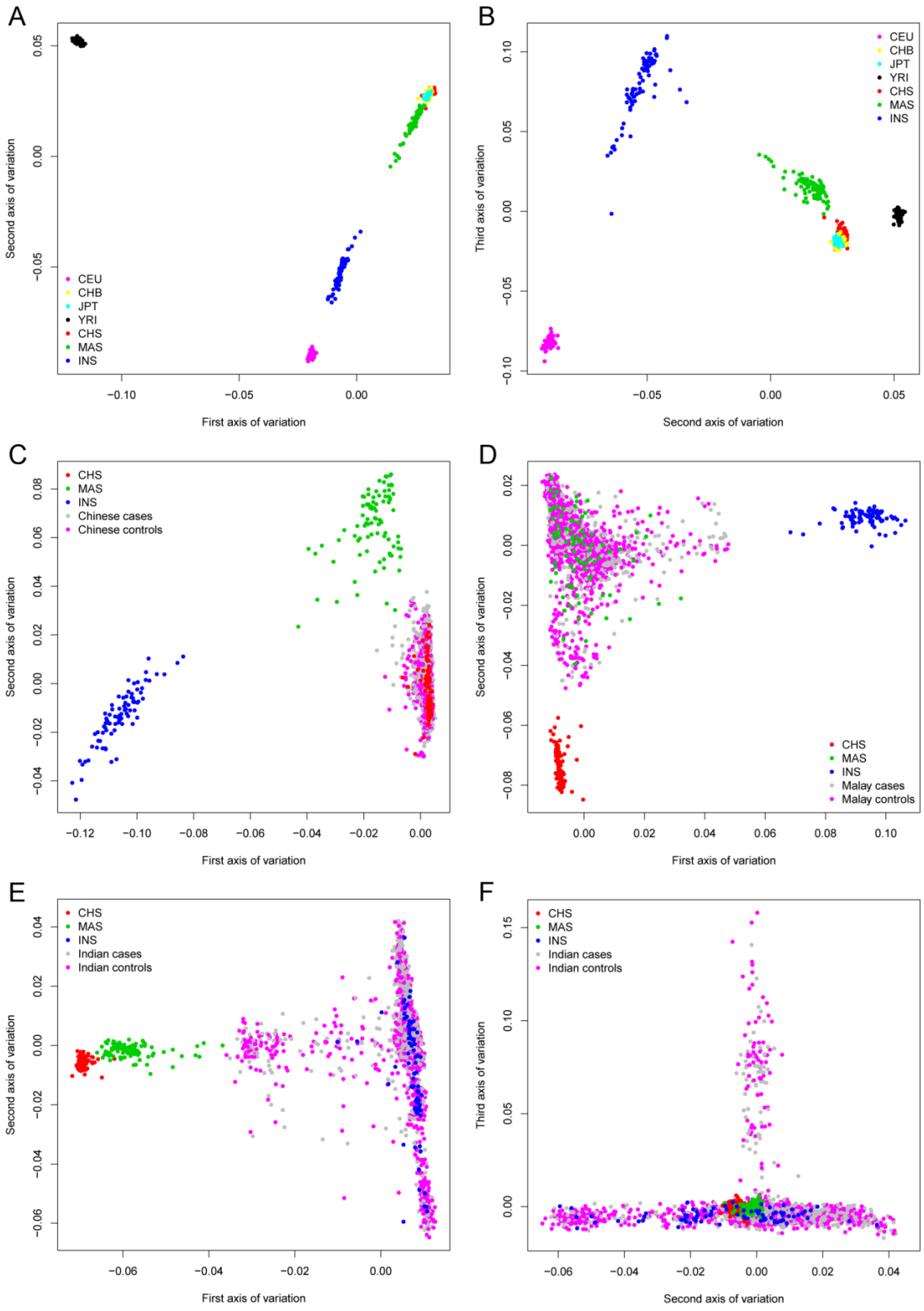


Figure 6. Principal components analysis plots of genetic variation. Each individual is mapped onto a pair of genetic variation coordinates represented by the first and second components or

second and third components. A) First two axes of variation of HapMap II (CEU: pink, CHB: yellow, JPT: cyan, YRI: black) and SGVP (CHS: red, MAS: green, INS: blue) and B) Second and third axes of variation of HapMap II and SGVP. Each of the Chinese, Malay and Indian Type 2 Diabetes case control study (cases: grey and controls: pink) are also superimposed onto SGVP. C) Chinese T2D cases and controls with SGVP; D) Malay T2D cases and controls with SGVP; E and F) Indian T2D cases and controls with SGVP [originally from references 70 and 115].

For Study 3, principal components analysis was also used to identify population outliers by comparison to reference populations from the HapMap and Indian samples from Reich and colleagues. Using the LOLIPOP data on the Illumina610 array, a set of 100,864 SNPs was pruned to reduce linkage disequilibrium across the SNPs and subsequently used in study-specific principal components analysis with HapMap II (Figure 7).

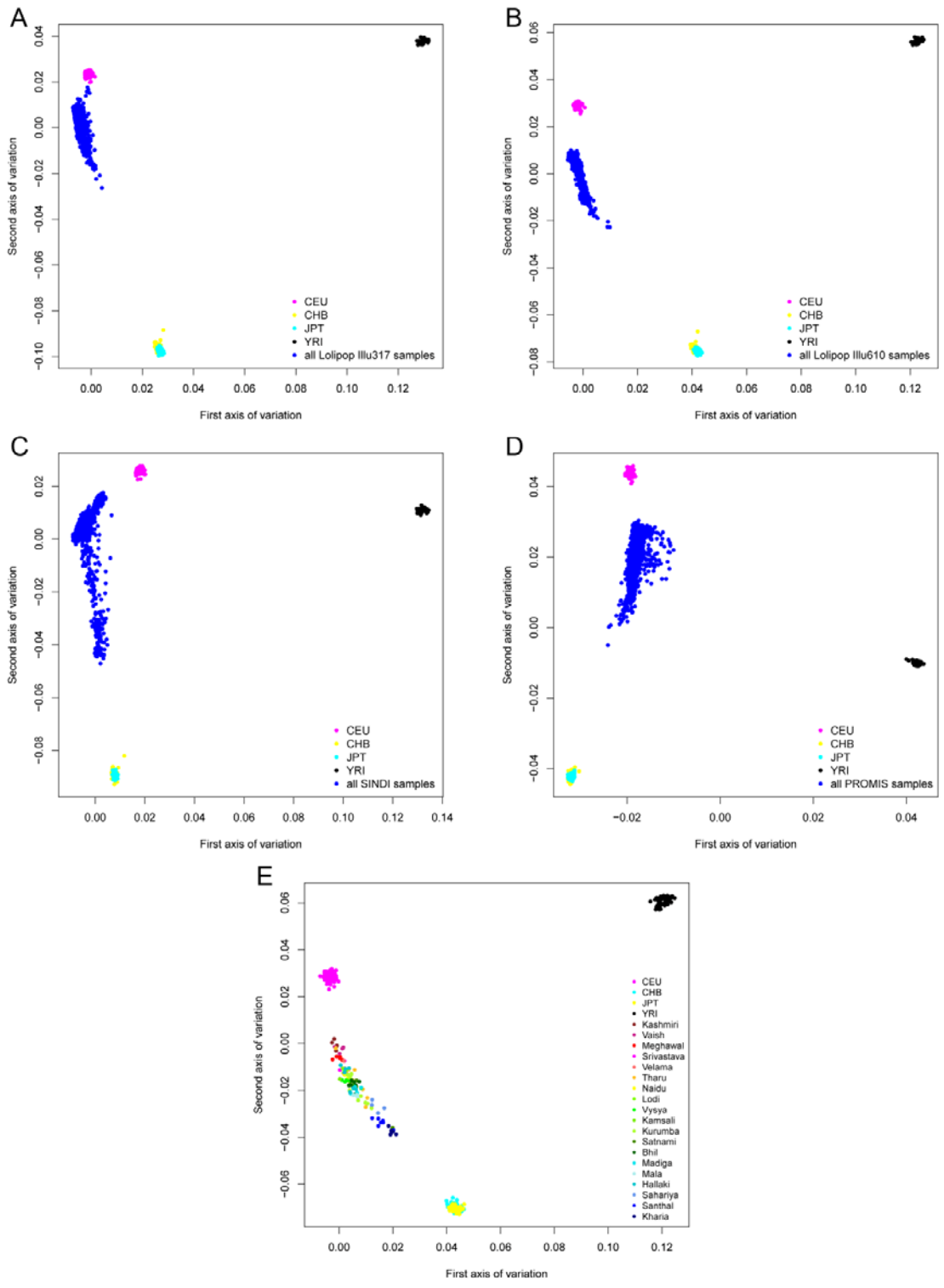


Figure 7. Principal components analysis plots of genetic variation in populations of South Asian ancestry. Each individual is mapped onto a pair of genetic variation coordinates represented by the first and second components or second and third components. A) First two axes of variation

of HapMap II (CEU: pink, CHB: yellow, JPT: cyan, YRI: black) and LOLIPOP samples genotyped on the Illumina317 array (blue); B) First two axes of variation of HapMap II and LOLIPOP samples genotyped on the Illumina610 array (blue); C) First two axes of variation of HapMap II and SINDI samples genotyped on the Illumina610 array (blue); D) First two axes of variation of HapMap II and PROMIS samples genotyped on the Illumina670 array (blue); E) First two axes of variation of HapMap II and Reich's Indian samples as reference (originally from reference 109).

3.3.6. Tests of association and conditional analyses

In assessing a SNP for association with the outcome, there are different modes of inheritance for the genetic load. Each of these modes of inheritance has the highest statistical power when the genetic model is concordant with the true effect of genetic burden. While the general model does not hold any prior assumption about the genetic load, being a two degree of freedom test, it is typically less powerful than the additive model with only one degree of freedom. In all our analyses, we only assume the additive mode of inheritance where each additional copy of the outcome-implicated allele increases the genetic risk by the same magnitude for all our association tests to avoid incurring multiple testing from different genetic models.

Logistic regression was used to test for association between Type 2 Diabetes (binary outcome) and autosomal SNPs in the Singapore genome-wide scans by applying SNPTESTv1.1.5[45] (http://www.stats.ox.ac.uk/~marchini/software/gwas/snptest.html). Chromosome X was not included in any of the analyses. The missing data likelihood method implemented in the –proper option in SNPTEST incorporated genotype uncertainties in the imputed data at the association tests. For SNPs directly observed on the genotyping arrays, the observed genotypes were reported and imputed results were not used in association testing. Association analyses were performed separately in the two genotyping arrays for the Chinese and later combined using meta-analysis (see section 3.3.7.) to obtain an overall association result for the Chinese.

A *P*-value cut-off of 5×10^{-8} was used to declare genome-wide significance for any polymorphism associated with Type 2 Diabetes. In 2007, it was proposed that a significance threshold of 5×10^{-7} at power of 0.5 corresponds to posterior odds in favour of a true positive at 10:1[45]. As genome-wide studies expanded the number of SNPs tested by imputation, this benchmark of significance threshold was further revised to 5×10^{-8}, which corresponded to more than a million independent tests across the genome[116-118]. Visual assessment of the genotype clusters was done on observed genotypes with statistical significance (*P*-value < 10^{-4}) in the meta-analysis (see section 3.3.7.) and observed SNPs in the top ten regions of each individual GWAS. SNPs with ambiguous assignment of genotype calls were removed from further analyses.

Primary analysis in Study 3 comprised of gender-specific association analyses and these analyses were combined with the meta-analysis method (see section 3.3.7.) to obtain a final summarized association result on common directly genotyped SNPs across the three studies. Secondary analyses included the following, i) all samples combined using whole genome imputed data on HapMap II; ii) males only; iii) females only; iv) gender specific, adjusted for BMI; v) lean Type 2 Diabetes cases (BMI < 25kg/m^2) versus overweight controls (BMI > 25kg/m^2).

3.3.7. Imputation

Analogous to solving a missing data problem, imputation techniques utilize the local linkage disequilibrium and population genetics models to infer unobserved genotypes[54,55,119,120]. In the instance where there is no population specific reference panel, the use of mixture panels had been recommended to provide more variation in the reference haplotypes.

In the Singapore genome-wide scans, the final post-QC set of genotype data was used as seed for imputation using IMPUTE v0.5.0[54] (https://mathgen.stats.ox.ac.uk/impute/impute.html). In the

Chinese, imputation was carried out using the JPT+CHB panel on build 36 release 22. All four panels of the HapMap II on build 36 release 22 were combined to form a mixture reference panel for the Malays and Indians. In addition, for the known Type 2 Diabetes loci, a second set of imputation was performed at these loci with IMPUTE v2.1.2 where it was possible to allow for more than one reference panel. This allowed the inclusion of population specific genotypes as an unphased reference panel which will hopefully reduce false negative findings that might have otherwise occurred with an inappropriate imputation reference panel being used. Specifically, we included the SGVP population specific genotypes as an unphased reference panel. For the Chinese, the phased reference panel was HapMap JPT+CHB and the unphased reference panel was SGVP CHS. For the Malays and Indians, all four HapMap population panels were the base reference with SGVP MAS and INS as the additional unphased panels respectively.

The genotype data was split into chunks of 10Mb and a 250kb buffer on both sides was implemented to avoid edge effects during imputation[54]. The effective population size of the YRI was used when imputing against the mixture reference panel. For imputation targeting specific loci, a 5Mb region was imputed around the index SNP for each locus with similar buffer size and effective population size. The imputation resulted in posterior probabilities for each of the genotype classes.

Similar imputation methods[54] were also applied on the LOLIPOP and PROMIS data, against pooled haplotypes from the HapMap II panels.

3.3.8. Meta-analysis

Aggregating evidences over multiple small GWAS increase the effective sample size and statistical power to detect any real associations. Using the fixed effects inverse variance model,

we pooled evidences, first over the genotyping arrays to obtain an overall association result in the Chinese and later across all the three ethnic groups/cohorts in Studies 2 and 4 respectively. This assumes a common effect size for a particular allele at the SNP across the studies and is implemented in the program METAL (http://www.sph.umich.edu/csg/abecasis/Metal/index.html). Similarly, the gender-specific association results from each of the genotyping array/study populations in South Asian populations were combined and summarized using meta-analysis, giving a summary association measure across males and females combined.

3.3.9. Replication and selection of SNPs

To confirm the findings in Study 3, we sought to reproduce the results in nine independent studies of South Asian ancestry. A total of 13,170 Type 2 Diabetes cases and 25,396 controls were available from the following studies: The COBRA study[94], the Chennai Urban Rural Epidemiology Study (CURES)[95], Diabetes Genetics in Pakistan (DGP) study[96], Mauritius cohort[97], Ragama Health Study (RHS)[98], Sikh Diabetes Study (SDS)[99], Singapore Consortium of Cohort Studies (SCCS), Sri Lankan Diabetes Study[100] and United Kingdom Asian Diabetes Study (UKADS)[101].

The SNP selection process is presented in the flow diagram below (Figure 8). SNPs within previously established Type 2 Diabetes loci were not considered for replication. All in all, 20 SNPs were selected for follow up based on the following selection criteria: i) seven SNPs from primary analysis with *P*-value < 10^{-5}; ii) twelve SNPs with *P*-values between 10^{-5} and 10^{-4} in primary analysis and had the lowest *P*-values in fixed effects inverse variance meta-analysis with populations of European ancestry from DIAGRAM+[12] and iii) one SNP reaching genome-wide significance of *P*-value < 5×10^{-8} in the female only secondary analysis.

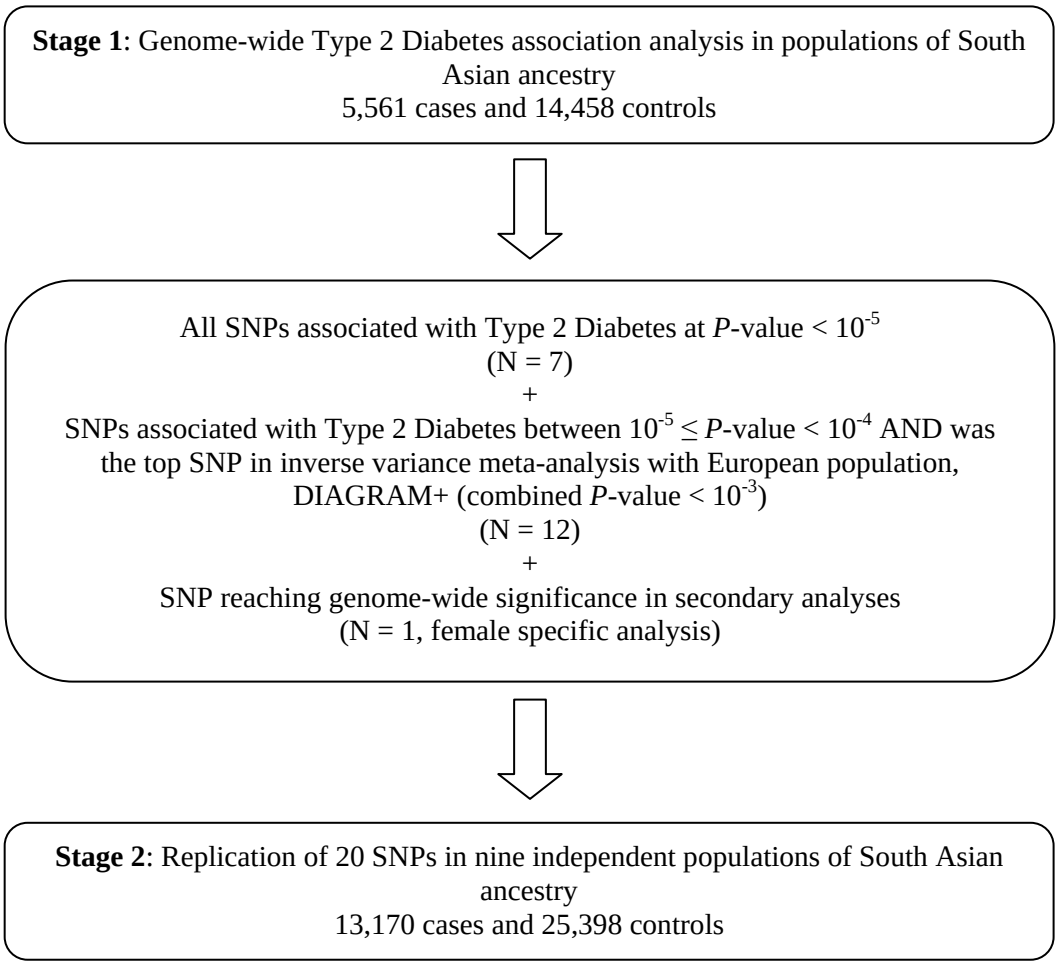


Figure 8. Summary of study design from the discovery stage to replication in Study 3.

De-novo genotyping of the SNPs in replication samples were performed by various methods such as KASPAR (K-Bioscience Ltd, UK), Sequenom MassArray or TaqMan assays. Similar to quality control filters for genome-wide scans, samples with poor call rates (< 90%) and SNPs with call rates < 95% or deviated from HWE at *P*-value < 2.5 x 10^{-3} were excluded. Study-specific association tests were performed and combined across replication cohorts using fixed effects meta-analysis. Threshold for significance was defined as *P*-value < 2.5 x 10^{-3}, after Bonferroni correction for 20 SNPs.

3.4. Methods for population genetics

3.4.1. Fixation index, Fst

The fixation index, Fst, is a measure of population differentiation, that quantifies the fraction of total genetic variation due to differences between populations[17]. The fixation index ranges from 0 to 1, with 0 indicating no differentiation and 1 indicating complete differentiation. The fixation index was implemented as the following:

a. Weighted Fst across the genome, accounting for differences in the number of chromosomes

$$Fst = 1 - \frac{\sum_j \binom{n_j}{2} \sum_i 2\frac{n_{ij}}{n_{ij-1}} x_{ij}(1 - x_{ij}) / \sum \binom{n_j}{2}}{\sum_i 2\frac{n_i}{n_i - 1} x_i(1 - x_i)}$$

where:

x_{ij} denotes the estimated MAF of SNP i in population j.
n_{ij} denotes the number of genotyped chromosomes at SNP i
n_j denotes the number of genotyped chromosomes in population j

b. SNP-specific Fst

SNP-specific Fst was computed for each pair of populations for every SNP that passed QC and common across all populations. The pairwise Fst values between HapMap II populations and SGVP and comparison within SGVP were summarized by summary statistics such as mean, median, 1st and 3rd quartiles. The Fst between two populations is defined as follows:

$$Fst = \frac{(p_1 - p_2)^2}{(p_1 + p_2)(2 - p_1 - p_2)}$$

where p_1 and p_2 denote the allele frequencies of a specific allele at a SNP in population 1 and 2 respectively.

3.4.2. Haplotype phasing

Statistical methods coupled with population genetics model are commonly used to infer phase information of genotype data and construct haplotypes containing the sequence of alleles on the

same chromosome[55,56,121-123]. The SGVP merged genotypes were phased using the fastPHASE program (version 1.3) for subsequent analyses[123]. Preliminary runs were performed to investigate the optimal choice of parameters in realistic computational running time. Assuming that similar haplotypes originate from the same cluster, the number of haplotype clusters K initiated ranged from 6 to 20 at the default setting of 20 expectation-maximization runs. Each chromosome was phased independently with ten iterations for error rate estimations. To determine K, 1,000 consecutive SNPs were randomly selected at each iteration, of which approximately 10% of the observed genotypes were masked and inferred by the algorithm. The discordance between the observed and inferred genotypes averaged over ten iterations constituted the error rate and the K with the lowest error rate was used as the eventual number of haplotype cluster. Based on the empirical error rates, the SGVP populations were phased within each subpopulations using K = 14.

3.4.3. Measures of linkage disequilibrium

Summary measures of linkage disequilibrium between any two SNPs were computed off the phased haplotypes of SGVP using Haploview[124]. Only SNPs with MAF greater than 5% were considered and we computed LD measures between a focal SNP and all other SNPs within 250kb upstream and downstream of the focal SNP. The two metrics are r^2 (square of genetic correlation coefficient) and D'[22,32].

3.4.4. Variation in LD (varLD)

Comparison of regional linkage disequilibrium between two populations was performed with the varLD algorithm[125,126]. This method could be implemented in two ways, (i) sliding window approach across the whole genome; or (ii) a targeted approach around a specified genomic region. The targeted approach tests the null hypothesis that the local regional pattern of correlation

between pairs of SNPs in the window is identical across the two populations compared and yields a Monte Carlo *P*-value that quantifies the statistical evidence of departure from the null.

All possible pairwise comparisons between the populations of HapMap II and SGVP were made. The sliding window approach defined windows of 50 consecutive SNPs common to both populations. For each of these windows, we compared the two 50 x 50 symmetric matrices of signed r^2 (r^2 linkage disequilibrium measure with the sign of the D' metric) from the two populations respectively. The sum of the absolute differences between the ranked eigenvalues of the two matrices constituted a score for each window. The extent of linkage disequilibrium in each window was assessed by comparing the relative rank of the score against the distribution of scores across the genome. Regions falling in the top 5% of the distribution were identified as candidate regions of variation in linkage disequilibrium. To further facilitate comparisons across multiple population-pairs, each pairwise distribution scores were standardized to have mean of zero and standard deviation of one.

In Study 2 on Singapore Type 2 Diabetes genome-wide association studies, varLD was implemented as a targeted approach centered on each of the established Type 2 Diabetes implicated loci. Comparisons were carried out between the three populations in SGVP and HapMap CEU. A region of 400kb centered on the index SNP was considered and a score was given by the difference of the trace of the eigen-decomposition of the signed r^2 matrices. To assess the statistical significance of each score, a Monte Carlo *P*-value was generated with 10,000 iterations of re-sampling from data combined over the two populations under the null of no differences in regional linkage disequilibrium.

The targeted varLD approach was also used in Study 3 to compare the regional pattern of linkage disequilibrium surrounding the index SNP of the six new loci implicated in Type 2 Diabetes in populations of South Asian ancestry. We performed 1,000 iterations of the Monte Carlo procedure within a 300kb window around the index SNP in each of the locus between any two of the following populations: (i) HapMap II CEU; (ii) SGVP INS and (iii) 60 randomly chosen control samples from the LOLIPOP study.

3.4.5. Detecting signatures of positive selection

Natural selection influences patterns of the human genome, through the removal of deleterious mutation (negative selection or purifying selection), the rapid rise and eventual fixation of positive selection, also known as Darwinian selection and the maintenance of multiple alleles through balancing selection or heterozygosity superiority. Positive selection is typically characterized by long haplotypes, which has risen to high frequency rapidly due to the selection pressure but relatively unbroken by recombination rates[127].

We first defined the extended haplotype homozygosity (EHH) as the probability of identity-by-descent for two randomly chosen haplotypes that were carrying the core SNP/haplotype of interest within 1Mb around the core region[128]. Briefly, EHH is a measure of haplotype identity in a core region as a function of distance. Using a single SNP as a unit, haplotype homozygosity starts at 1 and decays to 0 with increasing distance away from the focal SNP. EHH for a focal SNP is computed by comparing the probabilities of each distinct haplotypes (formed by extending away from the focal SNP in both directions) occurring in randomly selected chromosomes across samples. A recent selection relatively unbroken by recombination will have extended stretches of haplotype similarity due to unbroken transmission of an extended haplotype and EHH will likely remain close to 1 around the core region with a much slower rate of decay.

Both the integrated haplotype score (iHS)[129] and cross-population extended haplotype homozygosity (XP-EHH)[130] utilized the computation of EHH.

a. Integrated haplotype Score (iHS)

Ancestral allele information was obtained from the Haplotter website (http://haplotter.uchicago.edu/) and HapMap II averaged recombination rates was used. The SNPs considered for analyses had MAF greater than 5% with consistent allele information across ancestral/derived alleles designated for the SGVP genotype data. For each SNP, the EHH was computed for each focal SNP, extending in both directions until an EHH score of 0.05 or a 2.5 Mb region was reached. For adjacent SNPs with gaps of between 20kb and 200kb, a scaling factor was applied to the genetic distance so as to minimize the loss of SNPs and reduce spurious signals due to large gaps. Taking the integral under the EHH curve for each of the ancestral and derived alleles, we obtained the integrated EHH scores. Taking logarithm of the ratio of integrated EHH scores, the resultant raw iHS were then standardized to have mean zero and standard deviation one in 20 derived allele frequency bins, each spanning 5%. The iHS test is standardized across the genome and provides a measure of how unusual a region is compared to the rest of the genome, hence there is no formal significant test[129]. To identify regions of the genome with an unusual density of high standardized iHS scores, we computed the proportion of SNPs with $|iHS| > 2.0$[131] in non-overlapping windows of 100kb, containing at least ten SNPs. A less stringent absolute iHS score was applied to first isolate the unusual regions of the genome and subsequently windows containing top 1% proportion of the SNPs were identified.

b. Cross-population extended haplotype homozygosity (XP-EHH)

XP-EHH compares the evidence of selection across two populations at a core SNP, for instance, alleles that had risen to fixation in one population but not in the other populations. The use of a

reference population makes overlap between methods and findings across population difficult to interpret. Hence the HapMap YRI population with its greater genetic diversity was chosen as the reference population for all pairwise comparisons and only SNPs common to both populations within 1Mb of the core SNP were considered. Within the limits of 1Mb, the test was valid only if there exists another SNP with an EHH of between 0.3 and 0.5. In the case of multiple SNPs satisfying the above criterion, the SNP with an EHH closest to 0.04 was considered. For each population, the integral of the EHH at all SNPs bounded by the core SNP and the second SNP (with EHH between 0.3 and 0.5) were calculated and the logarithmic ratio of the two resultant integrals was defined as the XP-EHH log ratio. In a similar fashion as iHS, the log-ratios were standardized to have mean zero and standard deviation one. XP-EHH was primarily used to confirm differential iHS signals hence we defined a candidate selection region as a region with a cluster of SNPs with |XP-EHH scores| > 2.5. This corresponds to the top 1% of the standardized XP-EHH distribution.

CHAPTER 4 – SINGAPORE GENOME VARIATION PROJECT (SGVP)Motivation

The fundamental concept underlying genome-wide association studies relies on detecting indirect associations, where the genetic variants being queried are not directly responsible for the disease but are located near the causal variants and thus correlated (in linkage disequilibrium) with the causal variants. The HapMap Project cataloged a set of informative markers that capture common genetic variations across four populations from African, European and Asian ancestry. While this correlation reduces the amount of genotyping to perform, it can vary across the human genome and across different populations. As the success of genome-wide association studies in diverse populations leverage on assaying genetic variants in sufficiently strong linkage disequilibrium with the causal variants, how well association signals are transferable across diverse populations thus depends on how representative are the HapMap populations for the other populations.

Singapore is a small and young country with a migratory history predominantly consisting of immigrants with Chinese, Malay (indigenous) and Indian genetic ancestries from neighbouring countries such as China, Indonesia, Malaysia and India. As of 2010, there are 3.2 million citizens or permanent residents living in Singapore with the following ethnic composition, 74.1% Chinese, 13.4% Malays and 9.2% Indians[132]. The Chinese community consists mainly of descendents of early Han Chinese settlers from the southern provinces of China, such as Fujian and Guangdong as determined from their dialect groups (linguistic properties)[133]. Malays are indigenous to Singapore, with migration of Malays from the Peninsular Malaysia as well as Javanese and Boyanese people from Indonesia. Cultural and religious similarities between the indigenous and immigrant Malays have resulted in intermarriages and the descendents from these marriages are now collectively known as Malays. Indians in Singapore comprise of people with paternal ancestries tracing back to the India subcontinent. The British colonization of Singapore had

brought Indian migrants from the India subcontinent, with the majority consisting of Telugas and Tamils from southeastern India and a minority of Sikhs and Pathans from north India.

To embark on studies to identify genetic determinants of diseases within these three ethnic groups residing in Singapore, it would be essential to understand the genetic diversity of these populations and the extent of global diversity. The Singapore Genome Variation Project (SGVP) was thus initiated as a dense genetic resource that lay the foundation of genetic association studies in three ethnic groups residing in Singapore. Results presented are based on the post quality control set of 96 Singapore Chinese (CHS), 89 Singapore Malays (MAS) and 83 Singapore Indians (INS), combined over two genotyping arrays, Affymetrix 6.0 and Illumina1M.

4.2. Population structure

The principal components analysis was applied to post-QC data (see Chapter 3) from the three ethnic groups in Singapore, with various reference panels such as the HapMap and a catalog of Indian individuals collected by Reich and colleagues[52] (Figure 9).

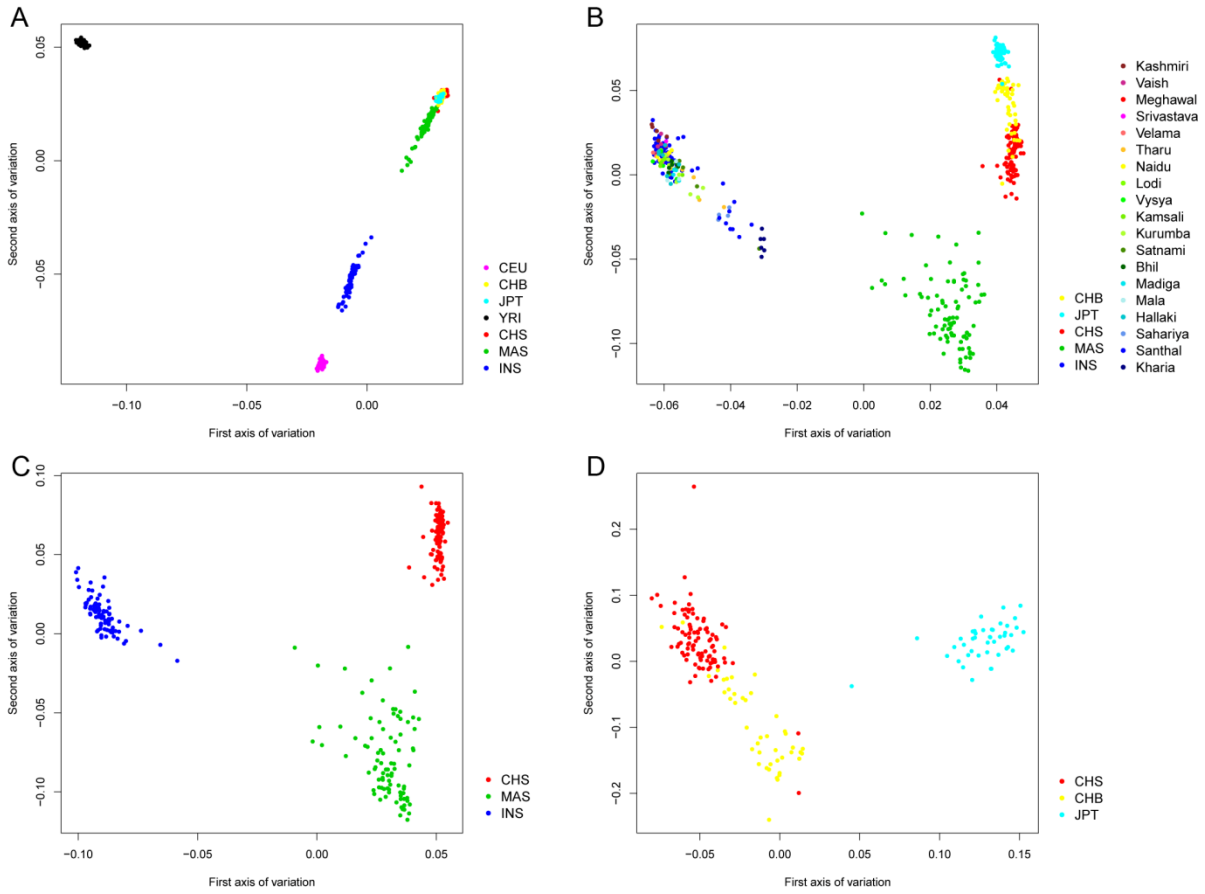


Figure 9. Principal components analysis maps of A) HapMap II and SGVP populations; B) Asia panels of HapMap II (CHB and JPT), SGVP and 19 diverse groups in India[52]; C) SGVP populations and D) Asia panels of HapMap II (CHB and JPT) with SGVP CHS. All plots show the second axis of variation against the first axis of variation (originally from reference 115).

First, we looked at the global diversity of the SGVP population superimposed on the global genetic diversity map, HapMap II including populations from European, African and Asian ancestries (Figure 9A). The first axis of variation at the global level effectively distinguished the African samples from the rest of the samples, while the second axis of variation addressed the differences between the European and Asian ancestries. Not surprisingly, at the global diversity scale, the Chinese in Singapore CHS were not distinguishable from the Han Chinese from Beijing CHB and Japanese from Tokyo JPT, which we collectively refer to as the Far East Asian cluster. The Malays in Singapore MAS were observed to be highly similar to the Far East Asian cluster,

although there appeared to be some degree of admixture, likely due to inter-marriages and migratory history. The Indians in Singapore also showed some admixture and were genetically closer to the CEPH with European ancestry CEU than to the Far East Asian cluster.

Together with HapMap Asian panels (CHB and JPT) and 98 samples from 19 diverse groups of Indian ancestry[52], the Asian diversity map is presented in Figure 9B. The first axis of variation distinguished the Indians from the India subcontinent (including INS) from the Far East Asian cluster comprising of the two Chinese populations (CHB and CHS), Japanese (JPT) and Malays (MAS). Within the Asian diversity, the second axis of variation appeared to coincide with a north-south cline that almost reflected their respective geographical locations on the physical world map. The Han Chinese CHB and Japanese JPT appeared to be reasonably distinct from each other. The Singapore Chinese CHS were closer to the Han Chinese in Beijing CHB but there appeared to be a cline, confirming the Northern and Southern cline of the Chinese and consistent with the migratory history of Singapore. The 98 samples from Reich and colleagues[52] were also separated along a north-south cline, with warmer colors representing those with a higher proportion of the Ancestral North Indians ANI ancestry (who are genetically closer to the Middle Easterners, Central Asians and Europeans) and the cooler colors representing those with a lesser proportion of the ANI ancestry. On the Asian diversity scale, Singapore Indians INS cluster with the 98 Indian samples from the India subcontinent.

Within Singapore, INS was more differentiated compared to CHS (Fst = 3.9%) and MAS (Fst = 2.7%), in contrast to the Fst = 0.6% between CHS and MAS (Figure 9C). Finally, comparing the Far East Asian cluster of CHB, JPT and CHS, there was a clear separation between CHB and CHS, although there were samples from CHB clustering together with CHS and vice versa

(Figure 9D). The Japanese JPT were more differentiated compared to CHB (Fst = 0.3%) and CHS (Fst = 0.4%) while Fst between CHB and CHS was 0.2%.

4.3. SNP and haplotype diversity and variation in linkage disequilibrium

The continuous axes of genetic variation from genome-wide principal components analyses suggested that the three ethnic groups residing in Singapore are pretty diverse, especially the Malays and Indians. We next look at how that diversity affects linkage equilibrium structure within these populations and compared against the HapMap, as these will have implications on the genetic association studies carried out in these three populations.

At the SNP level, comparing allele frequencies across pairwise panels using heatmaps in bins of 0.05, there was less variation in the allelic spectrum between CHS and MAS (Figure 10A), compared with CHS and INS (Figure 10B) and MAS with INS (Figure 10C). The pairwise allelic spectrums appeared almost symmetric between CHS and HapMap CHB (Figure 10D) and between CHS and MAS, indicating a greater degree of genetic similarity in these populations.

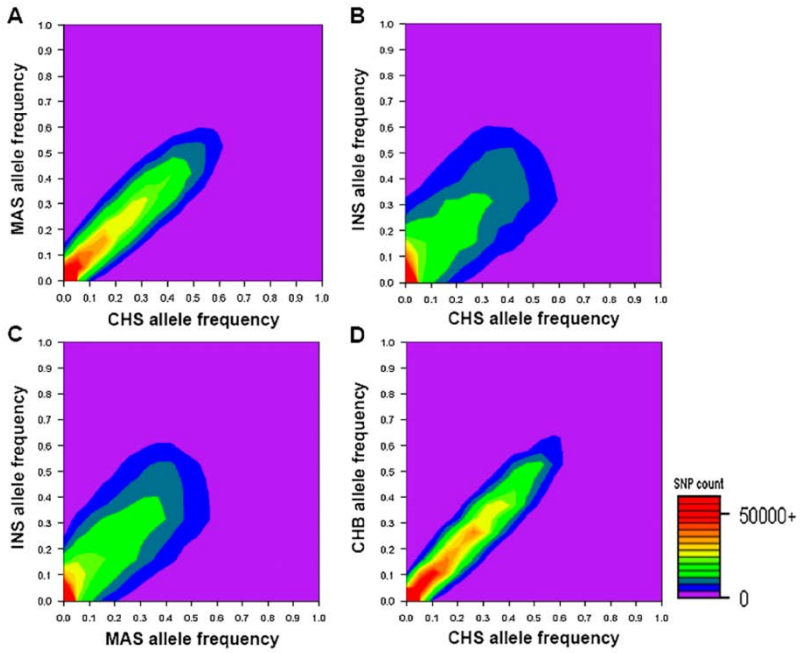


Figure 10. Allele frequency comparison between pairs of population: A) MAS against CHS; B) INS against CHS; C) INS against MAS; D) CHB against CHS. Each axis represents the allele frequencies for each population. For each SNP, the minor allele was defined across all the SGVP populations and subsequently the frequency of that allele was computed in each population. Twenty allele frequency bins each spanning 0.05 were constructed and the number of SNPs with MAF falling in each bin were tabulated/color-coded for each population (originally from reference 70).

Using the linkage disequilibrium metric r^2 computed on a single SNP basis across four HapMap populations and three SGVP populations, the degree of linkage disequilibrium decay was similar in the non-African populations, with linkage disequilibrium least conserved in the Africans (Figure 11).

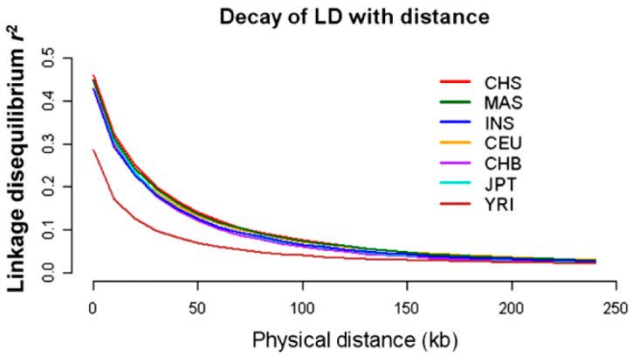


Figure 11. Decay of linkage disequilibrium with physical distance (kb) measured by r^2 with increasing distance up to 250kb for each of the HapMap and SGVP populations. 90 chromosomes were selected from each of the populations and only SNPs with MAF ≥ 5% were considered (originally from reference 70).

Next we looked at the haplotype diversity in 22 unlinked regions of 500kb long from each of the autosomal chromosomes in the same populations, thinned to the same SNP density spanning an average of 174 SNPs (Figure 12). For each population, we counted the number of distinct haplotypes in the region and the proportion of chromosomes accounted for by the corresponding distinct haplotype. The figure below showed the number of distinct haplotypes against the percentage of chromosomes considered for each of the HapMap and SGVP populations. Looking across the plot horizontally, 12 distinct haplotypes only accounted for slightly over 40% of the chromosomes in the YRI, the lowest across all populations. There was considerably higher haplotype diversity in YRI when compared with the rest, while the populations with Far East Asian ancestries (CHB, JPT and CHS) had the lowest haplotype diversity. There are a few factors affecting the haplotype diversity in these populations. The African population YRI had shorter linkage disequilibrium (also seen from the decay of linkage disequilibrium over distance) and hence more diversity in their haplotypes compared to the rest of the populations. Populations which are more heterogeneous would also likely have higher haplotype diversity. For instance, among the SGVP populations, INS had the greatest haplotype diversity, with 12 haplotypes

accounting for 57% of the INS chromosomes. This was followed by MAS, with 68% of the chromosomes accounted for by 12 haplotypes and 79% in the CHS.

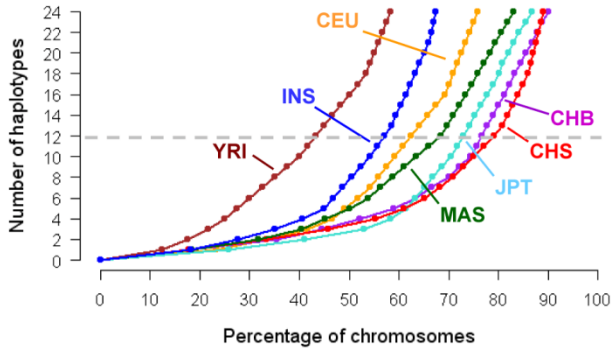


Figure 12. The plot showed the percentage of chromosomes that could be accounted for by the corresponding number of distinct haplotypes on the y-axis, over 22 unlinked regions of 500kb from each of the autosomal chromosomes (originally from reference 70).

In assessing linkage disequilibrium differences, heatmaps have been one of the conventional ways of visualizing differences in linkage disequilibrium across populations. These traditional heatmaps are difficult to assess since it is not easy to discern the significance of different linkage disequilibrium blocks across populations. We thus look at variation in linkage disequilibrium across pairs of populations using the varLD method, quantifying regional linkage disequilibrium differences.

A region that showed considerable signals of linkage disequilibrium variation from multiple pairs of populations coincided with reported association signals from genome-wide association analyses spanned the CDK5 regulatory subunit associated protein 1-like 1 (*CDKAL1*) gene on chromosome 6 (Figure 13). Our analysis indicated that the implicated index variant rs7754840 was found in a region with extensive linkage disequilibrium differences between populations and yet, associated with Type 2 Diabetes in many European[12,134-137] and Asian[138-149] populations. This would have implications on replication of the signals across populations and it is likely that this

index variant is in sufficient linkage disequilibrium with the causal variant(s) across all the populations. We will re-visit this locus with the Type 2 Diabetes data in Study 2.

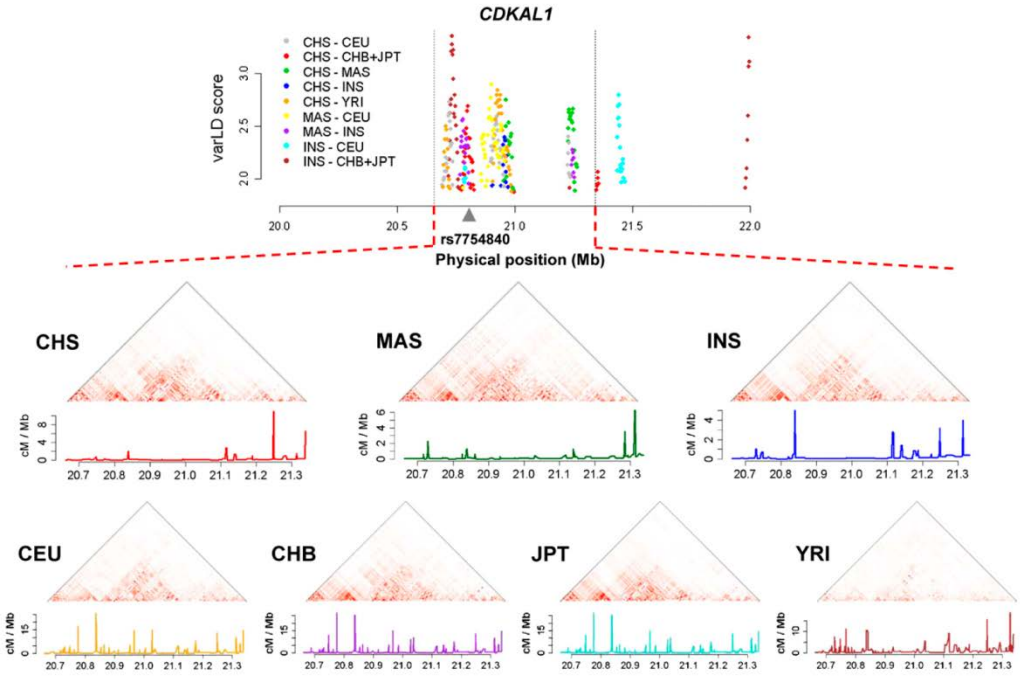


Figure 13. Variation in linkage disequilibrium scores at the *CDKAL1* locus, with r^2 heatmaps and population specific recombination rates (originally from reference 70).

Linkage differences across populations between the index SNPs and the causal variant is one of the reasons behind the failure to replicate association signals across populations. In an East Asian meta-analysis of blood pressure in 19,608 individuals[150], we looked up 13 established European implicated index SNPs from the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) with 29,136 individuals of European descent[151] and Global Blood Pressure Genetics (Global BPgen) with 34,333 individuals of European descent[152]. Of the 13 index variants, 7 of those replicated at $p < 0.05$ with consistent direction as the European results. We noted the smaller sample size in the East Asian meta-analysis compared to the European meta-analyses. In addition, we also applied varLD to look for evidences of LD differences at these 13 loci. Using the HapMap CEU and JPT+CHB populations, the following figure showed

the standardized varLD scores over a 200kb region centered on the index SNP, and a standardized varLD score of 2 indicated the empirical threshold across the genome. For the seven reported variants that showed an association with blood pressure in East Asians, there was limited evidence of linkage disequilibrium differences between European and East Asian populations at 5 (of 7) loci (unshaded in Figure 14) that harbor these variants. On the other hand, for the 6 reported variants for which we did not detect an association with the index SNP in East Asians, we found significant differences in linkage disequilibrium between the ethnic groups at 4 (of 6) loci (shaded red in Figure 14) harboring these variants.

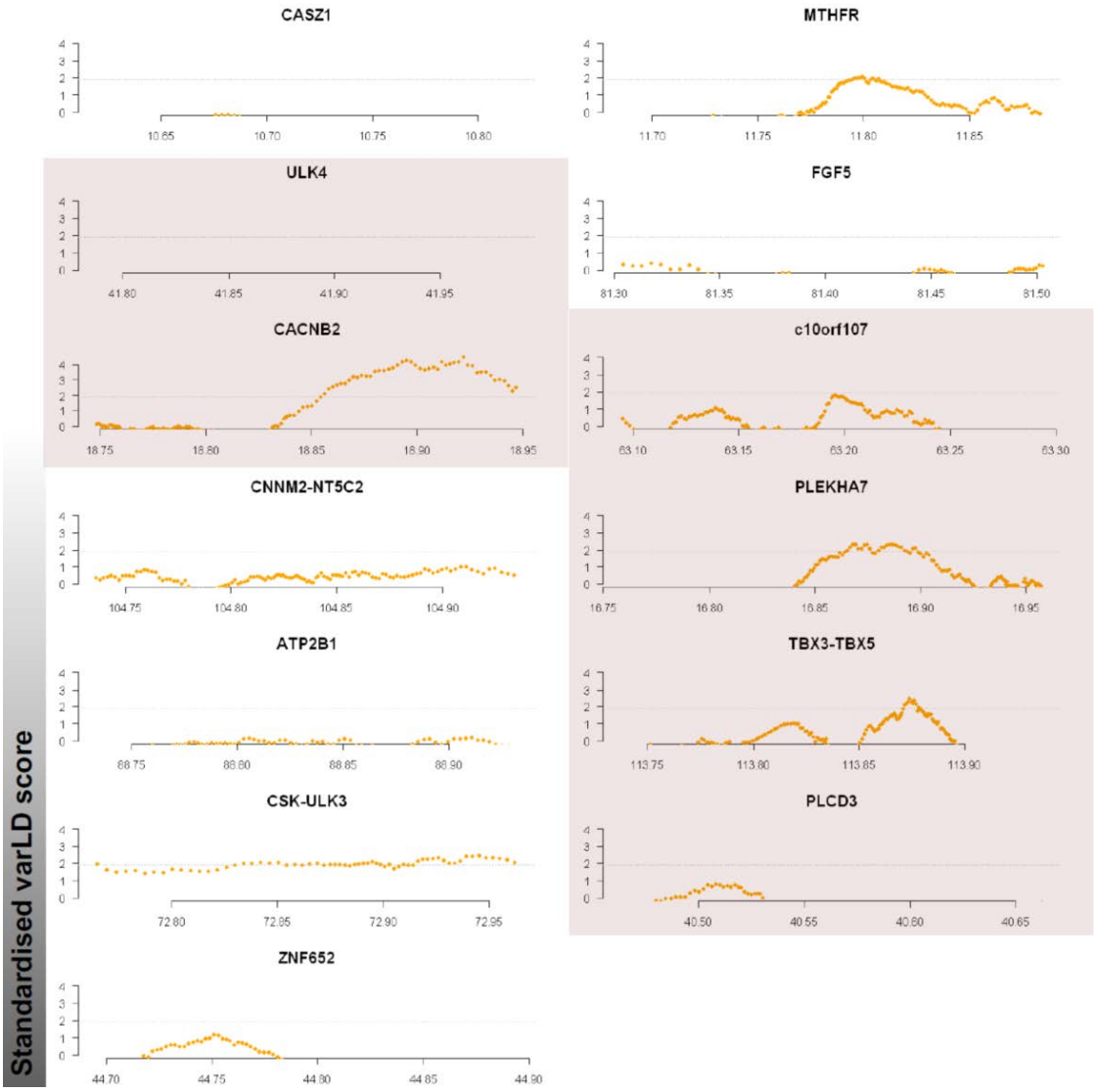


Figure 14. varLD assessment at 13 European established blood pressure loci, comparing HapMap CEU and JPT+CHB. Each plot illustrates the standardized varLD score (orange dotted circles) for 200kb region surrounding the index reported SNP. The horizontal gray dotted lines indicate the 5% empirical threshold at varLD score = 2 across the genome (originally from reference 150).

4.4. Signatures of positive selection

The availability of dense SNP data also allowed the survey of signatures of positive selection across the genome for these populations. Many of the selection signals found in the SGVP populations coincided with the selection signals found in HapMap[17,130]. Of these, many were well known to be implicated in alcohol dehydrogenase[153] (*ADH*) gene cluster and skin pigmentation[154-

[156] (solute carrier family 24, member 5 *SLC24A5* in INS; oculocutaneous albinism II *OCA2* in CHS and MAS; tyrosinase-related protein 1 *TYRP1* in CHS and MAS; myosin VA (heavy chain 12, myoxin) *MYO5A* in CHS, MAS and INS). Interestingly, one of top candidate selected regions in MAS was the kinesin family member 11/ hematopoietically expressed homeobox (*KIF11*/*HHEX)* locus that was implicated in the risk of Type 2 Diabetes in Europeans[12,134-137] and Asian[138,141,144,146,149] populations including Chinese and Japanese but not in the Malays (Table 5). We further looked at this locus in HapMap CEU and Type 2 Diabetic controls from SP2, SiMES and SINDI by identifying the most common haplotype in the region[126]. Using the index SNP with the highest iHS score in the Malays, the longest common haplotypes across the four populations were showed in Figure 15. Both the Chinese and Malays controls haplotypes carry the non-risk allele of the Type 2 Diabetes implicated SNP rs1111875. The selection signal appeared to be originating 0.5Mb downstream of *HHEX* near *CYP26A1*, suggesting possible genetic hitchhiking at this locus.

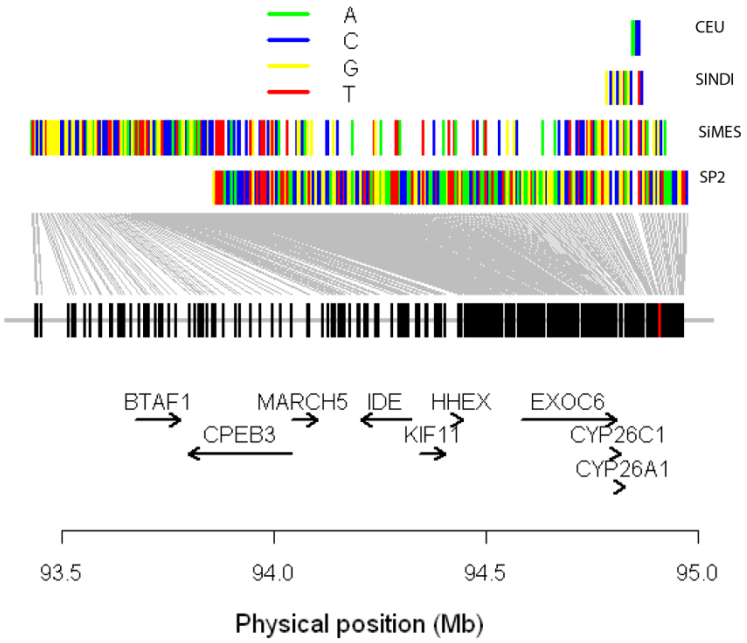


Figure 15. Visual representation of the haplotypes in Type 2 Diabetes controls of the Chinese (SP2), Malay (SiMES) and Indian (SINDI) cohorts and HapMap CEU.

Table 5. Top ten candidate regions of recent positive natural selection from the integrated haplotype score and if it had been previously observed in HapMap[18] (originally from 70).

Chr	Bin start	Bin end	Genes in bin	In HapMap
CHS				
2	17,400,000	17,800,000	*VSNL1, SMC6, GEN1*	--
2	25,800,000	26,400,000	*ASXL2, KIF3C, HADHA, RAB10, HADHB, GPR113*	--
2	108,300,000	108,500,000	*SULT1C2, GCC2*	Yes
2	125,200,000	126,100,000	*CNTNAP5*	--
2	197,300,000	197,500,000	*GTF3C3, PGAP1*	Yes
3	108,900,000	109,200,000	*BBX*	--
4	143,700,000	144,600,000	*USP38, GAB1*	Yes
7	5,400,000	5,800,000	*KIAA1856, FBXL18, ACTB, FSCN1, TRIAD3*	--
10	107,200,000	107,500,000	--	Yes
12	1,100,000	1,400,000	*ERC1*	--
MAS				
1	153,100,000	153,400,000	*PMVK, PBXIP1, PYGO2, SHC1, CKS1B, FLAD1, LENEP, ZBTB7B, DCST2, ADAM15, DCST1, EFNA4, EFNA3, EFNA1, RAG1AP1, DPM3*	--
2	84,300,000	84,900,000	*SUCLG1*	--
3	108,600,000	109,200,000	*BBX*	--
5	117,400,000	117,900,000	--	--
8	67,000,000	67,100,000	--	--
10	94,400,000	95,100,000	*KIF11, HHEX, EXOC6, CYP26A1, CYP26C1, FER1L3*	Yes
11	25,100,000	25,600,000	--	Yes
12	87,000,000	87,600,000	*CEP290, TMTC3, KITLG*	
15	61,600,000	62,600,000	*USP3, FBXL22, HERC1, DAPK2, FAM96A, SNX1, SNX22, PPIB, CSNK1G1, TRIP4, ZNF609*	Yes
17	25,000,000	25,500,000	*SSH2, EFCAB5, CCDC55*	--
INS				
2	82,800,000	83,100,000	--	--
2	96,300,000	97,100,000	*SNRNP200, NCAPH, ITRIPL1, NEURL3, ARID5A, FER1L5, CNNM4, CNNM3, SEMA4C, ANKRD23, ANKRD39, FAM178B*	--
4	29,100,000	30,000,000	--	
4	32,900,000	34,200,000	--	Yes
4	41,500,000	41,900,000	*TMEM33, WDR21B, SLC30A9, CCDC4*	Yes
7	119,500,000	120,300,000	*KCND2, TSPAN12*	--
8	42,600,000	42,800,000	*CHRNB3, CHRNA6*	--
11	60,600,000	61,000,000	*CD5, VPS37C, PGA3, PGA4, PGA5, VWCE, DOB1, DAK, CYBASC3, FLJ12529, C11orf79*	--
16	30,800,000	31,100,000	*CTF1, FBXL19, ORAI3, SETD1A, STX4, BCKDK, HSD3B7, STX1B2, ZNF668, ZNF646, VKORC1, PRSS8, TRIM72, PRSS36, MYST1, FUS*	--
17	24,900,000	25,900,000	*TP53I13, GIT1, ANKRD13B, CORO6, SSH2, EFCAB5, CCDC55, SLC6A4, BLMH, TMIGD1, CPD, GOSR1*	--

4.5. Summary

We now have a resource map of the genetic diversity of major ethnic groups in Singapore, which formed the basis of understanding ethnic differences in linkage equilibrium variations. Given a shared causal variant, there exist different index SNPs associated with disease identified in different populations. Replication and meta-analysis that rely on the transferability of the index implicated SNPs across populations will not be straightforward. Study 2 will explore these issues in greater details in the same populations, looking at Type 2 Diabetes in Chinese, Malays and Indians.

Key findings from Study 1:

I. Malays and Indians in Singapore showed some degree of admixture due to inter-marriages and migratory patterns. Chinese in Singapore formed a North-South cline with the HapMap Han Chinese from Beijing (CHB), again consistent with the migratory history of Singapore Chinese.

II. Populations which are more heterogeneous would have higher haplotype diversity. Within Singpaore, the Indians had the highest haplotype diversity, followed by the Malays and lastly the Chinese.

III. Genome-wide survey of signatures of positive selection coincided with selection signals found in HapMap.

CHAPTER 5 – TRANSFERABILITY OF TYPE 2 DIABETES LOCI IN MULTI-ETHNIC COHORTS FROM ASIA

5.1. Motivation

Type 2 Diabetes mellitus is a major chronic disease that affects more than 30 million people worldwide. The greatest increase in Type 2 Diabetes prevalence in the coming years is likely to be from Asia. With aging populations coupled with a more affluent and sedentary lifestyle from urbanization, the incidence of Type 2 Diabetes is expected to rise in these populations which already have higher rates of insulin resistance and metabolic syndrome[74,157,158]. In addition, these adverse effects on health also meant that Type 2 Diabetes develops earlier in Asians compared to the Europeans. The social economic burden of these younger diabetics, thus with a longer disease duration, and their co-morbidities such as cardiovascular diseases, diabetic nephropathy, diabetic retinopathy become a problem that needs to be addressed.

Type 2 Diabetes has benefited enormously from the advent of genome-wide genetic association studies. Since the discovery of the *TCF7L2* locus from linkage studies, and *PPARG* and *KCNJ11* from candidate gene studies, there has been at least forty other loci identified from genome-wide association studies associated with Type 2 Diabetes[12,67,68,76,91,134-137,146,149,159-162]. There is limited information on the transferability of the established loci in populations of other ethnicity as most of these studies had been carried out in populations of European descent and in large number of samples[12,76].

It is important to carry out genome-wide search for Type 2 Diabetes susceptibility locus in multiple populations and evaluate the transferability of established loci across populations. A common strategy is to genotype only the index implicated variants in other independent populations. If the causal variant(s) is(are) common across populations, variation in linkage

disequilibrium between index variant and causal variant among populations, allele frequency and effect size differences might potentially mask real associations across populations or result in the failure to detect genuine associations.

With the multi-ethnic demography of Singapore, we performed three population-based Type 2 Diabetes case control genome-wide association studies in three ethnic groups: Chinese (2,010 cases and 1,945 controls), Malays (794 cases and 1,240 controls) and Indians (977 cases and 1,169 controls) (Figure 16). Details of the quality control process are described in Chapter 3. The Chinese were genotyped on a combination of Illumina610 and Illumina1M arrays while the Malays and Indians were entirely genotyped on Illumina610 arrays. Table 6 below showed the summary characteristics of the study samples in each population. The cases were generally older than the controls, especially for the Chinese. Including age as a covariate was not effective in adjusting for the confounding effect as it resulted in spurious association from the disparate age distributions. All subsequent analyses presented were without covariate adjustment. The Malays and Indians had a higher BMI than the Chinese. Stratification by BMI status will be further discussed in Study 4 (Chapter 7).

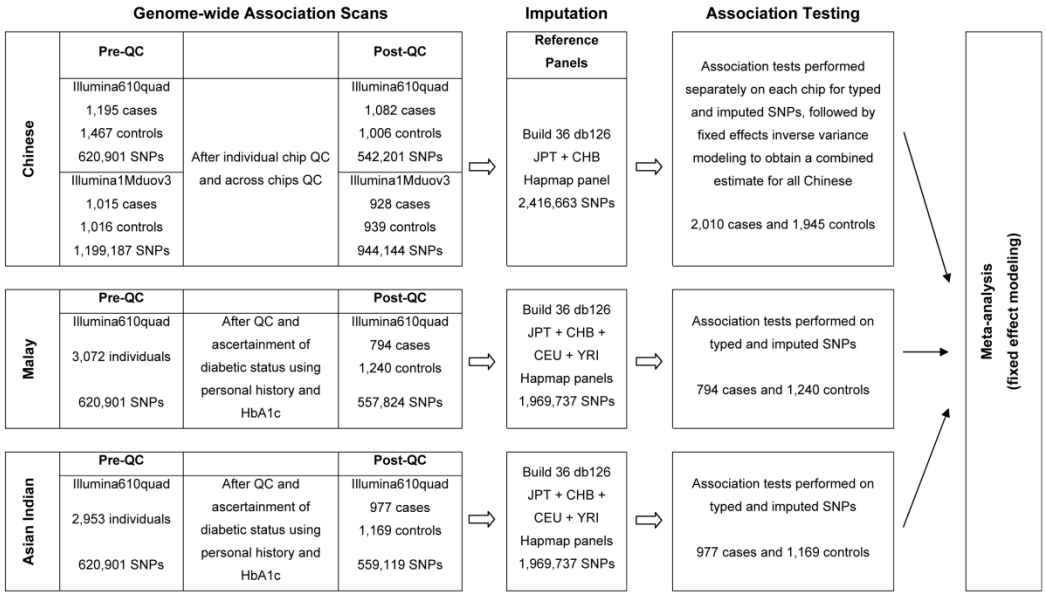


Figure 16. Diagram summarizing the study designs and analytical procedures for each of the genome-wide association studies (originally from reference 115)

Table 6. Summary characteristics of cases and controls stratified by their ethnic groups and genotyping arrays [originally from reference 115].

Characteristics	Chinese				Malay[a]		Asian Indian[a]	
	Illumina610quad		Illumina1Mduov3		Illumina610quad		Illumina610quad	
	Cases	Controls	Cases	Controls	Cases	Controls	Cases	Controls
N	1,082	1,006	928	939	794	1,240	977	1,169
Sex Ratio M/F (%)	402/680 (37.15/62.85)	217/789 (21.57/78.43)	602/326 (64.87/35.13)	599/340 (63.79/36.21)	405/389 (51.01/48.99)	645/595 (52.02/47.98)	531/466 (54.35/45.65)	566/603 (48.42/51.58)
Age[b] (yr)	65.07 (9.70)	47.69 (11.07)	63.67 (10.81)	46.74 (10.23)	62.27 (9.90)	56.89 (11.39)	60.71 (9.85)	55.73 (9.72)
Age at diagnosis[b] (yr)	55.65 (11.96)	--	52.15 (14.40)	--	54.35 (11.19)	--	51.35 (10.63)	--
Fasting glucose[b] (mmol/L)	--	4.67 (0.45)	--	4.73 (0.46)	--	--	--	--
HbA1C[b]	--	--	--	--	8.05 (1.84)	5.60 (0.30)	7.56 (1.52)	5.55 (0.28)
BMI[b] (kg/m^2)	25.27 (3.92)	22.30 (3.67)	25.42 (3.81)	22.84 (3.41)	27.82 (4.88)	25.13 (4.82)	27.06 (5.10)	25.33 (4.40)

[a] For Malay and Asian Indian samples, diabetic samples are defined as either with history of diabetes or HbA1c ≥ 6.5% while controls are defined as no history of diabetes and HbA1c < 6%.
[b] Mean(Standard Error).

5.2. Results from genome-wide scans

In the genome-wide scans by ethnic groups, we only identified a genome-wide significant SNP rs1048886 (*P*-value = 3.48 x 10^{-8}) on chromosome 6 in the Indians but it was not significant in the Chinese (*P*-value = 9.95 x 10^{-1}) nor Malays (*P*-value = 8.23 x 10^{-2}). This locus was also not found to be associated with Type 2 Diabetes in a larger South Asian meta-analysis at *P*-value < 10^{-4}in Study 3 (Chapter 6). It is highly likely that this locus was a false positive. Although no SNP showed genome-wide significance in the meta-analysis across the studies, we present the data for SNPs that showed suggestive evidence of association at *P*-value < 10^{-5}. Table 7 below showed the list of top regions from the fixed effects meta-analysis of the GWAS results across the ethnic groups. Among the top suggestive loci, two of the loci were well-established implicated loci for Type 2 Diabetes, *CDKAL1* and *HHEX*. Others include high mobility group 20A (*HMG20A*), zona pellucida-like domain containing 1 (*ZPLD1*) and hormonally upregulated Neu-associated kinase (*HUNK*) which showed no evidence of heterogeneity among the three ethnic groups and finally chromosome 6 open reading frame 57 (*C6orf57*) which was primarily driven by the Indians.

Table 7. Statistical evidence of the top regions (defined as $P < 10^{-5}$) that emerged from the fixed-effects meta-analysis of the GWAS results across Chinese, Malays and Asian Indians, with information on whether each SNP is a directly observed genotype (1) or is imputed (0). Combined minor allele frequencies of each index SNP is at least 5%. The I^2 statistic refers to the test of heterogeneity of the observed odds ratios for the risk allele in the three populations, and is expressed here as a percentage (originally from reference 115).

SNP	Chr	Pos (bp)	Nearest gene	Risk allele	Ref allele	Genotyped (1) or imputed (0)[a]	N	Chinese + Malays + Indians (3781 cases/4354 controls)			
								Risk allele frequency[b]	Fixed effects OR (95% CI)	Fixed effects *P*-value	I^2 (%)
rs7119	15	75564687	*HMG20A*	T	C	1111	8,135	0.188	1.24 (1.14-1.34)	5.24×10^{-7}	0
rs2063640	3	103685735	*ZPLD1*	A	C	1111	8,131	0.167	1.23 (1.13-1.34)	3.47×10^{-6}	0
rs2833610	21	32307057	*HUNK*	A	G	1111	8,127	0.567	1.17 (1.09-1.24)	3.90×10^{-6}	0
rs6583826	10	94337810	*KIF11*	G	A	1111	8,134	0.259	1.18 (1.10-1.27)	7.38×10^{-6}	0
rs1048886	6	71345910	*C6orf57*	G	A	1111	8,135	0.110	1.26 (1.14-1.39)	9.70×10^{-6}	85.40
rs9295474	6	20760696	*CKDAL1*	G	C	0000	8,079	0.357	1.16 (1.09-1.24)	8.59×10^{-6}	33.46

[a] This column shows whether each SNP is directly genotyped (1) or imputed (0) in each of the case control studies shown in Table 3. Each digit represents a case control study in the following order from left to right: Chinese on Illumina610, Chinese on Illumina1M, Malays on Illumina610 and Indians on Illumina610.
[b] Risk allele frequencies are sample size weighted frequencies across the three ethnic groups.

5.3. Evaluating transferability of known loci across populations

To evaluate the transferability of established disease implicated loci in our populations, we assessed the evidence of association at these loci in our populations, defining statistical significance as *P*-value < 0.05. For SNPs that were not directly observed, they were imputed with IMPUTEv2[58], which allowed the inclusion of population specific genotype information from SGVP to increase the haplotype diversity and reduce the chance of false negatives due to the use of an inappropriate reference panel. Among the 35 loci, only *KCNJ11* (Chinese *P*-value = 3.63 x 10^{-2} and Malay *P*-value = 2.26 x 10^{-2}), *CDKAL1* (Chinese *P*-value = 1.03 x 10^{-4} and Indian *P*-value = 3.60 x 10^{-2}) and *HHEX/IDE* (Chinese *P*-value = 2.79 x 10^{-2} and Indian *P*-value = 2.19 x 10^{-2}) showed evidence of association with Type 2 Diabetes in more than one population (Table 8). The majority of associations were present only in one population, namely, HNF1 homeobox B (*TCF2/HNF1B*), insulin-like growth factor 2 mRNA binding protein 2 (*IGF2BP2*), ArfGAP with RhoGAP domain, ankyrin repeat and PH domain 1 (*CENTD2*), C2 calcium-dependent domain containing 4B (*C2CD4A-C2CD4B*) and *FTO* in the Chinese; *KCNQ1* in the Malays and *TCF7L2*

and B-cell CLL/lymphoma 11A (*BCL11A*) in the Indians (Table 8). Finally, the meta-analysis across the three ethnic groups also exhibited associations at insulin receptor substrate 1 (*IRS1*) and solute carrier family 30 (zinc transporter), member 8 (*SLC30A8*).

As our study was underpowered with the relatively smaller sample sizes compared to the tens of thousands of samples used in the initial discovery of these loci, we sought to evaluate consistency in the direction of effects at index SNPs from each of the established loci. Many genome-wide scans tend to oversample affected relative to the proportion present in the general population. SNPs reaching genome-wide significance from the discovery studies tended to have effect size estimates that were upwardly biased[163]. This constitutes the winner's curse and subsequent replication efforts are dependent on the power in the original study. If the power to detect the original association is low, then the ascertainment effect on the replication efforts will be much more substantial[164]. Thus wherever possible, the published odds ratios (ORs) from combined discovery and replication stages were presented to avoid winner's curse. Figure 17 compared the published odds ratios from established Type 2 Diabetes implicated loci with the reported ORs from each of the individual genome-wide scans and meta-analysis. For the same allele that was associated with high risk in the original discovery studies, our ORs were generally consistent with what had been reported. Firstly, there was an over-representation of loci found to be associated with Type 2 Diabetes, where the number of nominally associated loci would be expected by chance under the null of *P*-value = 0.05 (Binomial test one-sided *P*-values: 2.85×10^{-4} for Chinese, 1.05×10^{-1} for Malays, 2.22×10^{-2} for Indians and 3.31×10^{-7} for meta-analysis). We observed that the same allele that conferred risk in the three populations were in accordance with the published results, with a two-sided binomial test for consistency of direction given by 5.92×10^{-3} for Chinese, 9.30×10^{-2} for Malays, 4.34×10^{-3} for Indians and 1.49×10^{-3} for meta-analysis. In addition, a greater proportion of SNPs displayed attenuated odds ratios in our populations

when compared to the effect sizes at each of the index SNPs, with two sided *P*-values given by Chinese: 5.22×10^{-2}; Malays: 3.47×10^{-2}; Indians: 8.55×10^{-4} and meta-analysis: 7.20×10^{-3}.

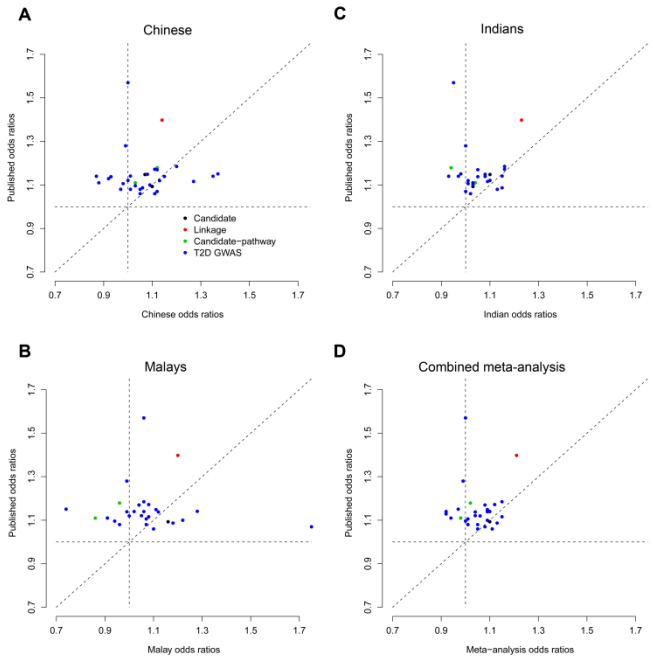


Figure 17. Bivariate plots comparing odds ratios established in populations of European ancestry against odds ratios observed in each of the ethnic groups (originally from reference 115).

Table 8. Known Type 2 Diabetes susceptibility loci tested for replication in three Singapore populations individually and combined meta-analysis. Published odds ratios (ORs) were obtained from European populations and correspond to the established ORs in **Figure 17**. Risk alleles were in accordance with previously established risk alleles. Information on whether each SNP was a directly observed genotype (1), or imputed (0) or not available for analysis (.) was presented in the table. Power (%) referred to the power for each of these individual studies to detect the published ORs at an α-level of 0.05, given the allele frequency and sample size for each study (originally from reference 115).

SNP	Chr	Pos (bp)	Nearest Gene	Risk allele	Ref allele	Pub-lished OR	Genotyped (1) Imputed (0) Not Available (.)[a]	Chinese (2010 cases/1945 controls)				Malays (794 cases/1240 controls)				Indians (977 cases/1169 controls)				Chinese + Malays + Indians (3781 cases/4354 controls)		
								Power	Risk allele freq	OR (95% CI)	*P*-value	Power	Risk allele freq	OR (95% CI)	*P*-value	Power	Risk allele freq	OR (95% CI)	*P*-value	Fixed effects OR (95% CI)	Fixed effects *P*-value	I^2 (%)
Identified Through Candidate Gene Study																						
rs1801282	3	12368125	*PPARG*	C	G	1.148	01.0	29	0.964	1.07 (0.84-1.35)	5.79×10^{-01}	--	--	--	--	39	0.889	1.10 (0.91-1.33)	3.22×10^{-01}	1.09 (0.94-1.26)	2.63×10^{-01}	--
rs5215	11	17365206	***KCNJ11***	C	T	1.093	1111	47	0.363	1.10 (1.01-1.21)	**3.63×10^{-02}**	27	0.401	1.16 (1.02-1.32)	**2.26×10^{-02}**	28	0.351	1.03 (0.91-1.16)	6.86×10^{-01}	1.10 (1.03-1.17)	**5.00×10^{-03}**	0
Identified Through Linkage Study																						
rs7903146	10	114748339	***TCF7L2***	T	C	1.398	1111	94	0.023	1.14 (0.84-1.53)	4.04×10^{-01}	70	0.043	1.20 (0.87-1.64)	2.62×10^{-01}	100	0.284	1.23 (1.08-1.40)	**2.10×10^{-03}**	1.21 (1.08-1.36)	**8.26×10^{-04}**	0
Identified Through Candidate Pathway Analysis																						
rs10010131	4	6343816	*WFS1*	G	A	1.11	0100	30	0.919	1.03 (0.88-1.21)	7.32×10^{-01}	26	0.839	0.86 (0.72-1.02)	8.17×10^{-02}	32	0.767	1.04 (0.90-1.20)	6.10×10^{-01}	0.98 (0.90-1.07)	6.88×10^{-01}	38.42
rs757210	17	33170628	*HNF1B (TCF2)*	T	C	1.179	1111	93	0.261	1.12 (1.01-1.24)	**2.51×10^{-02}**	70	0.338	0.96 (0.84-1.09)	4.99×10^{-01}	71	0.274	0.94 (0.82-1.07)	3.63×10^{-01}	1.02 (0.96-1.10)	4.80×10^{-01}	65.55
Identified Through Type 2 Diabetes GWAS																						
rs10923931	1	120319482	*NOTCH2*	T	G	1.138	0000	26	0.026	0.93 (0.71-1.22)	5.85×10^{-01}	24	0.054	1.12 (0.85-1.48)	4.21×10^{-01}	47	0.211	1.05 (0.91-1.22)	4.95×10^{-01}	1.04 (0.92-1.17)	5.17×10^{-01}	0
rs7578597	2	43586327	*THADA*	T	C	1.151	1111	29	0.995	1.37 (0.71-2.61)	3.45×10^{-01}	17	0.984	0.74 (0.45-1.23)	2.49×10^{-01}	39	0.879	0.98 (0.82-1.18)	8.27×10^{-01}	0.97 (0.82-1.15)	7.35×10^{-01}	7.07
rs243021	2	60438323	***BCL11A***	A	G	1.08	1111	38	0.669	1.05 (0.96-1.16)	2.87×10^{-01}	22	0.536	0.96 (0.85-1.09)	5.76×10^{-01}	24	0.482	1.13 (1.00-1.28)	4.75×10^{-02}	1.05 (0.98-1.12)	1.38×10^{-01}	36.97
rs2943641	2	226801989	***IRS1***	C	T	1.087	1111	22	0.931	1.06 (0.89-1.26)	5.29×10^{-01}	16	0.892	1.18 (0.96-1.44)	1.14×10^{-01}	21	0.805	1.15 (0.99-1.34)	6.64×10^{-02}	1.13 (1.02-1.24)	**1.92×10^{-02}**	0
rs6780569	3	23173478	*UBE2E2*	G	A	1.17	1111	81	0.817	1.12 (1.00-1.25)	5.97×10^{-02}	58	0.791	1.04 (0.89-1.22)	5.93×10^{-01}	66	0.701	1.05 (0.92-1.20)	4.75×10^{-01}	1.08 (1.00-1.16)	5.63×10^{-02}	0
rs4607103	3	64686944	*ADAMTS9*	C	T	1.096	000.	53	0.661	1.03 (0.94-1.13)	5.24×10^{-01}	28	0.709	0.94 (0.82-1.09)	4.19×10^{-01}	--	--	--	--	1.00 (0.93-1.08)	9.29×10^{-01}	--
rs1470579	3	187011774	***IGF2BP2***	C	A	1.139	1111	77	0.255	1.15 (1.04-1.28)	**5.80×10^{-03}**	50	0.329	0.99 (0.86-1.13)	8.59×10^{-01}	57	0.469	1.08 (0.95-1.22)	2.26×10^{-01}	1.09 (1.02-1.16)	**1.59×10^{-02}**	39.17
rs4457053	5	76460705	*ZBED3*	G	A	1.08	00..	18	0.055	1.01 (0.84-1.22)	9.31×10^{-01}	--	--	--	--	--	--	--	--	1.01 (0.84-1.22)	9.31×10^{-01}	--
rs7754840	6	20769229	***CDKAL1***	C	G	1.185	1111	97	0.369	1.20	**1.03×10^{-04}**	76	0.369	1.06	3.95×10^{-01}	71	0.245	1.16	**3.60×10^{-02}**	1.15	**2.34×10^{-05}**	12.52

										(1.09-1.31)				(0.93-1.21)				(1.01-1.34)		(1.08-1.23)		
rs864745	7	28147081	*JAZF1*	T	C	1.121	0000	60	0.785	1.00 (0.89-1.11)	9.36 x 10^{-01}	34	0.758	1.05 (0.91-1.22)	4.93 x 10^{-01}	37	0.752	1.10 (0.95-1.26)	2.07 x 10^{-01}	1.04 (0.96-1.12)	3.41 x 10^{-01}	0
rs972283	7	130117394	*KLF14*	G	A	1.07	….	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
rs896854	8	96029687	*TP53INP1*	T	C	1.06	1111	22	0.258	1.05 (0.95-1.16)	3.42 x 10^{-01}	13	0.29	1.10 (0.96-1.26)	1.80 x 10^{-01}	15	0.399	1.02 (0.90-1.15)	7.91 x 10^{-01}	1.05 (0.98-1.13)	1.47 x 10^{-01}	0
rs13266634	8	118253964	***SLC30A8***	C	T	1.149	1111	87	0.545	1.08 (0.98-1.18)	1.07 x 10^{-01}	58	0.573	1.11 (0.97-1.25)	1.20 x 10^{-01}	52	0.767	1.08 (0.94-1.25)	2.82 x 10^{-01}	1.09 (1.02-1.16)	**1.39 x 10^{-02}**	0
rs10811661	9	22124094	*CDKN2A/B*	T	C	1.191	….	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
rs13292136	9	81141948	*CHCHD9*	C	T	1.11	0000	30	0.913	0.88 (0.76-1.03)	1.25 x 10^{-01}	17	0.914	0.91 (0.73-1.13)	3.83 x 10^{-01}	24	0.855	1.03 (0.86-1.22)	7.64 x 10^{-01}	0.94 (0.85-1.04)	2.17 x 10^{-01}	0
rs17584499	9	8869118	*PTPRD*	T	C	1.57	1111	--	0.104	1.00 (0.87-1.16)	9.91 x 10^{-01}	--	0.227	1.06 (0.91-1.23)	4.43 x 10^{-01}	--	0.257	0.95 (0.83-1.09)	4.46 x 10^{-01}	1.00 (0.92-1.09)	9.78 x 10^{-01}	0
rs12779790	10	12368016	*CDC123/ CAMK1D*	G	A	1.092	….	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
rs1111875	10	94452862	***HHEX/IDE***	C	T	1.172	1111	90	0.3	1.11 (1.01-1.22)	**2.79 x 10^{-02}**	65	0.313	1.08 (0.94-1.23)	2.81 x 10^{-01}	71	0.381	1.16 (1.02-1.31)	**2.19 x 10^{-02}**	1.12 (1.05-1.19)	**1.09 x 10^{-03}**	0
rs2237892	11	2796327	***KCNQ1***	C	T	1.141	1111	80	0.686	1.01 (0.92-1.11)	8.71 x 10^{-01}	50	0.692	1.28 (1.12-1.47)	**3.42 x 10^{-04}**	16	0.976	1.15 (0.74-1.77)	5.33 x 10^{-01}	1.09 (1.01-1.18)	**2.40 x 10^{-02}**	75.43
rs231362	11	2648047	*KCNQ1*	G	A	1.08	000.	18	0.913	0.97 (0.83-1.13)	6.86 x 10^{-01}	14	0.856	1.07 (0.90-1.28)	4.51 x 10^{-01}	--	--	--	--	1.01 (0.90-1.14)	8.47 x 10^{-01}	--
rs1552224	11	72110746	***CENTD2***	A	C	1.14	1111	43	0.937	1.35 (1.12-1.63)	**1.38 x 10^{-03}**	24	0.923	1.06 (0.84-1.34)	6.23 x 10^{-01}	41	0.818	0.97 (0.83-1.14)	6.97X 10^{-01}	1.10 (0.99-1.23)	6.89 x 10^{-02}	72.68
rs10830963	11	92348358	*MTNR1B*	G	C	1.129	.1..	77	0.426	0.92 (0.81-1.05)	1.98 x 10^{-01}	--	--	--	--	--	--	--	--	0.92 (0.81-1.05)	1.99 x 10^{-01}	--
rs1531343	12	64461161	*HMGA2*	C	G	1.1	0000	34	0.104	1.09 (0.95-1.27)	2.26 x 10^{-01}	15	0.076	1.22 (0.96-1.55)	1.07 x 10^{-01}	24	0.187	1.03 (0.89-1.21)	6.72X 10^{-01}	1.09 (0.99-1.20)	8.49 x 10^{-02}	0
rs7961581	12	69949369	*TSPAN8/ LGR5*	C	T	1.106	0000	53	0.222	0.98 (0.88-1.09)	7.42 x 10^{-01}	30	0.219	1.07 (0.92-1.24)	3.93 x 10^{-01}	37	0.346	1.01 (0.89-1.14)	9.33 x 10^{-01}	1.01 (0.94-1.08)	8.20 x 10^{-01}	0
rs7957197	12	119945069	*HNF1A*	T	A	1.07	….	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
rs11634397	15	78219277	*ZFAND6*	G	A	1.06	01..	12	0.075	1.11 (0.94-1.31)	2.33 x 10^{-01}	--	--	--	--	--	--	--	--	1.11 (0.94-1.31)	2.33 x 10^{-01}	--
rs7172432	15	60183671	***C2CD4A-C2CD4B***	A	G	1.12	1111	68	0.672	1.13 (1.03-1.24)	**1.06 x 10^{-02}**	40	0.677	1.00 (0.87-1.14)	9.79 x 10^{-01}	45	0.593	1.01 (0.89-1.14)	9.08 x 10^{-01}	1.06 (1.00-1.13)	6.72 x 10^{-02}	37.7
rs8042680	15	89322341	*PRC1*	A	C	1.07	1111	10	0.997	1.12 (0.48-2.65)	7.90 x 10^{-01}	8	0.970	1.75 (1.21-2.53)	**3.09 x 10^{-03}**	16	0.766	1.00 (0.87-1.16)	9.60 x 10^{-01}	1.08 (0.95-1.23)	2.52 x 10^{-01}	73.39
rs9939609	16	52378028	***FTO***	A	T	1.116	0000	45	0.132	1.27 (1.11-1.45)	**3.79 x 10^{-04}**	37	0.300	1.08 (0.94-1.24)	2.57 x 10^{-01}	43	0.329	1.09 (0.96-1.24)	1.96 x 10^{-01}	1.15 (1.06-1.24)	**5.06 x 10^{-04}**	43.05
rs391300	17	2163008	*SRR*	C	A	1.28	1111	100	0.655	0.99 (0.90-1.08)	7.68 x 10^{-01}	97	0.466	0.99 (0.87-1.12)	8.48 x 10^{-01}	98	0.440	1.00 (0.89-1.13)	9.96 x 10^{-01}	0.99 (0.93-1.06)	7.63 x 10^{-01}	0
rs10425678	19	38669236	***PEPD***	C	T	1.14	1111	73	0.211	0.87 (0.78-0.97)	9.98 x 10^{-03}	43	0.224	1.02 (0.87-1.19)	8.05 x 10^{-01}	54	0.331	0.93 (0.82-1.05)	2.51 x 10^{-01}	0.92 (0.85-0.99)	**2.32 x 10^{-02}**	30.4

[a] This column shows whether each SNP is directly genotyped (1) or imputed (0) in each of the case control studies shown in Table 3. Each digit represents a case control study in the following order from left to right: Chinese on Illumina610, Chinese on Illumina1M, Malays on Illumina610 and Indians on Illumina610.

5.4. Power and related issues

While there was evidence of over-representation of association signals in a consistent direction in these Asian populations, we failed to observe statistically significant associations for a number of well-established loci, mostly in European populations. Our sample sizes were smaller compared to the large scale meta-analyses that discovered these variants. While meta-analysis boosted the sample size, the genetic heterogeneity across these populations implied that we were likely to detect those variants which showed little evidence of heterogeneity in meta-analysis, with similar effect sizes across the populations (Table 8). We were able to detect association at genetic variants from the earliest wave of the GWAS which tended to have smaller sample sizes and larger ORs, including *CDKAL1*, *KIF11/HHEX*, *IGFBP2*, *SLC30A8* and *FTO*[134-137]. The smaller effect sizes observed in our populations which affect the power of the study may explain why we fail to detect an association at the other variants.

TCF7L2 is by far the genetic variant that harbors the largest effect size for Type 2 Diabetes. However, we had not been able to observe any significant association in our Chinese and Malay populations. For the index SNP, rs7903146, the risk allele frequencies were less than 5% in the Chinese and Malays. However, the risk allele frequency of the Indians at this variant was 28.5%, improving our power to detect an association (P-value = 2.10×10^{-3}). In a recent large scale Type 2 Diabetes study conducted in Japanese, with the risk allele at 3.5% in 5,629 cases and 6,406 controls, they were able to detect an association at this variant with OR of 1.54 reaching genome-wide significance[162].

5.5. Allelic heterogeneity

We further looked at *CKDAL1*, a well-established Type 2 Diabetes implicated locus, across these three genetically heterogeneous populations. In Figure 18 below, the left panel shows regional

association plots of the index variant (from meta-analysis of the three populations) associated with Type 2 Diabetes in our populations. An association in this locus was observed only in the Chinese and Indians. The right panel shows the same locus, conditioned on rs7754840, the index SNP indentified in populations of European descent[134,135]. For the Chinese, we note that the conditional analysis (conditioned on rs7754840) effectively removed the association at this locus, suggesting that the observed association might be attributed to the same variant giving rise to an association in the European populations. On the other hand, in the Indians, conditional analysis attenuated signals near the index SNP but boosted signals upstream. It is possible that the European index SNP rs7754840 is a poor surrogate for the same causal variant in the Indians or there could be more than one casual variant in this region, i.e., allelic heterogeneity. Thus the European index SNP do not entirely account for the association signals observed at this locus in the Indians.

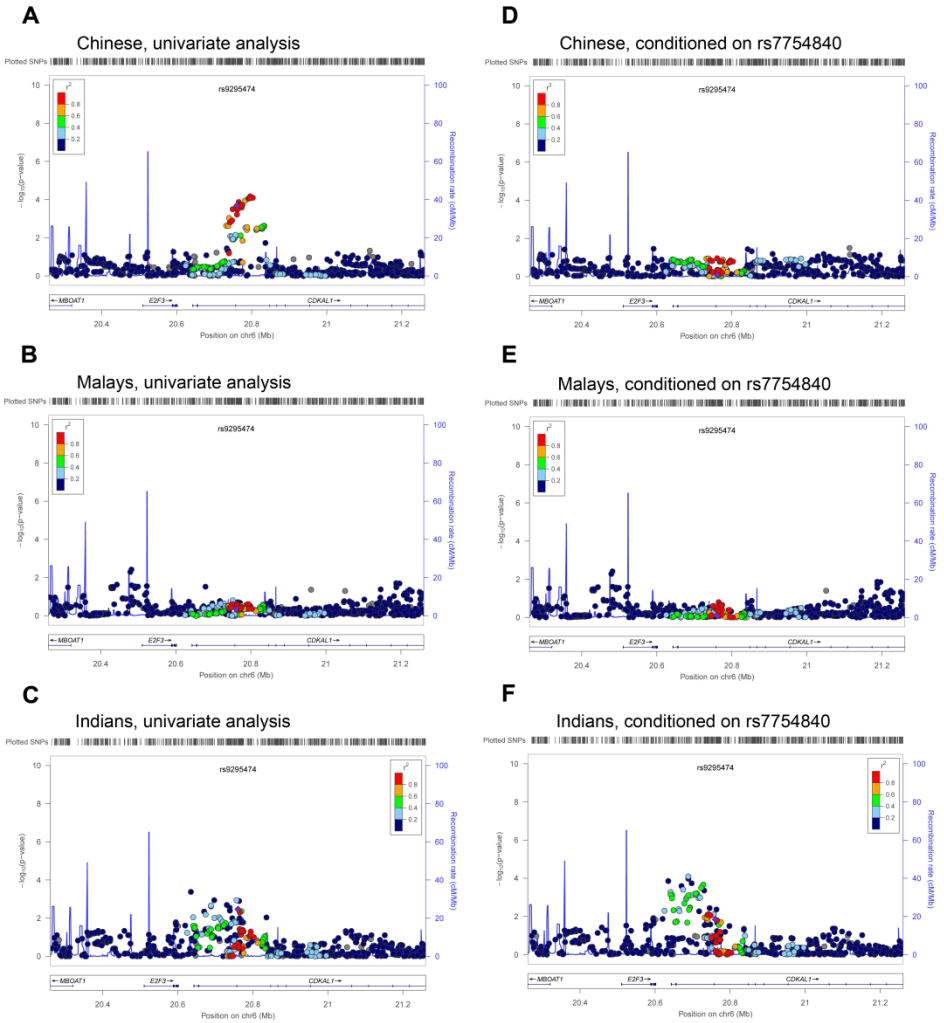


Figure 18. Regional association plots of the index SNP in *CDKAL1*. The left column of panels showed the univariate analysis while the right column of panels showed conditional analysis on the index SNP rs7754840 that was established in the Europeans. In each panel, the index SNP was represented by a purple diamond and the surrounding SNPs coloured based on their r^2 with the index SNP from the HapMap CHB+JPT reference panel. Estimated recombination rates reflect the local linkage disequilibrium structure in the 500kb buffer and gene annotations were obtained from the RefSeq track of the UCSC Gene Browser (refer to LocusZoom http://csg.sph.umich.edu/locuszoom/ for more details) (originally from reference 115).

The *KCNQ1* gene is another Type 2 Diabetes implicated locus exhibiting allelic heterogeneity between populations of European and Asian descent[12]. This regional plot (Figure 19) shows the association results centered at the *KCNQ1* gene, with the association signals in each of the ethnic groups distinguished by color (Chinese – red, Malay – green and Indians – blue). Also plotted

were two index SNPs, in purple alphabets representing each of the three ethnic groups. One of the index SNPs was rs2237892 which was found associated with Type 2 Diabetes in 6,800 case control pairs from Japanese, Korean and Chinese populations in 2008[67]. We also noted the newly established variant rs231362 from DIAGRAM+ that was 200kb upstream of the index variant of the Asian studies, which did not show any association in the Chinese and Malays, while the same index SNP was not available for Indians[12].

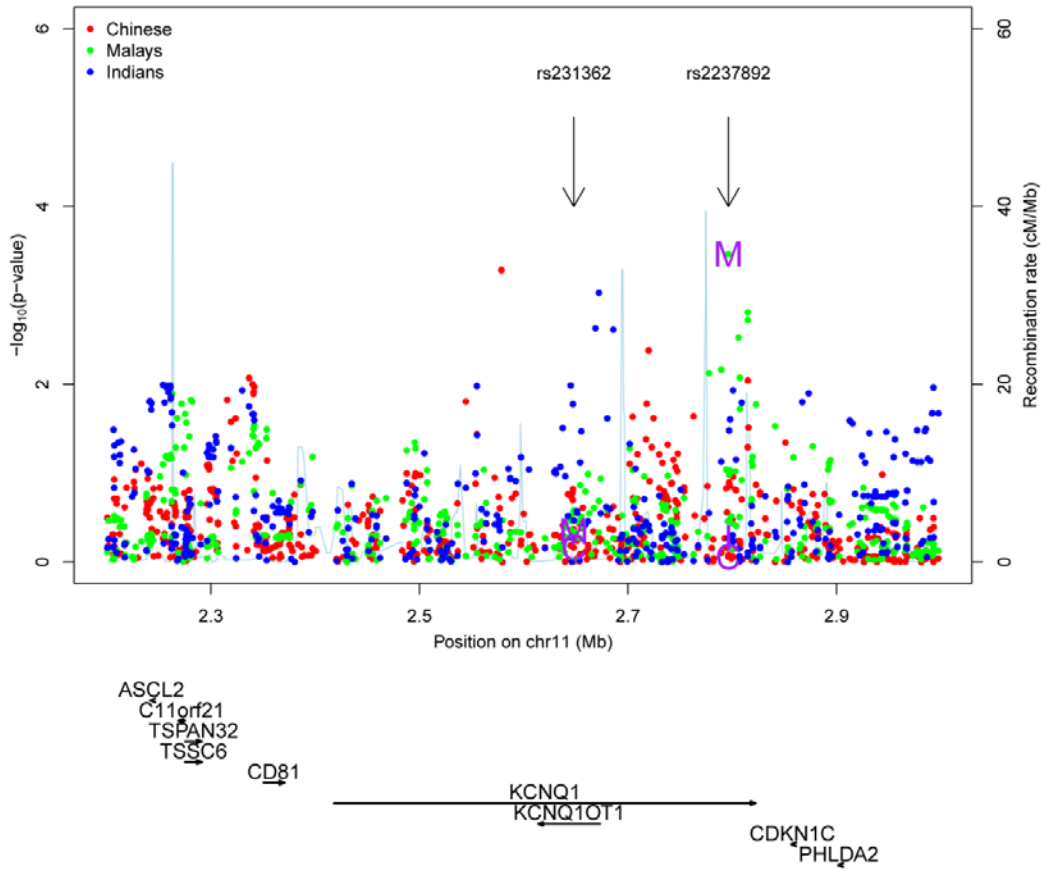


Figure 19. Regional association plots around the *KCNQ1* gene. The three ethnic groups are represented by three separate colors, red: Chinese, green: Malays and blue: Indians. Two index SNPs rs231362 and rs2237892 are plotted in purple and indicated by the first alphabet of the three ethnic groups. Note that rs231362 is not available for the Indians.

5.6. Summary

Despite the limited power of our studies, we showed that, among the associated variants established in the European populations for Type 2 Diabetes, there were more statistically significant associations detected in our studies that would be expected by chance, with consistency in direction of effects. Therefore we demonstrated that many of the variants established in the European populations were likely to be relevant in these Asian populations. We also discussed the possible reasons for failing to replicate association signals across populations: (i) power; (ii) variation in linkage disequilibrium and (iii) allelic heterogeneity. From the *KCNQ1* example, there is potential in non-European populations to detect novel loci, which might be common across populations or population specific, possibly due to different allelic frequencies or attributable to differences in environmental modifiers. In Study 3, which is in the next chapter, we look at a meta-analysis across three Type 2 Diabetes studies from individuals of South Asian ancestry.

Key findings from Study 2:

I. The individual population genome-wide association studies and meta-analyses failed to detect any new genome-wide association with Type 2 Diabetes.
II. There was an over-representation of established Type 2 Diabetes loci in the meta-analysis that would be expected by chance, with the same risk allele conferring risk across populations, indicating the presence of shared causal variants across world-wide populations.
III. Similarly, there were established Type 2 Diabetes loci where we failed to replicate in the Singapore populations. Possible reasons include the lack of power in these smaller studies, variation in linkage disequilibrium between populations at index SNPs and allelic heterogeneity at implicated loci.

CHAPTER 6 – GENOME-WIDE ASSOCIATION STUDY IDENTIFIES SIX TYPE 2 DIABETES LOCI IN INDIVIDUALS OF SOUTH ASIAN ANCESTRY

6.1. Motivation

South Asian Indians have one of the highest rates of Type 2 Diabetes in the world[74,165,166] and the number of South Asians affected by this chronic disease is projected to reach 80 million by 2030[166]. The prevalence differs across India, with prevalence lowest for non-obese, physically active Indians living in the rural regions and highest in the most urbanized states, where people tend to have a more sedentary lifestyle and are more likely to be obese[165]. While urbanization bringing changes in lifestyle and diet is one main driving force behind the increase in incidence and prevalence of Type 2 Diabetes in South Asians Indians, studies showed that they have a higher predisposition to insulin resistance, greater degree of central obesity and more visceral fat for any BMI[167-171]. Migrant South Asian populations also have a higher prevalence of Type 2 Diabetes than other populations residing in the same region, confirming that Type 2 Diabetes has an important genetic heritable component[97,165].

We thus carried out a genome-wide meta-analysis of three populations of South Asian origins from the LOLIPOP, PROMIS and SINDI studies, to identify common genetic variants underlying risk of Type 2 Diabetes in South Asians (people originating from the Indian subcontinent including India, Pakistan, Sri Lanka and Bangladesh). Across the three studies, there were 5,561 Type 2 Diabetes cases and 14,558 controls (Table 9). We noted that two of these cohorts, LOLIPOP and PROMIS, had an over-representation of coronary artery disease (CAD) cases. Diabetes is a risk factor for coronary artery diseases. The genetic association of diabetes could be confounded by the presence of CAD cases if the same variant is also associated with CAD, though it is more likely that variants act through the diabetes pathway toward CAD progression. However, if the variants discovered in this study were attributable to the presence of CAD cases

rather than Type 2 Diabetes, the association will fail to replicate in the independent replication studies due to the lower prevalence of CAD cases in these studies. Recently, LOLIPOP and PROMIS, together with two other European studies, identified five new loci associated with CAD[93]. These loci were not known to be associated with Type 2 Diabetes risk.

Table 9. Summary characteristics of Stage 1 discovery populations [(originally from reference 109)].

	LOLIPOP 610		LOLIPOP 317		SINDI		PROMIS	
	T2D	**Controls**	**T2D**	**Controls**	**T2D**	**Controls**	**T2D**	**Controls**
N	1783	4773	440	1699	977	1169	2361	6817
Age	59.4 (9.2)	53.9 (10.7)	54.1 (10.1)	46.8 (10.1)	60.7 (9.9)	55.7 (9.7)	55.0 (9.4)	52.9 (10.5)
Gender (% male)	82.9	84.8	100.0	100.0	54.4	48.4	76.5	83.0
SBP (mmHg)	140. 6 (20.4)	133.5 (18.9)	139.8 (20.6)	132.0 (20.2)	140.0 (19.7)	131.7 (19.2)	129.9 (21.5)	127.1 (20.5)
DBP (mmHg)	80.7 (10.9)	82.4 (10.7)	84.0 (11.7)	82.5 (12.1)	77.1 (10.1)	77.2 (9.9)	81.6 (11.9)	80.8 (11.6)
Weight (kg)	78.1 (14.0)	75.6 (13.0)	80.0 (15.9)	78.4 (13.9)	70.9 (14.0)	66.7 (12.5)	71.0 (12.5)	69.1 (13.2)
BMI (kg/m^2)	28.1 (4.6)	26.8 (4.2)	27.6 (4.7)	26.6 (4.2)	27.1 (5.1)	25.3 (4.4)	26.0 (4.0)	25.3 (3.9)
Waist	100.8 (11.5)	96.6 (10.9)	100.0 (12.2)	96.3 (11.4)	--	--	92.0 (12.0)	90.1 (11.7)
Waist-hip ratio	0.99 (0.07)	0.95 (0.07)	0.99 (0.07)	0.95 (0.07)	--	--	0.95 (0.06)	0.94 (0.07)
Cholesterol (mmol/L)	4.65 (1.20)	5.21 (1.12)	4.94 (1.09)	5.46 (1.04)	4.86 (1.11)	5.36 (0.98)	4.81 (1.39)	4.74 (1.29)
HDL chol (mmol/L)	1.16 (0.30)	1.22 (0.30)	1.17 (0.29)	1.247 (0.31)	1.02 (0.30)	1.10 (0.31)	0.88 (0.25)	0.89 (0.25)
Triglycerides (mmol/L)	1.94 (1.55)	1.77 (1.12)	1.82 (0.86)	1.65 (0.82)	2.09 (1.22)	1.85 (1.13)	2.51 (1.58)	2.20 (1.36)
Glucose (mmol/L)	8.6 (3.1)	5.2 (0.6)	8.9 (2.9)	5.1 (0.6)	9.71 (4.44)	5.38 (1.06)	13.31 (5.47)	6.89 (2.91)
HbA1c (%)	7.9 (1.7)	5.63 (0.549)	8.0 (1.8)	5.5 (0.5)	7.6 (1.5)	5.5 (0.28)	8.9 (1.93)	5.8 (0.45)
Coronary heart disease (%)	63.0	35.1	3.4	0.5	22.7	9.3	60.8	46.6
Ever smoked (%)	27.4	21.1	28.0	28.5	28.3	26.4	49.4	55.7
Hypertension (%)	80.1	56.1	70.7	40.6	74.2	46.2	36.5	25

Mean (Standard Error).

6.2. Six new loci associated with Type 2 Diabetes in people of South Asian ancestry

The primary analysis combining gender-specific association analysis using fixed effects meta-analysis across LOLIPOP, SINDI and PROMIS study populations identified one variant on the *TCF7L2* locus that reached genome-wide significance. The lead SNP was rs7903146 (*P*-value = 2.8×10^{-19}), the same index SNP widely reported across European populations[12,134-137].

Six new common variants were found to be associated with Type 2 Diabetes in people of South Asian ancestry. Table 10 below illustrates the association results of the index SNPs across the discovery (stage 1), replication (stage 2) and the combined analysis of the two stages (Stage 1 + 2). Regional plots of the six loci were shown in Figure 20, with the local linkage disequilibrium structure of HapMap CEU. In the selection process of SNPs to Stage 2 replication from a set of 43 SNPs exhibiting significance between *P*-value greater than 10^{-5} and *P*-value less than or equal to 10^{-4}, SNPs with evidence of association in the Europeans[12] were prioritized. This was done to generate a more manageable list of SNPs to Stage 2 replication. South Asians are genetically more similar to Europeans and many common disease implicated variants are shared across population of different ancestries. While this strategy might mean that we would fail to detect South Asian specific variants implicated with Type 2 Diabetes, this maximizes the discovery of genetic loci that are also present in European populations. Association of these variants with secondary glycemic (fasting insulin, fasting glucose, Homeostatic Model Assessment of beta-cell function and insulin sensitivity) and metabolic traits were also investigated in LOLIPOP and PROMIS. The HOMA indices provide an estimation of the glucose regulation as a feedback loop. For instance, fasting insulin levels will be elevated in direct proportion to diminished insulin sensitivity (HOMA-sensitivity) and elevated fasting glucose level reflect a feedback mechanism that maintained fasting insulin levels when there is a reduced insulin secretion (HOMA-B)[172].

On chromosome 2, rs3923113 showed the strongest association with Type 2 Diabetes (Stage 1 *P*-value = 3.7×10^{-7}; Stage 2 *P*-value = 6.7×10^{-4}; Stage 1+2 *P*-value = 1.0×10^{-8}), and the same risk allele was also associated with reduced insulin sensitivity in the meta-analysis of LOLIPOP and PROMIS studies (Insulin: beta = 4.71% change, *P*-value =1.0×10^{-3}; Glucose: beta = 1.62% change, *P*-value = 3.1×10^{-4}; HOMA-Sensitivity; beta = -4.5% change, *P*-value = 5.0×10^{-4}). The same allele was consistently associated with impairment of glucose homeostasis (elevated fasting glucose and insulin) and reduced insulin sensitivity in addition to higher risk of Type 2 Diabetes. The nearest gene to rs3923113 is growth factor receptor-bound protein 14 (*GRB14*) is an adapter protein which binds to insulin receptors and insulin-like growth-factor receptors, inhibiting tyrosine kinase signaling[173,174]. In addition, *GRB14* knockout mice had higher lean mass, better glucose homeostatisis despite lower insulin and improved insulin sensitivity[175].

The lead genotyped SNP on chromosome 3 was rs16861329 (Stage 1 *P*-value = 2.5×10^{-5}; Stage 2 *P*-value = 1.6×10^{-4}; Stage 1+2 *P*-value = 3.4×10^{-8}), intronic on ST6 beta-galactosamide alpha-2,6-sialyltranferase (*ST6GAL1*). This gene is involved in the post-translational modification of cell-surface components by glycosylation, and glycosylation through addition of sialic acid residues is reported to influence insulin action and cell surface trafficking[176]. This SNP was associated with decreased glucose levels in LOLIPOP and PROMIS (Glucose: beta = -1.37% change, *P*-value = 3.0×10^{-3}). Another potential candidate gene is the adiponectin, C1Q and collagen domain containing gene (*ADIPOQ*) encoding adiponectin (a hormone secreted by adipoctyes which promote insulin sensitivity), upstream of the index SNP. This index SNP was not in linkage disequilibrium ($r^2 < 0.1$) with reported *ADIPOQ* variants, which showed an association with adiponectin levels, obesity and Type 2 Diabetes[177], although adiponectin knockout mice showed severe insulin resistance[178].

On chromosome 10, the lead genotyped SNP rs1802295 (Stage 1 *P*-value = 1.9×10^{-6}; Stage 2 *P*-value = 6.6×10^{-4}; Stage 1+2 *P*-value = 4.1×10^{-8}) is in vacuolar protein sorting 26 homolog A (*VPS26A*) which has not been known to be associated with Type 2 Diabetes nor glucose metabolism. *VPS26A* is a multimeric protein involved in the transport of proteins from endosomes to the trans-Golgi network[179,180] and is also expressed in pancreatic, adipose and other tissues[181]. The same risk allele was associated with elevated glucose levels in LOLIPOP and PROMIS (Glucose: beta = 1.16% change, *P*-value = 8.0×10^{-3}).

There were two loci on chromosome 15 that showed an association with Type 2 Diabetes in South Asians. At 15q24, rs7178572 (Stage 1 *P*-value = 2.4×10^{-5}; Stage 2 *P*-value = 7.0×10^{-7}; Stage 1+2 *P*-value = 7.1×10^{-11}) is intronic on *HMG20A*, which is a non-histone chromosomal protein that may influence histone methylation and is involved in neuronal development[182,183]. This same locus also showed an association in the meta-analysis of three Asian ethnic groups of Chinese, Malays and Indians in Study 2, likely driven by the Indians. At 15q26, rs2028299 was associated with Type 2 Diabetes (Stage 1 *P*-value = 4.8×10^{-5}; Stage 2 *P*-value = 1.1×10^{-7}; Stage 1+2 *P*-value = 1.9×10^{-11}). This index SNP is located near a number of potential candidate genes which might be implicated in the biological pathway of Type 2 Diabetes. The nearest gene, adaptor-related protein complex 3, sigma 2 subunit (*AP3S2*), encodes a clathrin associated adaptor complex expressed in adipocytes, pancreatic islets and other tissues[184]. The perilipin 1 gene (*PLIN1*) 300kb upstream of the index SNP has been implicated with obesity in human and experimental mouse models[185,186]. It encodes Perilipin-1, a phosphoprotein which coats fat droplets in adipocytes and regulates lipolysis by hormone sensitive lipase[187]. Finally, the index SNP rs2028299 is 1.2MB away from the index SNP rs8042680 ($r^2 = 0$ for rs2028299 and rs8042680 in HapMap II CEU) associated with Type 2 Diabetes in Europeans on the protein

regulator of cytokinesis 1 (*PRC1*) gene[12]. Both these loci did not show any association with the glycemic traits.

Lastly, on chromosome 20, the lead SNP rs4812829 (Stage 1 *P*-value = 4.5×10^{-7}; Stage 2 *P*-value = 2.8×10^{-5}; Stage 1+2 *P*-value = 2.6×10^{-10}) with the strongest association among directly genotyped and imputed SNP is intronic in hepatocyte nuclear factor 4, alpha (*HNF4A*), known to be implicated in maturity-onset diabetes of the young (MODY), characterized by defective beta-cell function and impaired insulin secretion[188]. *HNF4A* is a nuclear transcription factor strongly expressed in the liver[189], and the risk allele was associated with reduced pancreatic beta cell function in South Asians (Glucose: beta = 2.33%, *P*-value = 1.0×10^{-6}; HOMA-Beta; beta = -4.5%, *P*-value = 1.0×10^{-3}).

Table 10. Association test results of the index SNPs from the six loci reaching genome-wide significance $P < 5 \times 10^{-8}$ in South Asians (originally from reference 109)

SNP	Chr position	Nearest Gene	Alleles (ref/risk)	South Asians							Europeans (DIAGRAM+)		
				Risk allele Freq	Stage 1 Genome-wide analysis		Stage 2 Replication		Combined Stage 1 + 2		Risk allele Freq	OR (95% CI)	*P*-value
					OR (95% CI)	*P*-value	OR (95% CI)	*P*-value	OR (95% CI)	*P*-value			
rs3923113	2 165210095	*GRB14*	C/A	0.74	1.15 (1.09-1.21)	3.7×10^{-7}	1.07 (1.03-1.11)	6.7×10^{-4}	1.09 (1.06-1.13)	1.0×10^{-8}	0.64	1.05 (1.01-1.10)	2.0×10^{-2}
rs16861329	3 188149155	*ST6GAL1*	A/G	0.75	1.12 (1.07-1.19)	2.3×10^{-5}	1.07 (1.03-1.11)	1.6×10^{-4}	1.09 (1.06-1.12)	3.4×10^{-8}	0.86	1.02 (0.95-1.09)	0.62
rs1802295	10 70601480	*VPS26A*	G/A	0.26	1.14 (1.08-1.20)	1.9×10^{-6}	1.06 (1.03-1.10)	6.6×10^{-4}	1.08 (1.05-1.12)	4.1×10^{-8}	0.31	1.04 (1.00-1.09)	6.0×10^{-2}
rs7178572	15 75534245	*HMG20A*	A/G	0.52	1.10 (1.05-1.15)	2.4×10^{-5}	1.08 (1.05-1.12)	7.0×10^{-7}	1.09 (1.06-1.12)	7.1×10^{-11}	0.71	1.07 (1.02-1.12)	2.6×10^{-3}
rs2028299	15 88175261	*AP3S2*	A/C	0.31	1.11 (1.05-1.16)	4.8×10^{-5}	1.09 (1.06-1.13)	1.1×10^{-7}	1.10 (1.07-1.13)	1.9×10^{-11}	0.31	1.05 (1.00-1.09)	4.0×10^{-2}
rs4812829	20 42422681	*HNF4A*	G/A	0.29	1.14 (1.08-1.19)	4.5×10^{-7}	1.07 (1.04-1.11)	2.8×10^{-5}	1.09 (1.06-1.12)	2.6×10^{-10}	0.19	1.08 (1.02-1.14)	1.0×10^{-2}

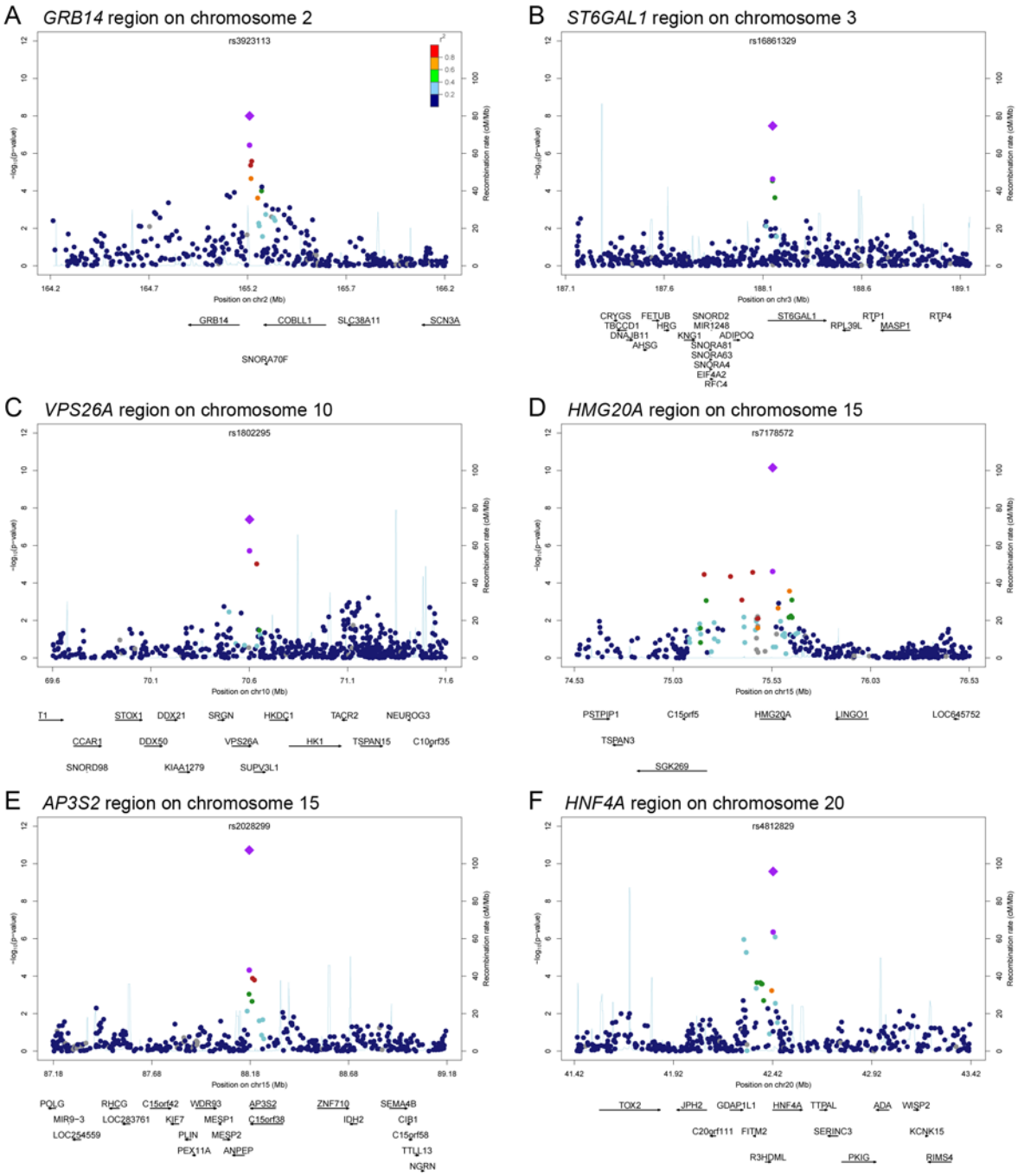


Figure 20. Regional association plots of observed genotyped SNPs at the six new loci associated with Type 2 Diabetes in individuals of South Asian ancestry. Results of the index SNPs in stage 1 were represented by a purple dot and combined analyses results of stage 1 and 2 were plotted as a purple diamond. The surrounding SNPs were colored based on their r^2 with the index SNP from the HapMap CEU reference panel [(originally from reference 109)].

6.3. Transferability of known Type 2 Diabetes to South Asians and assessment of linkage disequilibrium structure and heterogeneity compared to Europeans

While the meta-analysis discovered six new genetic loci associated with Type 2 Diabetes in individuals of South Asian ancestry, we applied varLD to quantify regional linkage disequilibrium differences between CEU Europeans from HapMap II and the South Asians populations. There was no evidence of linkage disequilibrium differences between LOLIPOP and SINDI but there appeared to be some evidence of linkage disequilibrium differences between Europeans and South Asians at the *VPS26A* locus (Table 11). In DIAGRAM+, the index SNP at the *VPS26A* was not statistically associated with Type 2 Diabetes in European populations (Table 11).

Table 11. Comparison of regional linkage disequilibrium structure between South Asians populations (LOLIPOP, SINDI) and CEU (HapMap2). Results were presented as Monte Carlo *P*-values for comparison of pairwise LD between SNPs at the loci by VarLD [(originally from reference 109)].

Lead SNP	Nearest Gene	CEU – LOLIPOP	CEU – SINDI	LOLIPOP – SINDI
rs3923113	*GRB14*	0.40	0.16	0.56
rs16861329	*ST6GAL1*	0.08	0.05	0.69
rs1802295	*VPS26A*	0.006	0.002	0.18
rs7178572	*HMG20A*	0.62	0.33	0.90
rs2028299	*AP3S2*	0.84	0.11	0.38
rs4812829	*HNF4A*	0.06	0.13	0.58

We further looked up 42 previously reported loci in the South Asian Indians. Table 12 showed a list of 42 variants at 41 loci implicated in Type 2 Diabetes, mostly in the Europeans. A total of 27 variants showed an association (*P*-value < 0.05) in the South Asians Stage 1 meta-analysis. A Binomial test for assessing whether the number of observed nominally significant association would be expected by chance under $P = 0.05$ indicated evidence of over-representation of associated loci (*P*-value < 2.2×10^{-16}). More than half of the previously reported loci showed nominal significance in the South Asians, further corroborating the observation in Study 2 that many of the European discovered loci were relevant across populations. This meta-analysis was better powered with a bigger sample size and lesser heterogeneity in the studies combined since

the discovery stage consisted only of individuals of South Asian. Under the null hypothesis that the proportion of variants showing association in the same direction was ½ by chance, 37 out of 42 variants showed consistency in the directions of association for the same alleles (*P*-value = 4.43×10^{-7}).

Three loci showed evidence of heterogeneity in effects between the Europeans and South Asians, namely, glucokinase (hexokinase 4) regulator (*GCKR*), *CDKAL1*, and Kruppel-like factor 14 (*KLF14*). *GCKR* encodes for the glucokinase regulator protein (GKRP), which regulates glycolysis primarily in liver hepatocytes. *GCKR* was associated with glycemic traits in multiple populations[76,91,98,190-192]. In 40,655 cases and 87,022 non-diabetic controls of European ancestry, *GCKR* was associated with Type 2 Diabetes (OR = 1.06, 95% CI = 1.04 – 1.08, *P*-value = 1.30×10^{-9})[76]. This association was not seen in the DIAGRAM+ study consisting of samples from European descent (total number of cases and controls = 22,570), probably due to the small effect size and comparatively smaller number of samples (Table 12). The effect size was larger in the South Asians with a total sample size of 10,816 (OR = 1.19, 95%CI = 1.11 – 1.19, *P*-value = 4.10×10^{-6}). There was also heterogeneity in the effect sizes at the *CDKAL1* locus, with a smaller effect size in the South Asians. The last locus was *KLF14*, which was only recently found to be implicated in Type 2 Diabetes[12]. There was no evidence of association in the South Asians, with the ORs trending in the opposite directions.

Table 12. Known Type 2 Diabetes loci and their index variants tested for replication in the South Asians meta-analysis. Risk alleles were in accordance with previously published risk alleles in the Europeans [originally from reference 109]. Index variants with association *P*-value < 0.05 in South Asians are shaded in grey

SNP	Chr	Pos (bp)	Nearest Gene	Risk allele	Ref allele	Europeans			South Asians			P_{heter} between European and South Asians
						Risk allele Freq	OR (95% CI)	*P*-value	Risk allele Freq	OR (95% CI)	*P*-value	
rs10923931	1	120230001	*NOTCH2*	T	G	0.12	1.14 (1.07 - 1.21)	6.50×10^{-05}	0.18	1.01 (0.95- 1.07)	8.30×10^{-01}	5.50×10^{-03}
rs340874	1	210547651	*PROX1*	C	T	0.51	1.07 (1.02-1.11)	2.00×10^{-03}	0.55	1.02 (0.98-1.07)	3.10×10^{-01}	2.00×10^{-01}
rs780094	2	27652888	*GCKR*	C	T	0.62	1.01 (0.97-1.05)	6.00×10^{-01}	0.74	1.19 (1.11-1.29)	4.10×10^{-06}	**1.20×10^{-04}**
rs11899863	2	43530470	*THADA*	C	T	0.93	1.17 (1.09-1.25)	1.00×10^{-05}	0.96	0.96 (0.80-1.15)	6.30×10^{-01}	4.40×10^{-02}
rs243021	2	60496470	*BCL11A*	A	G	0.46	1.09 (1.05-1.14)	8.10×10^{-06}	0.51	1.05 (1.00-1.10)	3.60×10^{-02}	1.90×10^{-01}
rs7593730	2	160996961	*RBMS1*	C	T	0.80	1.07 (1.02-1.13)	6.50×10^{-03}	0.80	1.01 (0.95-1.07)	7.80×10^{-01}	1.10×10^{-01}
rs7578326	2	226846158	*IRS1*	A	G	0.64	1.12 (1.07-1.17)	8.70×10^{-07}	0.77	1.08 (1.03-1.14)	4.40×10^{-03}	4.20×10^{-01}
rs13081389	3	12264800	*PPARG*	A	G	0.96	1.24 (1.14-1.35)	2.00×10^{-07}	0.93	1.07 (0.97-1.19)	1.70×10^{-01}	2.80×10^{-02}
rs6780569	3	23173488	*UBE2E2*	G	A	0.95	1.11 (1.04-1.18)	1.50×10^{-03}	0.74	1.07 (1.02-1.13)	8.90×10^{-03}	4.40×10^{-01}
rs6795735	3	64680405	*ADAMTS9*	C	T	0.54	1.09 (1.04-1.13)	8.40×10^{-05}	0.28	1.06 (1.01-1.12)	2.40×10^{-02}	5.10×10^{-01}
rs11708067	3	124548468	*ADCY5*	A	G	0.78	1.10 (1.05-1.16)	1.70×10^{-04}	0.78	1.10 (1.04-1.17)	1.40×10^{-03}	9.80×10^{-01}
rs1470579	3	187011782	*IGF2BP2*	C	A	0.29	1.14 (1.09-1.19)	2.20×10^{-09}	0.42	1.12 (1.07-1.17)	2.50×10^{-06}	5.80×10^{-01}
rs1801214	4	6421094	*WFS1*	T	C	0.73	1.13 (1.08-1.18)	3.20×10^{-08}	0.68	1.08 (1.02-1.13)	7.20×10^{-03}	1.60×10^{-01}
rs4457053	5	76460705	*ZBED3*	G	A	0.26	1.16 (1.10-1.23)	4.20×10^{-08}	0.22	0.97 (0.88-1.08)	6.30×10^{-01}	3.00×10^{-03}
rs10440833	6	20796100	*CDKAL1*	A	T	0.25	1.25 (1.20-1.31)	1.80×10^{-22}	0.26	1.08 (1.02-1.14)	4.90×10^{-03}	**2.80×10^{-05}**
rs2191349	7	14837549	*DGKB*	T	G	0.47	1.07 (1.03-1.11)	1.20×10^{-03}	0.62	1.05 (1.00-1.11)	3.50×10^{-02}	6.70×10^{-01}
rs849134	7	27969462	*JAZF1*	A	G	0.53	1.13 (1.08-1.17)	2.80×10^{-09}	0.69	1.06 (1.01-1.12)	2.30×10^{-02}	6.50×10^{-02}
rs4607517	7	44008908	*GCK*	A	G	0.2	1.03 (0.97-1.09)	3.10×10^{-01}	0.13	1.01 (0.95-1.08)	7.10×10^{-01}	7.30×10^{-01}
rs972283	7	129924109	*KLF14*	G	A	0.55	1.10 (1.06-1.15)	1.80×10^{-06}	0.61	0.98 (0.94-1.03)	4.60×10^{-01}	**2.70×10^{-04}**
rs896854	8	96029687	*TP53INP1*	T	C	0.48	1.10	1.20×10^{-06}	0.41	1.08	1.80×10^{-03}	4.50×10^{-01}

							(1.06-1.15)			(1.03-1.13)		
rs3802177	8	118254206	*SLC30A8*	G	A	0.75	1.15 (1.10-1.21)	1.50×10^{-08}	0.76	1.13 (1.07-1.19)	2.40×10^{-05}	5.70×10^{-01}
rs17584499	9	8869118	*PTPRD*	T	C	0.25	1.03 (0.96-1.10)	3.80×10^{-01}	0.25	0.98 (0.93-1.04)	5.50×10^{-01}	2.80×10^{-01}
rs10965250	9	22123284	*CDKN2A/B*	G	A	0.81	1.20 (1.13-1.27)	1.20×10^{-10}	0.87	1.20 (1.10-1.30)	2.60×10^{-05}	9.60×10^{-01}
rs13292136	9	79181682	*CHCD9*	C	T	0.93	1.20 (1.11-1.29)	1.50×10^{-06}	0.86	1.10 (1.03-1.17)	7.50×10^{-03}	8.00×10^{-02}
rs12779790	10	12368016	*CDC123*	G	A	0.23	1.09 (1.04-1.15)	6.80×10^{-04}	0.17	1.12 (1.05-1.20)	5.90×10^{-04}	4.90×10^{-01}
rs5015480	10	94455539	*HHEX*	C	T	0.57	1.18 (1.13-1.23)	1.30×10^{-15}	0.45	1.08 (1.03-1.13)	2.10×10^{-03}	3.30×10^{-03}
rs7903146	10	114748339	*TCF7L2*	T	C	0.25	1.40 (1.34-1.46)	2.20×10^{-51}	0.31	1.25 (1.19-1.32)	3.40×10^{-19}	1.10×10^{-03}
rs2334499	11	1653425	*DUSP8*	T	C	0.44	1.08 (1.04-1.13)	1.20×10^{-04}	0.28	1.02 (0.97-1.07)	4.20×10^{-01}	7.80×10^{-02}
rs231362	11	2648047	*KCNQ1*	G	A	0.52	1.11 (1.06-1.16)	6.40×10^{-06}	0.73	1.09 (1.03-1.15)	3.00×10^{-03}	5.90×10^{-01}
rs163184	11	2803645	*KCNQ1*	G	T	0.44	1.09 (1.04-1.13)	6.80×10^{-05}	0.53	1.08 (1.03-1.13)	1.20×10^{-03}	8.40×10^{-01}
rs5215	11	17365206	*KCNJ11*	C	T	0.41	1.09 (1.05-1.14)	1.60×10^{-05}	0.37	1.04 (0.99-1.09)	1.10×10^{-01}	1.10×10^{-01}
rs1552224	11	72110746	*CENTD2*	A	C	0.88	1.13 (1.07-1.19)	7.00×10^{-06}	0.83	1.04 (0.98-1.10)	2.20×10^{-01}	4.20×10^{-02}
rs1387153	11	92313476	*MTNR1B*	T	C	0.28	1.12 (1.07-1.17)	1.00×10^{-06}	0.39	1.07 (1.02-1.12)	8.70×10^{-03}	1.30×10^{-01}
rs1531343	12	64461161	*HMGA2*	C	G	0.1	1.20 (1.12-1.29)	1.70×10^{-07}	0.18	1.07 (1.01-1.13)	3.40×10^{-02}	1.00×10^{-02}
rs4760790	12	69921061	*TSPAN8*	A	G	0.23	1.11 (1.06-1.16)	3.60×10^{-06}	0.34	1.06 (1.01-1.11)	1.80×10^{-02}	1.60×10^{-01}
rs7957197	12	119923406	*HNF1A*	T	A	0.85	1.14 (1.08-1.19)	4.60×10^{-07}	0.95	1.14 (0.97-1.34)	1.20×10^{-01}	9.90×10^{-01}
rs7172432	15	60183681	*C2CD4A/B*	A	G	0.52	1.07 (1.03-1.12)	1.10×10^{-03}	0.61	1.05 (1.01-1.11)	2.60×10^{-02}	6.40×10^{-01}
rs11634397	15	78219277	*ZFAND6*	G	A	0.6	1.11 (1.06-1.16)	5.10×10^{-06}	0.53	1.05 (1.00-1.12)	7.50×10^{-02}	1.80×10^{-01}
rs8042680	15	89322341	*PRC1*	A	C	0.22	1.10 (1.06-1.15)	8.20×10^{-06}	0.63	1.06 (1.01-1.11)	2.80×10^{-02}	2.00×10^{-01}
rs11642841	16	52402988	*FTO*	A	C	0.45	1.13 (1.08-1.18)	3.40×10^{-08}	0.32	1.07 (1.02-1.14)	1.20×10^{-02}	1.70×10^{-01}
rs391300	17	2163008	*SRR*	C	T	0.64	1.00 (0.96-1.04)	9.50×10^{-01}	0.51	0.99 (0.94-1.03)	6.10×10^{-01}	6.70×10^{-01}
rs4430796	17	33172153	*HNF1B*	G	A	0.53	1.14 (1.08-1.20)	1.50×10^{-06}	0.37	1.07 (1.02-1.13)	4.10×10^{-03}	1.10×10^{-01}

6.4. Obesity and Type 2 Diabetes in South Asians

Obesity is a major risk factor for Type 2 Diabetes. Individuals with Type 2 Diabetes tended to have higher BMI than non-diabetic individuals. As obesity is one of the pathways leading to predisposition to Type 2 Diabetes, it was possible that the association at some of the genetic loci identified could be mediated via obesity. We carried out the following secondary analyses: i) genome-wide study in BMI extremes (lean Type 2 Diabetes cases with BMI < 25kg/m^2 versus overweight controls with BMI > 25kg/m^2); ii) association of the six index SNPs with BMI and WHR in the LOLIPOP and PROMIS and iii) adjusting for BMI or waist-hip-ratio (WHR) at the six new loci across the Stage 2 replication cohorts where WHR was available across all cohorts.

In the genome-wide analysis of BMI extremes, only *TCF7L2* reached genome-wide significance. The six index SNPs were not associated with anthropometric measures BMI and WHR in LOLIPOP and PROMIS (Table 13). Adjustments for BMI and WHR did not remove the association of the six index SNPs with Type 2 Diabetes in the replication cohorts (Table 13). These findings suggest the associations of these SNPs with Type 2 Diabetes were independent of obesity.

Table 13. Association of the six index SNPs with [originally from reference 109]

i) Secondary quantitative anthropometric traits in the LOLIPOP and PROMIS cohorts, as change in phenotype per copy of risk allele in the Type 2 Diabetes association and adjusted for age and gender. Associations were computed in each study separately and combined by inverse variance meta-analysis.

ii) Type 2 Diabetes in the Stage 2 replication cohorts, with no adjustment for adiposity measures, adjustment for BMI and adjustment for WHR. Results are presented as OR (95% CI) for each copy of Type 2 Diabetes risk allele. All analyses were adjusted for age and gender in each individual cohort and combined by inverse variance meta-analysis.

SNP	Chr position	Nearest Gene	Risk allele	Stage 1 Discovery cohorts (LOLIPOP + PROMIS)				Stage 2 Replication cohorts					
				BMI		WHR		Unadjusted		BMI adjusted		WHR adjusted	
				Beta	P	Beta	P	OR	P	OR	P	OR	P
rs3923113	2 165210095	*GRB14*	A	-0.09	6.9×10^{-02}	-0.001	2.4×10^{-01}	1.07 (1.03-1.11)	6.7×10^{-04}	1.09 (1.04-1.14)	1.0×10^{-04}	1.05 (1.01-1.10)	2.6×10^{-02}
rs16861329	3 188149155	*ST6GAL1*	G	0.08	1.3×10^{-01}	0.001	3.9×10^{-01}	1.07 (1.03-1.11)	1.6×10^{-04}	1.07 (1.02-1.11)	3.4×10^{-03}	1.07 (1.03-1.12)	1.7×10^{-03}
rs1802295	10 70601480	*VPS26A*	A	0.07	1.7×10^{-01}	0.000	8.8×10^{-01}	1.06 (1.03-1.10)	6.6×10^{-04}	1.07 (1.02-1.11)	1.9×10^{-03}	1.06 (1.02-1.11)	3.7×10^{-03}
rs7178572	15 75534245	*HMG20A*	G	-0.07	1.5×10^{-01}	-0.001	3.0×10^{-01}	1.08 (1.05-1.12)	7.0×10^{-07}	1.07 (1.02-1.11)	1.4×10^{-03}	1.07 (1.03-1.11)	1.1×10^{-03}
rs2028299	15 88175261	*AP3S2*	C	-0.02	7.0×10^{-01}	0.000	6.0×10^{-01}	1.09 (1.06-1.13)	1.1×10^{-07}	1.09 (1.05-1.13)	6.1×10^{-06}	1.09 (1.05-1.13)	2.7×10^{-05}
rs4812829	20 42422681	*HNF4A*	A	-0.02	7.0×10^{-01}	0.000	6.0×10^{-01}	1.07 (1.04-1.11)	2.8×10^{-05}	1.08 (1.03-1.12)	2.1×10^{-04}	1.08 (1.04-1.12)	2.0×10^{-04}

6.5. Summary

This meta-analysis across large samples of individuals of South Asian ancestry exhibit the potential of non-European genome-wide association efforts to detect new loci associated with Type 2 Diabetes, or even in other diseases. As European data was used to prioritize the selection of SNPs to the replication phase, four out of the six loci were also associated with Type 2 Diabetes in European populations. More than half of the currently established Type 2 Diabetes implicated loci were also associated with the same outcome in South Asians, due to increased power from a larger sample size and the relative homogeneity of the populations (as compared to Study 2).This further supports the observations made in Study 2 that many of these common variants are largely shared across populations.

The key findings from Study 3 were:

I. Six new common variants were found to be associated with Type 2 Diabetes in people of South Asian ancestry.
II. These new loci were not associated with secondary anthropometric traits and including these anthropometric traits as covariates did not remove any association at these six loci. These findings suggested that these associations were independent of obesity.
III. Of 42 previously implicated Type 2 Diabetes loci, 27 showed as association (*P*-value < 0.05) in the South Asians. The observation is unlikely to happen by chance (Binomial test *P*-value < 2.2×10^{-16}).

CHAPTER 7 – TYPE 2 DIABETES AND OBESITY

7.1. Motivation

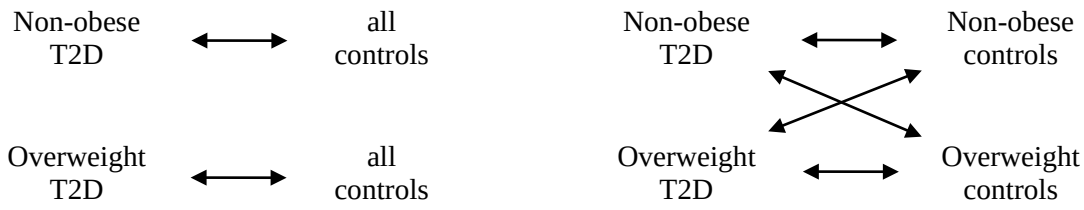
Type 2 Diabetes is a complex chronic disease and closely related to the metabolic syndrome which is a clustering of obesity, dyslipidemia, hypertension and glucose intolerance, now known as risk factors for cardiovascular diseases. There are several pathways leading to disease manifestation of Type 2 Diabetes. One of these pathways is commonly linked to obesity dependent abnormalities of muscle, fat or liver responses to insulin. The incidence of Type 2 Diabetes worldwide has been linked to rising rates of obesity, brought on by affluence, sedentary lifestyle and over-nutrition[193].

The World Health Organization (WHO) of the United Nations classifies individuals with BMI ≥ 25 kg/m^2 as overweight and BMI ≥ 30 kg/m^2 as obese, using BMI as a surrogate for adiposity. While there have been recommendations to lower the cutoffs for Asian populations as it is likely that the existing cutoffs might underestimate the overall risk attributable to obesity[111,194,195], the use of ethnic-specific BMI cutoffs might be complicated by health management, variation in prevalence, environmental and nutritional changes in increasingly multi-ethnic metropolitan populations in the world[110]. In 2004, WHO identified additional cut-points for Asians with a threshold of 23 kg/m^2 to differentiate between healthy and overweight and a threshold of 27.5 kg/m^2 for obese[110]. These cut-offs at 23kg/m^2 will generate much clearer 'extreme' phenotypes of hypercases and hypercontrols in these Asian populations compared to the 25kg/m^2 cutoff but will have an impact on the sample sizes, especially in the Chinese (Table 14).

Early genome-wide association studies established the association of the *FTO* gene with Type 2 Diabetes, mediated through the effects of obesity[196]. The risk allele predisposing to Type 2 Diabetes was also associated with increased BMI, but the association signal for Type 2 Diabetes

was attenuated when BMI was included as a covariate in the association analyses. In addition, the *FTO* association was not consistently replicated in genome-wide association scans of European descent despite considerable power[134,135,197]. A likely explanation has to do with the ascertainment of lean Type 2 Diabetes subjects in some of these studies, while certain studies prioritized mainly diabetic cases with considerably higher BMI[136]. Timpson and colleagues further investigated the disease susceptibility heterogeneity of non-obese and overweight Type 2 Diabetes cases through stratified analyses[198].

We followed up with Study 2 (Chapter 5), by refining case and controls phenotype in the following ways: (i) performing association between non-obese cases and all controls; and overweight cases and all controls (ii) performing association between all pairwise combinations of non-obese cases/controls and overweight cases/controls.



7.2. Summary characteristics by obesity status

It can be seen from Table 14 below that there are differences in the obesity status across Type 2 Diabetes case ascertainment and ethnic groups. Within the cases, there were higher percentages of overweight cases in the Malays and Indians. Chinese controls tended to be non-obese.

Table 14. Number of Type 2 Diabetes case controls stratified by BMI status.

BMI stratification	N	Type 2 Diabetes Disease Status							
		Type 2 Diabetes Cases				Type 2 Diabetes Controls			
	non-obese BMI < 25 kg/m^2	Illu610 Chinese SDCS	Illu1M Chinese SDCS	Illu610 Malays SiMES	Illu610 Indians SINDI	Illu610 Chinese SP2	Illu1M Chinese SP2	Illu610 Malays SiMES	Illu610 Indians SINDI
		531 (49.4)	443 (48.3)	237 (30.3)	356 (36.6)	816 (81.1)	704 (75.2)	643 (52.2)	581 (50.5)
	overweight BMI ≥ 25 kg/m^2	Illu610 Chinese SDCS	Illu1M Chinese SDCS	Illu610 Malays SiMES	Illu610 Indians SINDI	Illu610 Chinese SP2	Illu1M Chinese SP2	Illu610 Malays SiMES	Illu610 Indians SINDI
		544 (50.6)	474 (51.7)	545 (69.7)	618 (63.4)	190 (18.9)	232 (24.8)	589 (47.8)	570 (49.5)

(): within column percentages.

7.3. Heterogeneity in association signal by obesity status

The results from the BMI-stratified analyses with all controls are presented in Figure 21. Due to the small sample size, these stratified analyses had little power to detect associations at genome-wide significance. We defined *P*-value < 10^{-6} as suggestive evidence of association.

Comparing the non-obese cases with all controls as the reference group, there were suggestive evidences of association at two loci: *CDKAL1* at chromosome 6 and insulin-like growth factor 1 receptor (*IGF1R*) on chromosome 15. Established index SNP rs7754840 at *CDKAL1* was the top ranking SNP (OR = 1.27, 95% CI = 1.26 – 1.38, *P*-value = 2.86 x 10^{-7}). This SNP had *P*-value at 10^{-7} while *P*-value using all cases and controls in Study 2 was at 10^{-5}. Odds ratio was also higher compared to 1.15 (1.08 – 1.23) in Study 2. Risk allele at *CDKAL1* was associated with reduced beta-cell glucose sensitivity[199], suggesting that this has effects on insulin secretion[73,200,201]. The signal at the index SNP rs7180435 was mainly driven by the Indians (Meta-analysis: OR = 1.82, 95% CI = 1.43 – 2.30, *P*-value = 7.65 x 10^{-7} and Indians: OR = 1.81, 95% CI = 1.82 – 2.33, *P*-value = 4.42 x 10^{-6}). This SNP was almost monomorphic in the Chinese (risk allele 0.002) and Malay (risk allele 0.021) while risk allele frequency was 0.132 in the Indians. These findings and study design however, were not replicated in Study 4. Nevertheless, *IGF1R* is a transmembrane receptor that is activated by the two growth factors, insulin-like growth receptor 1 and 2 (*IGF-1*

and *IGF-2*). This receptor has been associated with several cancers such as breast[202], prostate[203] and lung[204]. Interestingly, heterozygous knockout mice showed longer lifespans with small decrease in their growth[205].

In the association analysis between overweight cases and all controls, suggestive association signals between *FTO* and Type 2 Diabetes were detected similar to what Timpson et al. reported[198]. Index SNP rs9939609 (established association with Type 2 Diabetes and obesity)[45,196,206] was associated with Type 2 Diabetes in overweight cases versus all controls (OR = 1.26, 95% CI = 1.15 – 1.38, *P*-value = 6.22 x 10^{-7}) while it ranked 2,346,240 in the non-obese cases versus all controls association analysis (OR = 1.00, 95% CI = 0.90 – 1.12, *P*-value = 9.44 x 10^{-1}).

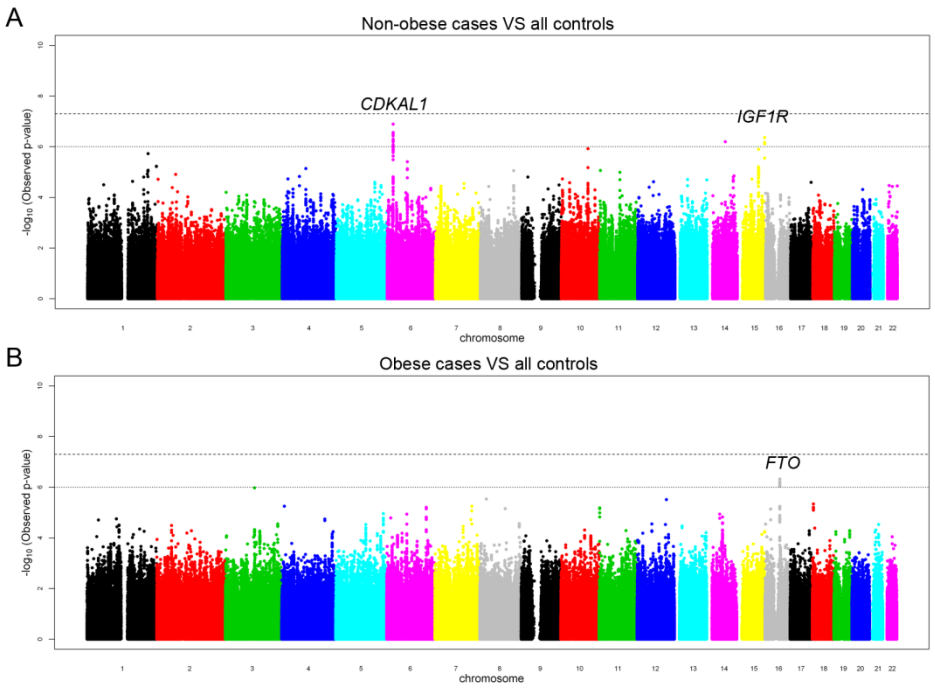


Figure 21. Manhattan plots of genome-wide association analyses. A) Association between non-obese cases and all controls; B) Association between overweight cases and all controls.

To investigate possible interaction effects by obesity status, we further stratified controls, in addition to cases, by their obesity status (Figure 22). We only detected an association at *CDKAL1* in non-obese cases and non-obese controls (Index SNP rs7754840: OR = 1.28, 95% CI: 1.16 – 1.41, *P*-value = 3.88 x 10^{-7}). These results were similar to the association between non-obese cases and all controls. In comparing extremes of overweight cases and non-obese controls, only *FTO* showed suggestive evidence of association, with the top SNP rs7185735 at *P*-value = 7.21 x 10^{-8}, almost reaching genome-wide significance. The use of more extreme controls (non-obese) resulted in a 10 fold more significant *P*-value for rs9939609 and larger effect size compared to using all controls (OR = 1.32, 95% CI = 1.19 – 1.46, *P*-value = 8.97 x 10^{-8}). In the extremes of non-obese cases and overweight controls, we identified suggestive signals on chromosome 6 *C6orf57*/*KIAA1411* which was also present in Study 2 and driven mainly by the Indians, where there was no stratification by obesity status. In the stratified analysis, the top SNP was 200kb upstream of the association signal in the un-stratified analysis and was associated in both Malays (*P*-value = 7.60 x 10^{-4}) and Indians (*P*-value = 1.02 x 10^{-4}), with consistent direction of effect in the Chinese (*P*-value = 0.09). While the *FTO* variant was detected in overweight cases versus non-obese controls, no evidence of association was seen in the stratum of overweight cases and overweight controls.

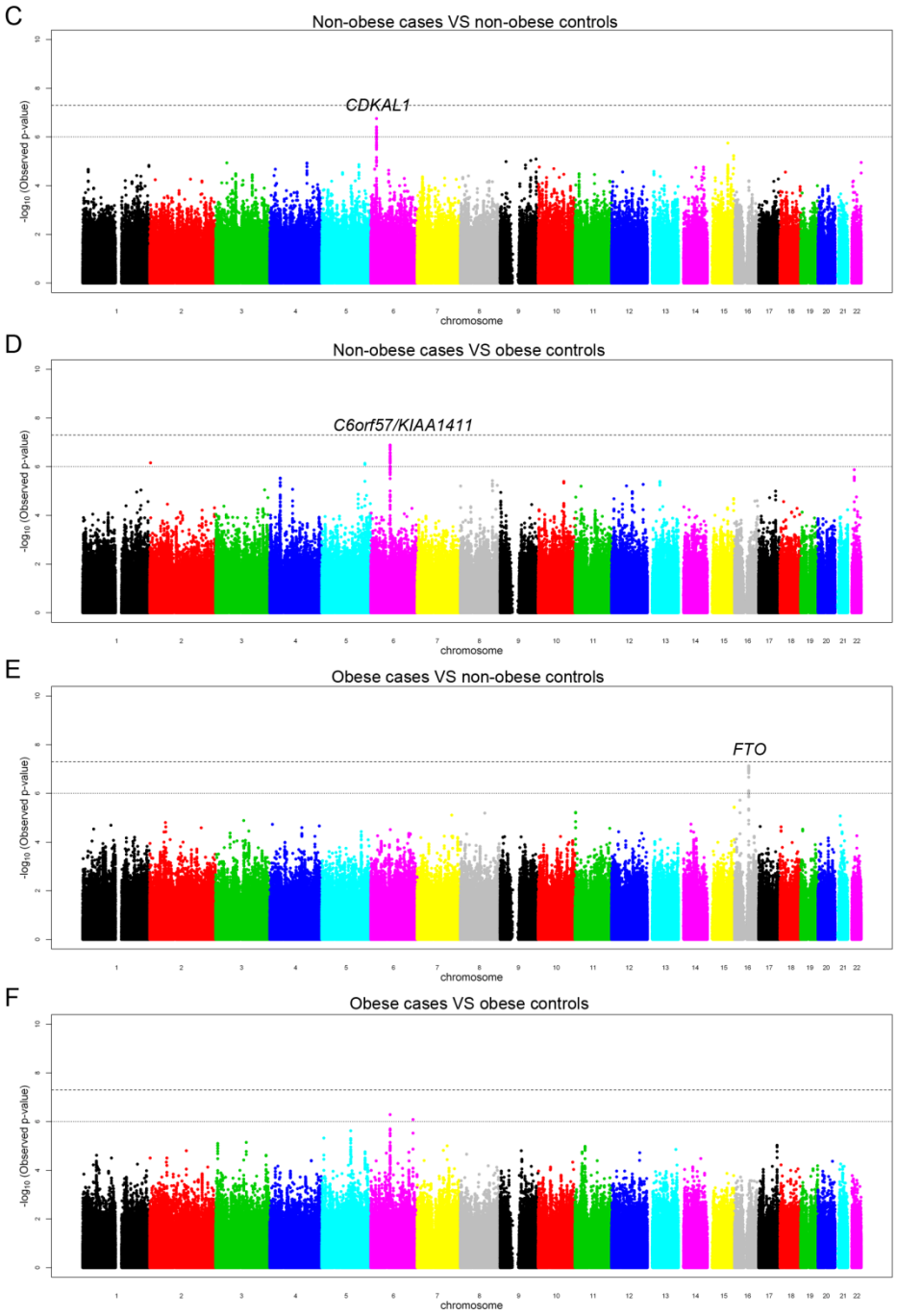


Figure 22. Manhattan plots of genome-wide association analyses. C) Association between non-obese cases and non-obese controls; D) Association between non-obese cases and overweight controls; E) Association between overweight cases and non-obese controls and F) Association between overweight cases and overweight controls.

We further looked at the two loci, *CDKAL1* (index SNP rs7754840) and *FTO* (index SNP rs8050136) by using the multinomial logistic regression in the Chinese, to test whether the effect sizes differ across the strata of cases (non-obese and overweight cases) against a common control group (all controls) (Table 15)[198,207]. rs8050136 was chosen as it was a directly genotyped SNP. The genotype data was combined over the two arrays, Illumina610 and Illumina1M. Consistent with the stratum-specific results, association signal in *CDKAL1* was primarily driven by the non-obese cases while the association signal in *FTO* was due to the overweight cases.

Table 15. Selected stratified Type 2 Diabetes association results for two index SNPs, rs7754840 and rs8050136, in Chinese.

SNP	Non-obese cases vsall controls		Overweight cases vs all controls		*P*-value* (heterogeneity between strata)
	OR (95% CI)	*P*-value	OR (95% CI)	*P*-value	
CDKAL1 (rs7754840)	1.35 (1.21 – 1.52)	6.69×10^{-8}	1.07 (0.96 – 1.19)	2.51×10^{-1}	< 0.0001
FTO (rs8050136)	1.10 (0.93 – 1.29)	2.67×10^{-1}	1.40 (1.20 – 1.64)	1.51×10^{-5}	< 0.0001

* Test of heterogeneity between strata using a likelihood ratio test of nested multinomial models where model 1 assumed the same beta for all strata while model 2 assumed different beta at different strata.

7.4. Summary

In this BMI-stratified analysis in Asians, we showed that *CDKAL1* is implicated in the non-obese cases, along the reduced insulin pathway of Type 2 Diabetes. Similar to the findings in the European studies, the *FTO* variant was only detected in overweight cases with all controls/non-obese controls. As the *FTO* variant was associated with increasing BMI, overweight cases and overweight controls likely had much more similar allelic frequencies. The refining of cases and controls also suggested *IGF1R* as a likely Type 2 Diabetes implicated locus, although it still needs to be validated in other independent studies. These phenotypic refining supplements our limited knowledge in the physiological pathways of Type 2 Diabetes.

I. Key findings of Study 4:We showed in BMI stratified and multinomial regression analyses that *CDKAL1* is implicated in the reduced insulin pathway of Type 2 Diabetes in Asian populations.
II. Consistent with what was established in populations of European descent, *FTO* affects Type 2 Diabetes along the obesity pathway.
III. Refining of cases and controls definition provides a better understanding of association signals and the pathways leading to disease manifestation.

CHAPTER 8 – DISCUSSION

8.1. Bringing it all together

The genetic architecture of common diseases and complex traits, at its most fundamental, looks at the relationship of genetic variants with the phenotype. As with any epidemiological study, it requires some basic understanding of the phenotype of interest and in this case, an agnostic search for the genetic variants that could increase or decrease the risk of the phenotype. Evolutionary histories, origins and migratory patterns of the populations studied influenced the ease at which this phenotype-genotype relationship can be detected.

We have shown the importance of understanding linkage disequilibrium and genetic diversity in multi-ethnic populations for valid and sensible interpretations of genetic association studies (Study 1). In Singapore, Malays and Indians showed greater genetic heterogeneity within their own ethnic group, likely due to inter-marriages and migratory history. This provided essential basis for understanding the transferability of association signals of Type 2 Diabetes in populations of European ancestry to Asian populations. Under the assumption of common causal variants across populations, reproducibility of association signals depends on the (i) linkage disequilibrium patterns of the index variants in different populations; (ii) the power to detect these association, which relates to the frequencies and effect sizes of these variants in different populations; (iii) the presence of different causal variants in the same implicated locus (Study 2). Certainly, due to evolutionary history and environmental pressures, some variants are expected to be specific to populations or may occur at higher frequencies in specific populations. Through a genome-wide association study of Type 2 Diabetes in South Asians, we also showed the potential for non-European populations to discover new susceptibility loci that might have been missed in the well-studied Europeans (Study 3). Six new loci were implicated in Type 2 Diabetes in this South Asian ancestry which is more genetically diverse and possesses one of the world's highest

prevalence of Type 2 Diabetes. Lastly, we also demonstrated the importance of finer phenotyping in elucidating the roles of genetic predisposition variants and supplementing existing knowledge on disease physiology (Study 4). These are summarized in Figure 23.

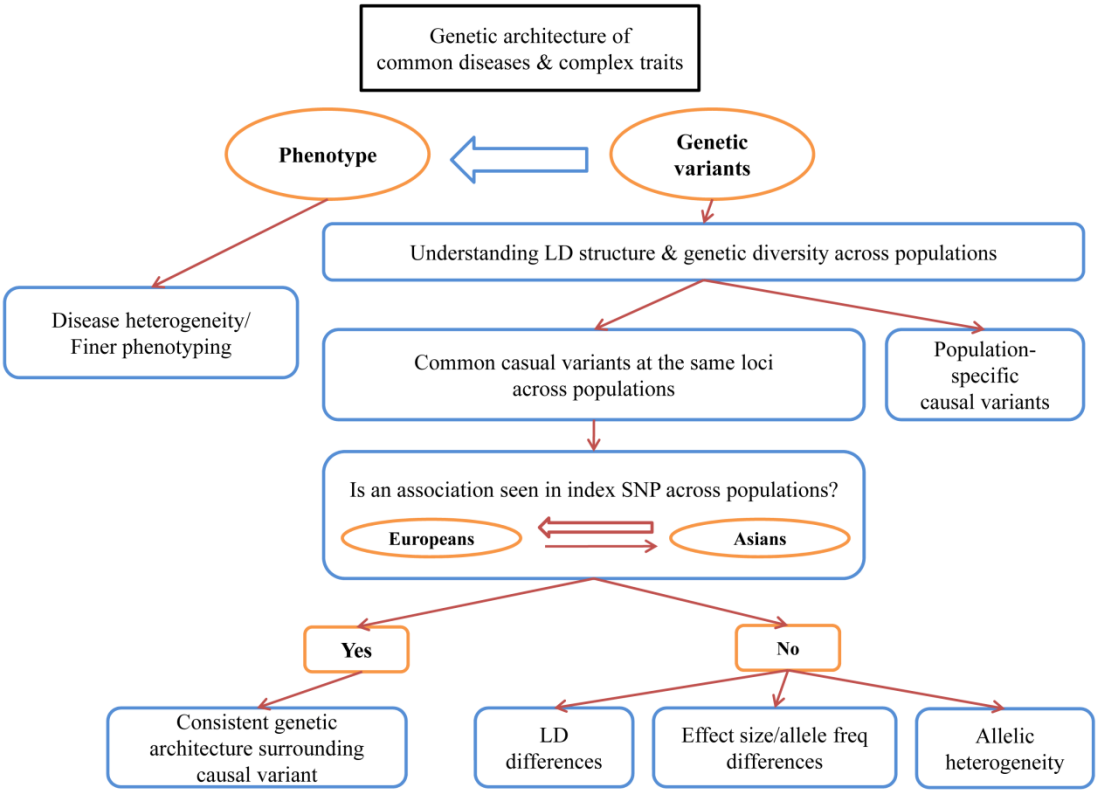


Figure 23. Schematic diagram unifying the four studies from Chapter 4 to Chapter 7.

We observed that implicated variants from published genome-wide association studies tended to display greater evidence of inter-population heterogeneity in effect sizes, particularly so for variants detected by assimilating tens to hundreds of thousands of samples.

8.2. What's next? / Future Work

The eventual aim of genetic studies is translational, from variant discovery to biology to clinical practice. Broadly, the support and endeavor for genetic research stems from two translational

aspects: (i) to better understand the pathogenesis of diseases for improved diagnosis (diagnostic tools), treatment (drug targets and prognosis to targeted therapy) and prevention and (ii) for risk prediction of inherited individual predisposition (personalized medicine)[208]. Thus far, progress has been less prominent for common diseases, with greater successes seen mainly in highly familial monogenic diseases.

Genome-wide association study represents an important advancement beyond the candidate gene and linkage studies and is a critical tool for genetic mapping of common diseases and complex traits in populations. In performing meta-analysis across populations of Chinese, Malay and Asian Indian ancestry, we did not account for population diversity differences between them, especially between the South Asian Indians and the two other East Asian populations. We are likely to detect association signals that are common across these populations or driven strongly by one or more of the populations carrying the implicated variants. In addition, the SNP-based method has its limitation that assumes similar genetic architecture surrounding the causal variant across populations. Methods that focus on a genomic region would be a possible advancement beyond the single SNP method to assimilate statistical evidence across populations. Instead of looking for evidence at index SNPs, Xu et al. introduced a novel method of looking at regional evidence of disease association, taking linkage disequilibrium into account[209]. Across a pre-defined region in the genome (e.g. moving window or around a gene), the method quantifies an over-representation of independent associated SNPs through eigen-decomposition of the linkage disequilibrium matrix. A plethora of genetic variants have been found to be implicated in common diseases, infectious diseases and cancers. These findings have improved our understanding of the genetic architecture of disease but the important question remains, how can these findings bring us closer to translational genomics.

8.2.1. Fine mapping

Linkage disequilibrium has been instrumental in the design of genotyping arrays but long stretches of linkage disequilibrium hinder fine mapping to localize the causal variants. Due to different evolutionary history and migratory patterns, different populations exhibit varying degrees of linkage disequilibrium, with the least conserved linkage disequilibrium in African populations. While explicitly assuming a common causal variant across populations, the causal variants could take the form of similar haplotype structure but of differing lengths across populations or the presence of different dominant haplotypes across populations[210]. Sufficiently dense genotyping across populations with appropriate reference panels could provide a means of localizing potential genomic regions for further investigation.

An early example of success was the discovery of the *TCF7L2* locus where a large region on chromosome 10 was implicated in Type 2 Diabetes by linkage and subsequent fine-mapping efforts through sequencing localized the susceptibility region to an intron within the gene[10,11]. The HapMap project had catalogued over 3.5 million common variants across three populations of European, Asian and African ancestry. More recently, the 1000 Genomes project aim to discover and provide haplotype information on multiple form of human genetic variation through sequencing major population groups from Europe, East Asia, South Asia, West Africa and Americas[40]. This will provide a wider spectrum of genetic variants beyond commercial genotyping arrays and allow association tests at low frequency variants not previously discovered. As with the course of genome-wide association analysis, these sequenced reference panels will provide a less costly mean of performing association analyses by imputing the sequenced variants into previously genotyped samples. While imputation accuracy decreases as allele frequency decreases, the more complete catalogue of putative functional annotation in these variants offers a

more comprehensive resource in understanding and localizing putative functional variants in known disease implicated regions[40,211].

8.2.2. Missing heritability

Despite the number of genetic variants discovered and large numbers of individuals studied, the amount of heritability explained remains low[212]. Many explanations have been proposed, including structural variants, rare variants with larger effects and lack of understanding of the role of effect modifier of other genes (gene-gene interactions) or the environment and lifestyle (gene-environment interactions). The ability to estimate heritability has also been widely debated. In general, heritability refers to the narrow sense heritability which is the amount of phenotypic variation attributable to the additive effects of genetic variants. The prevailing view is that many additional variants remained to be discovered such as low frequency disease implicated variants. However, this narrow sense heritability fails to take into account dominance genetic effects and genetic interactions, and does not include the other important contributor to phenotypic variation, the degree of environmental variations[213,214].

a. Structural variants

Structural variants are genetic variants in the genome that are typically span 1kb or larger, taking the form of insertion/deletion (copy number variants CNVs), inversion and translocation. Copy number variants account for a major proportion of genetic polymorphism that are not attributed to SNPs and has been implicated in monogenic genomic disorder such as Charcot-Marie-Tooth disease[215] and more recently in schizophrenia[216,217], autism[218] and obesity[219,220]. As with SNP variation, there is substantial variation in copy number variants across populations[36]. It has been noted that existing SNPs on genotyping arrays do tag common CNVs in the European populations[221], though this is less so in non-European populations[222]. CNVs are poorly captured by genotyping arrays, especially in defining the breakpoints of the variants. While next

generation sequencing (process of determining the exact order of nucleotides in the DNA) aims to revolutionize structural variant studies, each method of sequencing still has their bias and much more computational work is needed to improve existing algorithms to map structural variants with greater confidence[223].

b. Rare variants and sequencing

The first generation genotyping arrays have focused on common variants typically occurring at 5% and above in a population. Could some of the missing heritability be explained by those rarer variants that occur in less than 5% of the populations? The ability to statistically detect an association becomes increasingly difficult as the allele frequency decreases, unless effect sizes are large[212]. In many instances, the variants discovered in genome-wide studies are found in non-coding regions of the genome. Individuals carrying rare variants with large effect size are more likely to be in the extreme spectrum of the complex traits or enriched in frequency in disease cases compared to controls[224-228]. Sanna and colleagues showed that by sequencing exons and flanking regions of seven LDL-C implicated genes in individuals with extremely high or low LDL-C values, the combination of common and rare variants doubled the amount of heritability explained[229]. These successful targeted sequencing examples have showed that individuals at the extreme spectrum tended to carry an excess of rare variants, with increased genetic heritability.

It is possible that the same rare variants could have differential effects on the phenotype and might be located across large stretches of the genome. Rather than to look at each rare variant on its own, similar to the single SNP based approach, aggregating rare variants to look at cumulative burden of rare variants on diseases has becomes more appealing[212]. There have been less successful examples of targeted sequencing at implicated genes[230]. A synthetic association theory has been proposed, that rare-disease causing variants with much bigger effect size (here referring

to variants less commonly found in genome-wide scans) could lead to the genome-wide association signals detected in common variants and these disease causing variants could be located megabases away from the original genome-wide signals[230]. These rare variants could be found on the same disease-causing haplotype, that is tagged by some common-occurring variant[228,230]. In Type 2 Diabetes that has been studied across multiple populations with different ancestries, the transferability of the association signals across global populations suggest that the hypothesis that rare variants are driving these common shared associations is less likely to be true. With increasingly high sequencing throughput and rapidly falling costs, it is now possible to sequence whole exomes, or even whole genomes, to improve our catalog of human diversity and to look for disease implicated variants. These new variants discovered could be responsible for the genome-wide association signals and thus bringing us a step closer to the causal/functional variants.

c. Gene-gene interactions / gene-environment interactions / epigenetics

The joint effects of genetic variants on common diseases, and the interplay of genetic variants in specific environmental factors have not been fully explored[231]. The difficulty in interaction analyses lies in the power and the environmental exposure measurements. Besides the effect size of interaction and allele frequency of the variant, the prevalence of environmental exposure and burden of multiple testing influence the ability to detect any statistical interaction. While methods to ascertain genotypes have greatly improved the accuracy and validity of genotype calling, harmonization of exposure measurements across studies and in particular across heterogeneous populations remains a challenge. Methods to measure environmental exposures accurately have been less successful than in the genetic field. The fundamental bias problems in traditional epidemiological studies also return to haunt us, such as confounding, information bias; selection bias and reverse causation. Biological interactions could include genetic variants that exhibit

synergistic effects without marginal main effects, or environmental factors that are only "turned on" in certain genetically susceptible individuals, or interaction in the presence of both genetic and environmental risk factors. Current analytical methods include looking at risk dosages accumulated over implicated variants, candidate genes which showed an association with phenotype and also in case only designs that assumes independence in the original source populations.

More recently, epigenetic that describes the heritable gene expression without equivalent information stored in DNA has garnered widespread interest, and has been linked to gene-environment interaction[232]. Epigenetic modifies gene activities without corresponding changes in the underlying DNA code and has been suggested to be a dynamic and reversible process that is triggered by environmental influences. In Type 2 Diabetes, a common disease with strong genetic component and environmental influences, epigenetic mechanisms such as maternal nutrition and metabolism[233] and growth regulation[234] are of growing interest in the long term progression of the disease and affiliated risk factors like obesity.

d. Pleiotropy

Pleiotrophy is a phenomenon where a genetic locus is associated with multiple phenotypes. Carlson in 2004 had suggested that the use of intermediate quantitative phenotypes would increase the proportion of variance explained by a given locus than in the eventual clinical endpoint[75]. On the other hand, the correlation between these intermediate phenotypes and clinical endpoints could induce correlation or interaction between the genetic variants and phenotypes. The inability to differentiate induced correlation and interaction with true association findings will inflate or deflate heritability measures. In lipids phenotypes, many loci are also associated with more than one lipid traits[63]. On chromosome 12q24, there has been a variety of association

signals in celiac disease[235], blood pressure[150-152], hematological parameters such as platelets[236], red blood cells[237] and leukocytes[238], retinal venular caliber[239], myocardial infarction[238] and coronary heart disease[236]. These are blood vessels related complex traits which are inter-related, for instance, blood pressure is a pre-cursor to myocardial infarction and coronary heart disease, however, the locus points towards an inflammatory signaling pathway in endothelial cells that has an effect on blood pressure regulation and development of atherosclerosis[152,240].

CHAPTER 9 – CONCLUSION

The field of genetic mapping is moving beyond finding association between phenotypic variation and genetic variants to establish biological mechanisms and fine map the causal variants. These studies are timely in highlighting (i) the importance of understanding inter-population genetic diversity; (ii) the transferability and consistency of association signals across populations and (iii) the potential for non-European populations to discover disease implicated variants. Genome-wide association studies of common diseases and complex traits have showed some degree of shared genetic susceptibility across global populations, suggesting shared causal variants underlying disease pathogenesis. Different evolutionary history and migratory patterns in worldwide populations result in different allelic spectrum, where some variants are more common in some populations than others. Though sequencing costs have decreased rapidly, it is still not affordable to look at every single genetic variant in the genome for association with diseases and traits in substantial number of subjects. Exome sequencing is a more efficient strategy to whole genome sequencing that targets the genetic variants in the coding regions (2- 3% of the genome), motivated by understanding of functional changes in the genome sequences. Studies assimilating multi-ethnic populations will be in a better position to discover casual variants that are relatively common across populations or multiple low frequency variants in some or all of the populations. These developments are crucial and should work in conjunction with rapid technological advancements in the genomics field. I believe these studies emphasize the significance of studying multi-ethnic populations that elucidates the underlying genetic architecture of common disease and complex traits.

References

1. Tsui L.C., Buchwald M., Barker D., Braman J.C., Knowlton R., Schumm J.W., et al. *Cystic fibrosis locus defined by a genetically linked polymorphic DNA marker*. Science, 1985. **230**(4729): p. 1054-7.

2. Badano J.L. and Katsanis N. *Beyond Mendel: an evolving view of human genetic disease transmission.* Nat Rev Genet, 2002. **3**(10): p. 779-89.

3. Fisher R.A. *The Correlation Between Relatives on the Supposition of Mendelian Inheritance* Trans Roy Soc Edinb, 1918. **52**: p. 399-433.

4. Risch N.J. *Searching for genetic determinants in the new millennium.* Nature, 2000. **405**(6788): p. 847-56.

5. Altshuler D., Hirschhorn J.N., Klannemark M., Lindgren C.M., Vohl M.C., Nemesh J., et al. *The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes.* Nat Genet, 2000. **26**(1): p. 76-80.

6. Gloyn A.L., Weedon M.N., Owen K.R., Turner M.J., Knight B.A., Hitman G., et al. *Large-scale association studies of variants in genes encoding the pancreatic beta-cell KATP channel subunits Kir6.2 (KCNJ11) and SUR1 (ABCC8) confirm that the KCNJ11 E23K variant is associated with type 2 diabetes.* Diabetes, 2003. **52**(2): p. 568-72.

7. Barroso I., Gurnell M., Crowley V.E., Agostini M., Schwabe J.W., Soos M.A., et al. *Dominant negative mutations in human PPARgamma associated with severe insulin resistance, diabetes mellitus and hypertension.* Nature, 1999. **402**(6764): p. 880-3.

8. Gloyn A.L., Pearson E.R., Antcliff J.F., Proks P., Bruining G.J., Slingerland A.S., et al. *Activating mutations in the gene encoding the ATP-sensitive potassium-channel subunit Kir6.2 and permanent neonatal diabetes.* N Engl J Med, 2004. **350**(18): p. 1838-49.

9. Reynisdottir I., Thorleifsson G., Benediktsson R., Sigurdsson G., Emilsson V., Einarsdottir A.S., et al. *Localization of a susceptibility gene for type 2 diabetes to chromosome 5q34-q35.2.* Am J Hum Genet, 2003. **73**(2): p. 323-35.

10. Grant S.F., Thorleifsson G., Reynisdottir I., Benediktsson R., Manolescu A., Sainz J., et al. *Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes.* Nat Genet, 2006. **38**(3): p. 320-3.

11. Helgason A., Palsson S., Thorleifsson G., Grant S.F., Emilsson V., Gunnarsdottir S., et al. *Refining the impact of TCF7L2 gene variants on type 2 diabetes and adaptive evolution.* Nat Genet, 2007. **39**(2): p. 218-25.

12. Voight B.F., Scott L.J., Steinthorsdottir V., Morris A.P., Dina C., Welch R.P., et al. *Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis.* Nat Genet, 2010.

13. Risch N. and Merikangas K. *The future of genetic studies of complex human diseases.* Science, 1996. **273**(5281): p. 1516-7.

14. *The International HapMap Project.* Nature, 2003. **426**(6968): p. 789-96.

15. Bodmer W. and Bonilla C. *Common and rare variants in multifactorial susceptibility to common diseases.* Nat Genet, 2008. **40**(6): p. 695-701.

16. Wang W.Y., Barratt B.J., Clayton D.G., and Todd J.A. *Genome-wide association studies: theoretical and practical concerns.* Nat Rev Genet, 2005. **6**(2): p. 109-18.

17. *A haplotype map of the human genome.* Nature, 2005. **437**(7063): p. 1299-320.

18. Frazer K.A., Ballinger D.G., Cox D.R., Hinds D.A., Stuve L.L., Gibbs R.A., et al. *A second generation human haplotype map of over 3.1 million SNPs.* Nature, 2007. **449**(7164): p. 851-61.

19. Klein R.J., Zeiss C., Chew E.Y., Tsai J.Y., Sackler R.S., Haynes C., et al. *Complement factor H polymorphism in age-related macular degeneration.* Science, 2005. **308**(5720): p. 385-9.

20. Hindorff L.A., Junkins H.A., Hall P.N., Mehta J.P., and Manolio T.A. *A Catalog of Published Genome-Wide Association Studies.* 2011 06/07/2011]; Available from: http://www.genome.gov/gwastudies/.

21. Hindorff L.A., Sethupathy P., Junkins H.A., Ramos E.M., Mehta J.P., Collins F.S., et al. *Potential etiologic and functional implications of genome-wide association loci for human diseases and traits.* Proc Natl Acad Sci U S A, 2009. **106**(23): p. 9362-7.

22. Hartl D.L. and Clark A.R. eds. *Principles of Population Genetics*. 2007, Sinauer Associates: Massachusetts.

23. Rosenberg N.A., Pritchard J.K., Weber J.L., Cann H.M., Kidd K.K., Zhivotovsky L.A., et al. *Genetic structure of human populations.* Science, 2002. **298**(5602): p. 2381-5.

24. Reich D.E., Cargill M., Bolk S., Ireland J., Sabeti P.C., Richter D.J., et al. *Linkage disequilibrium in the human genome.* Nature, 2001. **411**(6834): p. 199-204.

25. Daly M.J., Rioux J.D., Schaffner S.F., Hudson T.J., and Lander E.S. *High-resolution haplotype structure in the human genome.* Nat Genet, 2001. **29**(2): p. 229-32.

26. Gabriel S.B., Schaffner S.F., Nguyen H., Moore J.M., Roy J., Blumenstiel B., et al. *The structure of haplotype blocks in the human genome.* Science, 2002. **296**(5576): p. 2225-9.

27. Carlson C.S., Eberle M.A., Rieder M.J., Smith J.D., Kruglyak L., and Nickerson D.A. *Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans.* Nat Genet, 2003. **33**(4): p. 518-21.

28. Phillips M.S., Lawrence R., Sachidanandam R., Morris A.P., Balding D.J., Donaldson M.A., et al. *Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots.* Nat Genet, 2003. **33**(3): p. 382-7.

29. Carlson C.S., Eberle M.A., Rieder M.J., Yi Q., Kruglyak L., and Nickerson D.A. *Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium.* Am J Hum Genet, 2004. **74**(1): p. 106-20.

30. de Bakker P.I., Yelensky R., Pe'er I., Gabriel S.B., Daly M.J., and Altshuler D. *Efficiency and power in genetic association studies.* Nat Genet, 2005. **37**(11): p. 1217-23.

31. Lewontin R.C. *The Interaction of Selection and Linkage. Ii. Optimum Models.* Genetics, 1964. **50**: p. 757-82.

32. Lewontin R.C. *The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models.* Genetics, 1964. **49**(1): p. 49-67.

33. Hill W.G. and Robertson A. *Linkage disequilibrium in finite populations.* Theor App Genet, 1968(38): p. 226-231.

34. de Bakker P.I., Burtt N.P., Graham R.R., Guiducci C., Yelensky R., Drake J.A., et al. *Transferability of tag SNPs in genetic association studies in multiple populations.* Nat Genet, 2006. **38**(11): p. 1298-303.

35. Conrad D.F., Jakobsson M., Coop G., Wen X., Wall J.D., Rosenberg N.A., et al. *A worldwide survey of haplotype variation and linkage disequilibrium in the human genome.* Nat Genet, 2006. **38**(11): p. 1251-60.

36. Altshuler D.M., Gibbs R.A., Peltonen L., Dermitzakis E., Schaffner S.F., Yu F., et al. *Integrating common and rare genetic variation in diverse human populations.* Nature, 2010. **467**(7311): p. 52-8.

37. Affymetrix. *Genome-Wide Human SNP Array 6.0*. 2009. http://media.affymetrix.com/support/technical/datasheets/genomewide_snp6_datasheet.pdf.

38. Affymetrix. *GeneChip® Human Mapping 10K Array Xba 142 2.0* 2004. http://media.affymetrix.com/support/technical/datasheets/10k2_datasheet.pdf.

39. Illumina. *Genome-Wide DNA Analysis BeadChips*. 2010. http://www.illumina.com/Documents/products/datasheets/datasheet_infiniumhd.pdf.

40. *A map of human genome variation from population-scale sequencing.* Nature, 2011. **467**(7319): p. 1061-73.

41. Illumina. *The Omni Family of Microarrays*. 2010. http://www.illumina.com/documents/products/datasheets/datasheet_gwas_roadmap.pdf.

42. Di X., Matsuzaki H., Webster T.A., Hubbell E., Liu G., Dong S., et al. *Dynamic model based algorithms for screening and genotyping over 100 K SNPs on oligonucleotide microarrays.* Bioinformatics, 2005. **21**(9): p. 1958-63.

43. Affymetrix. *BRLMM: an improved genotype calling method for the GeneChip Human Mapping 500K Array Set.* 2006.

44. Rabbee N. and Speed T.P. *A genotype calling algorithm for affymetrix SNP arrays.* Bioinformatics, 2006. **22**(1): p. 7-12.

45. *Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.* Nature, 2007. **447**(7145): p. 661-78.

46. Korn J.M., Kuruvilla F.G., McCarroll S.A., Wysoker A., Nemesh J., Cawley S., et al. *Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs.* Nat Genet, 2008. **40**(10): p. 1253-60.

47. Oliphant A., Barker D.L., Stuelpnagel J.R., and Chee M.S. *BeadArray technology: enabling an accurate, cost-effective approach to high-throughput genotyping.* Biotechniques, 2002. **Suppl**: p. 56-8, 60-1.

48. Fan J.B., Oliphant A., Shen R., Kermani B.G., Garcia F., Gunderson K.L., et al. *Highly Parallel SNP Genotyping*. in *Cold Spring Harbor Symposia on Quantitative Biology*. 2004. Cold Spring Harbor Laboratory Press.

49. Teo Y.Y., Inouye M., Small K.S., Gwilliam R., Deloukas P., Kwiatkowski D.P., et al. *A genotype calling algorithm for the Illumina BeadArray platform.* Bioinformatics, 2007. **23**(20): p. 2741-6.

50. Rosenberg N.A., Huang L., Jewett E.M., Szpiech Z.A., Jankovic I., and Boehnke M. *Genome-wide association studies in diverse populations.* Nat Rev Genet, 2010. **11**(5): p. 356-66.

51. Chakravarti A. *Human genetics: Tracing India's invisible threads.* Nature, 2009. **461**(7263): p. 487-8.

52. Reich D., Thangaraj K., Patterson N., Price A.L., and Singh L. *Reconstructing Indian population history.* Nature, 2009. **461**(7263): p. 489-94.

53. Clark A.G., Hubisz M.J., Bustamante C.D., Williamson S.H., and Nielsen R. *Ascertainment bias in studies of human genome-wide polymorphism.* Genome Res, 2005. **15**(11): p. 1496-502.

54. Marchini J., Howie B., Myers S., McVean G., and Donnelly P. *A new multipoint method for genome-wide association studies by imputation of genotypes.* Nat Genet, 2007. **39**(7): p. 906-13.

55. Li Y., Willer C.J., Ding J., Scheet P., and Abecasis G.R. *MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes.* Genet Epidemiol, 2010. **34**(8): p. 816-34.

56. Browning B.L. and Browning S.R. *A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals.* Am J Hum Genet, 2009. **84**(2): p. 210-23.

57. Guan Y. and Stephens M. *Practical issues in imputation-based association mapping.* PLoS Genet, 2008. **4**(12): p. e1000279.

58. Howie B.N., Donnelly P., and Marchini J. *A flexible and accurate genotype imputation method for the next generation of genome-wide association studies.* PLoS Genet, 2009. **5**(6): p. e1000529.

59. Huang L., Li Y., Singleton A.B., Hardy J.A., Abecasis G., Rosenberg N.A., et al. *Genotype-imputation accuracy across worldwide human populations.* Am J Hum Genet, 2009. **84**(2): p. 235-50.

60. Pei Y.F., Li J., Zhang L., Papasian C.J., and Deng H.W. *Analyses and comparison of accuracy of different genotype imputation methods.* PLoS One, 2008. **3**(10): p. e3551.

61. Jallow M., Teo Y.Y., Small K.S., Rockett K.A., Deloukas P., Clark T.G., et al. *Genome-wide and fine-resolution association analysis of malaria in West Africa.* Nat Genet, 2009. **41**(6): p. 657-65.

62. Kathiresan S., Melander O., Guiducci C., Surti A., Burtt N.P., Rieder M.J., et al. *Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans.* Nat Genet, 2008. **40**(2): p. 189-97.

63. Teslovich T.M., Musunuru K., Smith A.V., Edmondson A.C., Stylianou I.M., Koseki M., et al. *Biological, clinical and population relevance of 95 loci for blood lipids.* Nature, 2010. **466**(7307): p. 707-13.

64. Teo Y.Y. and Sim X. *Patterns of linkage disequilibrium in different populations: implications and opportunities for lipid-associated loci identified from genome-wide association studies.* Curr Opin Lipidol, 2010. **21**(2): p. 104-15.

65. Kooner J.S., Chambers J.C., Aguilar-Salinas C.A., Hinds D.A., Hyde C.L., Warnes G.R., et al. *Genome-wide scan identifies variation in MLXIPL associated with plasma triglycerides.* Nat Genet, 2008. **40**(2): p. 149-51.

66. Johnson A.D., Handsaker R.E., Pulit S.L., Nizzari M.M., O'Donnell C.J., and de Bakker P.I. *SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap.* Bioinformatics, 2008. **24**(24): p. 2938-9.

67. Yasuda K., Miyake K., Horikawa Y., Hara K., Osawa H., Furuta H., et al. *Variants in KCNQ1 are associated with susceptibility to type 2 diabetes mellitus.* Nat Genet, 2008. **40**(9): p. 1092-7.

68. Unoki H., Takahashi A., Kawaguchi T., Hara K., Horikoshi M., Andersen G., et al. *SNPs in KCNQ1 are associated with susceptibility to type 2 diabetes in East Asian and European populations.* Nat Genet, 2008. **40**(9): p. 1098-102.

69. Tai E.S., Sim X.L., Ong T.H., Wong T.Y., Saw S.M., Aung T., et al. *Polymorphisms at newly identified lipid-associated loci are associated with blood lipids and cardiovascular disease in an Asian Malay population.* J Lipid Res, 2009. **50**(3): p. 514-20.

70. Teo Y.Y., Sim X., Ong R.T., Tan A.K., Chen J., Tantoso E., et al. *Singapore Genome Variation Project: a haplotype map of three Southeast Asian populations.* Genome Res, 2009. **19**(11): p. 2154-62.

71. Rother K.I. *Diabetes treatment--bridging the divide.* N Engl J Med, 2007. **356**(15): p. 1499-501.

72. Prokopenko I., McCarthy M.I., and Lindgren C.M. *Type 2 diabetes: new genes, new understanding.* Trends Genet, 2008. **24**(12): p. 613-21.

73. McCarthy M.I. *Genomics, type 2 diabetes, and obesity.* N Engl J Med, 2010. **363**(24): p. 2339-50.

74. Ramachandran A., Ma R.C., and Snehalatha C. *Diabetes in Asia.* Lancet, 2010. **375**(9712): p. 408-18.

75. Carlson C.S., Eberle M.A., Kruglyak L., and Nickerson D.A. *Mapping complex disease loci in whole-genome association studies.* Nature, 2004. **429**(6990): p. 446-52.

76. Dupuis J., Langenberg C., Prokopenko I., Saxena R., Soranzo N., Jackson A.U., et al. *New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk.* Nat Genet, 2010. **42**(2): p. 105-16.

77. Heid I.M., Jackson A.U., Randall J.C., Winkler T.W., Qi L., Steinthorsdottir V., et al. *Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution.* Nat Genet, 2010. **42**(11): p. 949-60.

78. Speliotes E.K., Willer C.J., Berndt S.I., Monda K.L., Thorleifsson G., Jackson A.U., et al. *Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index.* Nat Genet, 2010. **42**(11): p. 937-48.

79. Koo S.H., Ho W.F., and Lee E.J. *Genetic polymorphisms in KCNQ1, HERG, KCNE1 and KCNE2 genes in the Chinese, Malay and Indian populations of Singapore.* British journal of clinical pharmacology, 2006. **61**(3): p. 301-8.

80. Ng D.P., Fukushima M., Tai B.C., Koh D., Leong H., Imura H., et al. *Reduced GFR and albuminuria in Chinese type 2 diabetes mellitus patients are both independently associated with activation of the TNF-alpha system.* Diabetologia, 2008. **51**(12): p. 2318-24.

81. Hughes K., Yeo P.P., Lun K.C., Thai A.C., Sothy S.P., Wang K.W., et al. *Cardiovascular diseases in Chinese, Malays, and Indians in Singapore. II. Differences in risk factor levels.* J Epidemiol Community Health, 1990. **44**(1): p. 29-35.

82. Tan C.E., Emmanuel S.C., Tan B.Y., and Jacob E. *Prevalence of diabetes and ethnic differences in cardiovascular risk factors. The 1992 Singapore National Health Survey.* Diabetes Care, 1999. **22**(2): p. 241-7.

83. Hughes K., Aw T.C., Kuperan P., and Choo M. *Central obesity, insulin resistance, syndrome X, lipoprotein(a), and cardiovascular risk in Indians, Malays, and Chinese in Singapore.* J Epidemiol Community Health, 1997. **51**(4): p. 394-9.

84. Cutter J., Tan B.Y., and Chew S.K. *Levels of cardiovascular disease risk factors in Singapore following a national intervention programme.* Bull World Health Organ, 2001. **79**(10): p. 908-15.

85. Nang E.E., Khoo C.M., Tai E.S., Lim S.C., Tavintharan S., Wong T.Y., et al. *Is there a clear threshold for fasting plasma glucose that differentiates between those with and without neuropathy and chronic kidney disease?: the Singapore Prospective Study Program.* Am J Epidemiol, 2009. **169**(12): p. 1454-62.

86. Leow B. *Singapore Census of Population 2000: Statistical Release 1 - Demographic Characteristics.*, Statistics D.o., Editor 2001: Singapore.

87. Foong A.W., Saw S.M., Loo J.L., Shen S., Loon S.C., Rosman M., et al. *Rationale and methodology for a population-based study of eye diseases in Malay people: The Singapore Malay eye study (SiMES).* Ophthalmic Epidemiol, 2007. **14**(1): p. 25-35.

88. *Standards of medical care in diabetes--2011.* Diabetes Care, 2011. **34 Suppl 1**: p. S11-61.

89. Lavanya R., Jeganathan V.S., Zheng Y., Raju P., Cheung N., Tai E.S., et al. *Methodology of the Singapore Indian Chinese Cohort (SICC) eye study: quantifying ethnic variations in the epidemiology of eye diseases in Asians.* Ophthalmic Epidemiol, 2009. **16**(6): p. 325-36.

90. Chambers J.C., Zhao J., Terracciano C.M., Bezzina C.R., Zhang W., Kaba R., et al. *Genetic variation in SCN10A influences cardiac conduction.* Nat Genet, 2010. **42**(2): p. 149-52.

91. Chambers J.C., Zhang W., Zabaneh D., Sehmi J., Jain P., McCarthy M.I., et al. *Common genetic variation near melatonin receptor MTNR1B contributes to raised plasma glucose and increased risk of type 2 diabetes among Indian Asians and European Caucasians.* Diabetes, 2009. **58**(11): p. 2703-8.

92. Saleheen D., Zaidi M., Rasheed A., Ahmad U., Hakeem A., Murtaza M., et al. *The Pakistan Risk of Myocardial Infarction Study: a resource for the study of genetic, lifestyle and other determinants of myocardial infarction in South Asia.* Eur J Epidemiol, 2009. **24**(6): p. 329-38.

93. *A genome-wide association study in Europeans and South Asians identifies five new loci for coronary artery disease.* Nature genetics, 2011. **43**(4): p. 339-44.

94. Jafar T.H., Hatcher J., Poulter N., Islam M., Hashmi S., Qadri Z., et al. *Community-based interventions to promote blood pressure control in a developing country: a cluster randomized trial.* Ann Intern Med, 2009. **151**(9): p. 593-601.

95. Chidambaram M., Radha V., and Mohan V. *Replication of recently described type 2 diabetes gene variants in a South Indian population.* Metabolism, 2010. **59**(12): p. 1760-6.

96. Rees S.D., Islam M., Hydrie M.Z., Chaudhary B., Bellary S., Hashmi S., et al. *An FTO variant is associated with Type 2 diabetes in South Asian populations after accounting for body mass index and waist circumference.* Diabet Med, 2011. **28**(6): p. 673-680.

97. Jowett J.B., Diego V.P., Kotea N., Kowlessur S., Chitson P., Dyer T.D., et al. *Genetic influences on type 2 diabetes and metabolic syndrome related quantitative traits in Mauritius.* Twin Res Hum Genet, 2009. **12**(1): p. 44-52.

98. Takeuchi F., Katsuya T., Chakrewarthy S., Yamamoto K., Fujioka A., Serizawa M., et al. *Common variants at the GCK, GCKR, G6PC2-ABCB11 and MTNR1B loci are associated with fasting glucose in two Asian populations.* Diabetologia, 2010. **53**(2): p. 299-308.

99. Sanghera D.K., Bhatti J.S., Bhatti G.K., Ralhan S.K., Wander G.S., Singh J.R., et al. *The Khatri Sikh Diabetes Study (SDS): study design, methodology, sample collection, and initial results.* Hum Biol, 2006. **78**(1): p. 43-63.

100. Katulanda P., Constantine G.R., Mahesh J.G., Sheriff R., Seneviratne R.D., Wijeratne S., et al. *Prevalence and projections of diabetes and pre-diabetes in adults in Sri Lanka--Sri Lanka Diabetes, Cardiovascular Study (SLDCS).* Diabet Med, 2008. **25**(9): p. 1062-9.

101. Bellary S., O'Hare J.P., Raymond N.T., Gumber A., Mughal S., Szczepura A., et al. *Enhanced diabetes care to patients of south Asian ethnic origin (the United Kingdom Asian Diabetes Study): a cluster randomised controlled trial.* Lancet, 2008. **371**(9626): p. 1769-76.

102. Illumina. *Human1M-Duo DNA Analysis BeadChip Kits*. 2011. http://www.illumina.com/products/human1m_duo_dna_analysis_beadchip_kits.ilmn.

103. Illumina. *Sentrix® HumanHap300 Genotyping BeadChip.* 2006.

104. Teo Y.Y. *Exploratory data analysis in large-scale genetic studies.* Biostatistics, 2010. **11**(1): p. 70-81.

105. Purcell S., Neale B., Todd-Brown K., Thomas L., Ferreira M.A., Bender D., et al. *PLINK: a tool set for whole-genome association and population-based linkage analyses.* Am J Hum Genet, 2007. **81**(3): p. 559-75.

106. R D.C.T. *R: A language and environment for statistical computing*.

107. Price A.L., Patterson N.J., Plenge R.M., Weinblatt M.E., Shadick N.A., and Reich D. *Principal components analysis corrects for stratification in genome-wide association studies.* Nat Genet, 2006. **38**(8): p. 904-9.

108. Patterson N., Price A.L., and Reich D. *Population structure and eigenanalysis.* PLoS Genet, 2006. **2**(12): p. e190.

109. Kooner J.S., Saleheen D., Sim X., Sehmi J., Zhang W., Frossard P., et al. *Genome-wide association study in individuals of South Asian ancestry identifies six new type 2 diabetes susceptibility loci.* Nat Genet, 2011.

110. *Appropriate body-mass index for Asian populations and its implications for policy and intervention strategies.* Lancet, 2004. **363**(9403): p. 157-63.

111. Low S., Chin M.C., Ma S., Heng D., and Deurenberg-Yap M. *Rationale for redefining obesity in Asians.* Ann Acad Med Singapore, 2009. **38**(1): p. 66-9.

112. Marchini J., Cardon L.R., Phillips M.S., and Donnelly P. *The effects of human population structure on large genetic association studies.* Nat Genet, 2004. **36**(5): p. 512-7.

113. Devlin B., Roeder K., and Wasserman L. *Genomic control, a new approach to genetic-based association studies.* Theor Popul Biol, 2001. **60**(3): p. 155-66.

114. Reich D.E. and Goldstein D.B. *Detecting association in a case-control study while correcting for population stratification.* Genet Epidemiol, 2001. **20**(1): p. 4-16.

115. Sim X., Ong R.T., Suo C., Tay W.T., Liu J., Ng D.P., et al. *Transferability of type 2 diabetes implicated loci in multi-ethnic cohorts from Southeast Asia.* PLoS Genet, 2011. **7**(4): p. e1001363.

116. McCarthy M.I., Abecasis G.R., Cardon L.R., Goldstein D.B., Little J., Ioannidis J.P., et al. *Genome-wide association studies for complex traits: consensus, uncertainty and challenges.* Nature reviews. Genetics, 2008. **9**(5): p. 356-69.

117. Lettre G., Lange C., and Hirschhorn J.N. *Genetic model testing and statistical power in population-based association studies of quantitative traits.* Genetic epidemiology, 2007. **31**(4): p. 358-62.

118. Pe'er I., Yelensky R., Altshuler D., and Daly M.J. *Estimation of the multiple testing burden for genomewide association studies of nearly all common variants.* Genetic epidemiology, 2008. **32**(4): p. 381-5.

119. Servin B. and Stephens M. *Imputation-based analysis of association studies: candidate regions and quantitative traits.* PLoS Genet, 2007. **3**(7): p. e114.

120. Browning B.L. and Browning S.R. *Haplotypic analysis of Wellcome Trust Case Control Consortium data.* Hum Genet, 2008. **123**(3): p. 273-80.

121. Stephens M., Smith N.J., and Donnelly P. *A new statistical method for haplotype reconstruction from population data.* Am J Hum Genet, 2001. **68**(4): p. 978-89.

122. Li N. and Stephens M. *Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data.* Genetics, 2003. **165**(4): p. 2213-33.

123. Scheet P. and Stephens M. *A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase.* Am J Hum Genet, 2006. **78**(4): p. 629-44.

124. Barrett J.C., Fry B., Maller J., and Daly M.J. *Haploview: analysis and visualization of LD and haplotype maps.* Bioinformatics, 2005. **21**(2): p. 263-5.

125. Teo Y.Y., Fry A.E., Bhattacharya K., Small K.S., Kwiatkowski D.P., and Clark T.G. *Genome-wide comparisons of variation in linkage disequilibrium.* Genome Res, 2009. **19**(10): p. 1849-60.

126. Ong R.T., Liu X., Poh W.T., Sim X., Chia K.S., and Teo Y.Y. *A method for identifying haplotypes carrying the causative allele in positive natural selection and genome-wide association studies.* Bioinformatics, 2011. **27**(6): p. 822-8.

127. Sabeti P.C., Schaffner S.F., Fry B., Lohmueller J., Varilly P., Shamovsky O., et al. *Positive natural selection in the human lineage.* Science, 2006. **312**(5780): p. 1614-20.

128. Sabeti P.C., Reich D.E., Higgins J.M., Levine H.Z., Richter D.J., Schaffner S.F., et al. *Detecting recent positive selection in the human genome from haplotype structure.* Nature, 2002. **419**(6909): p. 832-7.

129. Voight B.F., Kudaravalli S., Wen X., and Pritchard J.K. *A map of recent positive selection in the human genome.* PLoS Biol, 2006. **4**(3): p. e72.

130. Sabeti P.C., Varilly P., Fry B., Lohmueller J., Hostetter E., Cotsapas C., et al. *Genome-wide detection and characterization of positive selection in human populations.* Nature, 2007. **449**(7164): p. 913-8.

131. Pickrell J.K., Coop G., Novembre J., Kudaravalli S., Li J.Z., Absher D., et al. *Signals of recent positive selection in a worldwide sample of human populations.* Genome research, 2009. **19**(5): p. 826-37.

132. Department of S. *Advance Census Release*. Census of Population 2010, 2010. http://www.singstat.gov.sg/pubn/popn/c2010acr.pdf.

133. Saw S.H. *The population of Singapore*. 2nd ed2007, Singapore: Institute of South East Asian Studies.

134. Saxena R., Voight B.F., Lyssenko V., Burtt N.P., de Bakker P.I., Chen H., et al. *Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels.* Science, 2007. **316**(5829): p. 1331-6.

135. Scott L.J., Mohlke K.L., Bonnycastle L.L., Willer C.J., Li Y., Duren W.L., et al. *A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants.* Science, 2007. **316**(5829): p. 1341-5.

136. Zeggini E., Weedon M.N., Lindgren C.M., Frayling T.M., Elliott K.S., Lango H., et al. *Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes.* Science, 2007. **316**(5829): p. 1336-41.

137. Zeggini E., Scott L.J., Saxena R., Voight B.F., Marchini J.L., Hu T., et al. *Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes.* Nat Genet, 2008. **40**(5): p. 638-45.

138. Han X., Luo Y., Ren Q., Zhang X., Wang F., Sun X., et al. *Implication of genetic variants near SLC30A8, HHEX, CDKAL1, CDKN2A/B, IGF2BP2, FTO, TCF2, KCNQ1, and WFS1 in type 2 Diabetes in a Chinese population.* BMC Med Genet, 2010. **11**: p. 81.

139. Herder C., Rathmann W., Strassburger K., Finner H., Grallert H., Huth C., et al. *Variants of the PPARG, IGF2BP2, CDKAL1, HHEX, and TCF7L2 genes confer risk of type 2 diabetes independently of BMI in the German KORA studies.* Horm Metab Res, 2008. **40**(10): p. 722-6.

140. Hu C., Zhang R., Wang C., Wang J., Ma X., Lu J., et al. *PPARG, KCNJ11, CDKAL1, CDKN2A-CDKN2B, IDE-KIF11-HHEX, IGF2BP2 and SLC30A8 are associated with type 2 diabetes in a Chinese population.* PLoS One, 2009. **4**(10): p. e7643.

141. Lee Y.H., Kang E.S., Kim S.H., Han S.J., Kim C.H., Kim H.J., et al. *Association between polymorphisms in SLC30A8, HHEX, CDKN2A/B, IGF2BP2, FTO, WFS1, CDKAL1, KCNQ1 and type 2 diabetes in the Korean population.* J Hum Genet, 2008. **53**(11-12): p. 991-8.

142. Lin Y., Li P., Cai L., Zhang B., Tang X., Zhang X., et al. *Association study of genetic variants in eight genes/loci with type 2 diabetes in a Han Chinese population.* BMC Med Genet, 2010. **11**: p. 97.

143. Liu Y., Yu L., Zhang D., Chen Z., Zhou D.Z., Zhao T., et al. *Positive association between variations in CDKAL1 and type 2 diabetes in Han Chinese individuals.* Diabetologia, 2008. **51**(11): p. 2134-7.

144. Ng M.C., Park K.S., Oh B., Tam C.H., Cho Y.M., Shin H.D., et al. *Implication of genetic variants near TCF7L2, SLC30A8, HHEX, CDKAL1, CDKN2A/B, IGF2BP2, and FTO in type 2 diabetes and obesity in 6,719 Asians.* Diabetes, 2008. **57**(8): p. 2226-33.

145. Omori S., Tanaka Y., Takahashi A., Hirose H., Kashiwagi A., Kaku K., et al. *Association of CDKAL1, IGF2BP2, CDKN2A/B, HHEX, SLC30A8, and KCNJ11 with susceptibility to type 2 diabetes in a Japanese population.* Diabetes, 2008. **57**(3): p. 791-5.

146. Tsai F.J., Yang C.F., Chen C.C., Chuang L.M., Lu C.H., Chang C.T., et al. *A genome-wide association study identifies susceptibility variants for type 2 diabetes in Han Chinese.* PLoS Genet, 2010. **6**(2): p. e1000847.

147. Wen J., Ronn T., Olsson A., Yang Z., Lu B., Du Y., et al. *Investigation of type 2 diabetes risk alleles support CDKN2A/B, CDKAL1, and TCF7L2 as susceptibility genes in a Han Chinese cohort.* PLoS One, 2010. **5**(2): p. e9153.

148. Wu Y., Li H., Loos R.J., Yu Z., Ye X., Chen L., et al. *Common variants in CDKAL1, CDKN2A/B, IGF2BP2, SLC30A8, and HHEX/IDE genes are associated with type 2 diabetes and impaired fasting glucose in a Chinese Han population.* Diabetes, 2008. **57**(10): p. 2834-42.

149. Yamauchi T., Hara K., Maeda S., Yasuda K., Takahashi A., Horikoshi M., et al. *A genome-wide association study in the Japanese population identifies susceptibility loci for type 2 diabetes at UBE2E2 and C2CD4A-C2CD4B.* Nat Genet, 2010. **42**(10): p. 864-8.

150. Kato N., Takeuchi F., Tabara Y., Kelly T.N., Go M.J., Sim X., et al. *Meta-analysis of genome-wide association studies identifies common variants associated with blood pressure variation in east Asians.* Nat Genet, 2011. **43**(6): p. 531-8.

151. Levy D., Ehret G.B., Rice K., Verwoert G.C., Launer L.J., Dehghan A., et al. *Genome-wide association study of blood pressure and hypertension.* Nat Genet, 2009. **41**(6): p. 677-87.

152. Newton-Cheh C., Johnson T., Gateva V., Tobin M.D., Bochud M., Coin L., et al. *Genome-wide association study identifies eight loci associated with blood pressure.* Nat Genet, 2009. **41**(6): p. 666-76.

153. Osier M.V., Pakstis A.J., Soodyall H., Comas D., Goldman D., Odunsi A., et al. *A global perspective on genetic variation at the ADH genes reveals unusual patterns of linkage disequilibrium and diversity.* Am J Hum Genet, 2002. **71**(1): p. 84-99.

154. Lamason R.L., Mohideen M.A., Mest J.R., Wong A.C., Norton H.L., Aros M.C., et al. *SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans.* Science, 2005. **310**(5755): p. 1782-6.

155. Lao O., de Gruijter J.M., van Duijn K., Navarro A., and Kayser M. *Signatures of positive selection in genes associated with human skin pigmentation as revealed from analyses of single nucleotide polymorphisms.* Ann Hum Genet, 2007. **71**(Pt 3): p. 354-69.

156. Sulem P., Gudbjartsson D.F., Stacey S.N., Helgason A., Rafnar T., Magnusson K.P., et al. *Genetic determinants of hair, eye and skin pigmentation in Europeans.* Nat Genet, 2007. **39**(12): p. 1443-52.

157. Chan J.C., Malik V., Jia W., Kadowaki T., Yajnik C.S., Yoon K.H., et al. *Diabetes in Asia: epidemiology, risk factors, and pathophysiology.* JAMA, 2009. **301**(20): p. 2129-40.

158. Yang W., Lu J., Weng J., Jia W., Ji L., Xiao J., et al. *Prevalence of diabetes among men and women in China.* N Engl J Med, 2010. **362**(12): p. 1090-101.

159. Bouatia-Naji N., Bonnefond A., Cavalcanti-Proenca C., Sparso T., Holmkvist J., Marchand M., et al. *A variant near MTNR1B is associated with increased fasting plasma glucose levels and type 2 diabetes risk.* Nat Genet, 2009. **41**(1): p. 89-94.

160. Lyssenko V., Nagorny C.L., Erdos M.R., Wierup N., Jonsson A., Spegel P., et al. *Common variant in MTNR1B associated with increased risk of type 2 diabetes and impaired early insulin secretion.* Nat Genet, 2009. **41**(1): p. 82-8.

161. Prokopenko I., Langenberg C., Florez J.C., Saxena R., Soranzo N., Thorleifsson G., et al. *Variants in MTNR1B influence fasting glucose levels.* Nat Genet, 2009. **41**(1): p. 77-81.

162. Takeuchi F., Serizawa M., Yamamoto K., Fujisawa T., Nakashima E., Ohnaka K., et al. *Confirmation of multiple risk Loci and genetic impacts by a genome-wide association study of type 2 diabetes in the Japanese population.* Diabetes, 2009. **58**(7): p. 1690-9.

163. Xiao R. and Boehnke M. *Quantifying and correcting for the winner's curse in genetic association studies.* Genetic epidemiology, 2009. **33**(5): p. 453-62.

164. Zollner S. and Pritchard J.K. *Overcoming the winner's curse: estimating penetrance parameters from case-control data.* American journal of human genetics, 2007. **80**(4): p. 605-15.

165. Diamond J. *Medicine: diabetes in India.* Nature, 2011. **469**(7331): p. 478-9.

166. Shaw J.E., Sicree R.A., and Zimmet P.Z. *Global estimates of the prevalence of diabetes for 2010 and 2030.* Diabetes Res Clin Pract, 2010. **87**(1): p. 4-14.

167. McKeigue P.M., Shah B., and Marmot M.G. *Relation of central obesity and insulin resistance with high diabetes prevalence and cardiovascular risk in South Asians.* Lancet, 1991. **337**(8738): p. 382-6.

168. Ramachandran A., Snehalatha C., Viswanathan V., Viswanathan M., and Haffner S.M. *Risk of noninsulin dependent diabetes mellitus conferred by obesity and central adiposity in different ethnic groups: a comparative analysis between Asian Indians, Mexican Americans and Whites.* Diabetes Res Clin Pract, 1997. **36**(2): p. 121-5.

169. Raji A., Seely E.W., Arky R.A., and Simonson D.C. *Body fat distribution and insulin resistance in healthy Asian Indians and Caucasians.* J Clin Endocrinol Metab, 2001. **86**(11): p. 5366-71.

170. Chandalia M., Abate N., Garg A., Stray-Gundersen J., and Grundy S.M. *Relationship between generalized and upper body obesity to insulin resistance in Asian Indian men.* J Clin Endocrinol Metab, 1999. **84**(7): p. 2329-35.

171. Deepa R., Sandeep S., and Mohan V. *Abdominal obesity, visceral fat and type 2 diabetes - Asian Indian phenotype*, in *Type 2 diabetes in South Asians:Epidemiology, risk factors and prevention*, Mohan V. and Rao G.H.R., Editors. 2006, Jaypee Brothers Medical Publishers (P) Ltd: New Delhi. p. 138-52.

172. Matthews D.R., Hosker J.P., Rudenski A.S., Naylor B.A., Treacher D.F., and Turner R.C. *Homeostasis model assessment: insulin resistance and beta-cell function from fasting plasma glucose and insulin concentrations in man.* Diabetologia, 1985. **28**(7): p. 412-9.

173. Dufresne A.M. and Smith R.J. *The adapter protein GRB10 is an endogenous negative regulator of insulin-like growth factor signaling.* Endocrinology, 2005. **146**(10): p. 4399-409.

174. Depetris R.S., Wu J., and Hubbard S.R. *Structural and functional studies of the Ras-associating and pleckstrin-homology domains of Grb10 and Grb14.* Nat Struct Mol Biol, 2009. **16**(8): p. 833-9.

175. Holt L.J., Lyons R.J., Ryan A.S., Beale S.M., Ward A., Cooney G.J., et al. *Dual ablation of Grb10 and Grb14 in mice reveals their combined role in regulation of insulin signaling and glucose homeostasis.* Mol Endocrinol, 2009. **23**(9): p. 1406-14.

176. Woodard-Grice A.V., McBrayer A.C., Wakefield J.K., Zhuo Y., and Bellis S.L. *Proteolytic shedding of ST6Gal-I by BACE1 regulates the glycosylation and function of alpha4beta1 integrins.* J Biol Chem, 2008. **283**(39): p. 26364-73.

177. Siitonen N., Pulkkinen L., Lindstrom J., Kolehmainen M., Eriksson J.G., Venojarvi M., et al. *Association of ADIPOQ gene variants with body weight, type 2 diabetes and serum adiponectin concentrations: the Finnish Diabetes Prevention Study.* BMC Med Genet, 2011. **12**: p. 5.

178. Maeda N., Shimomura I., Kishida K., Nishizawa H., Matsuda M., Nagaretani H., et al. *Diet-induced insulin resistance in mice lacking adiponectin/ACRP30.* Nat Med, 2002. **8**(7): p. 731-7.

179. Seaman M.N., Harbour M.E., Tattersall D., Read E., and Bright N. *Membrane recruitment of the cargo-selective retromer subcomplex is catalysed by the small GTPase Rab7 and inhibited by the Rab-GAP TBC1D5.* J Cell Sci, 2009. **122**(Pt 14): p. 2371-82.

180. Seaman M.N., Marcusson E.G., Cereghino J.L., and Emr S.D. *Endosome to Golgi retrieval of the vacuolar protein sorting receptor, Vps10p, requires the function of the VPS29, VPS30, and VPS35 gene products.* J Cell Biol, 1997. **137**(1): p. 79-92.

181. Kim E., Lee J.W., Baek D.C., Lee S.R., Kim M.S., Kim S.H., et al. *Identification of novel retromer complexes in the mouse testis.* Biochem Biophys Res Commun, 2008. **375**(1): p. 16-21.

182. Artegiani B., Labbaye C., Sferra A., Quaranta M.T., Torreri P., Macchia G., et al. *The interaction with HMG20a/b proteins suggests a potential role for beta-dystrobrevin in neuronal differentiation.* J Biol Chem, 2010. **285**(32): p. 24740-50.

183. Sumoy L., Carim L., Escarceller M., Nadal M., Gratacos M., Pujana M.A., et al. *HMG20A and HMG20B map to human chromosomes 15q24 and 19p13.3 and constitute a distinct class of HMG-box genes with ubiquitous expression.* Cytogenet Cell Genet, 2000. **88**(1-2): p. 62-7.

184. Dell'Angelica E.C., Ohno H., Ooi C.E., Rabinovich E., Roche K.W., and Bonifacino J.S. *AP-3: an adaptor-like protein complex with ubiquitous expression.* EMBO J, 1997. **16**(5): p. 917-28.

185. Beller M., Bulankina A.V., Hsiao H.H., Urlaub H., Jackle H., and Kuhnlein R.P. *PERILIPIN-dependent control of lipid droplet structure and fat storage in Drosophila.* Cell Metab, 2010. **12**(5): p. 521-32.

186. Qi L., Corella D., Sorli J.V., Portoles O., Shen H., Coltell O., et al. *Genetic variation at the perilipin (PLIN) locus is associated with obesity-related phenotypes in White women.* Clin Genet, 2004. **66**(4): p. 299-310.

187. Brasaemle D.L., Rubin B., Harten I.A., Gruia-Gray J., Kimmel A.R., and Londos C. *Perilipin A increases triacylglycerol storage by decreasing the rate of triacylglycerol hydrolysis.* J Biol Chem, 2000. **275**(49): p. 38486-93.

188. Yamagata K., Furuta H., Oda N., Kaisaki P.J., Menzel S., Cox N.J., et al. *Mutations in the hepatocyte nuclear factor-4alpha gene in maturity-onset diabetes of the young (MODY1).* Nature, 1996. **384**(6608): p. 458-60.

189. Battle M.A., Konopka G., Parviz F., Gaggl A.L., Yang C., Sladek F.M., et al. *Hepatocyte nuclear factor 4alpha orchestrates expression of cell adhesion proteins during the epithelial transformation of the developing liver.* Proc Natl Acad Sci U S A, 2006. **103**(22): p. 8419-24.

190. Orho-Melander M., Melander O., Guiducci C., Perez-Martinez P., Corella D., Roos C., et al. *Common missense variant in the glucokinase regulatory protein gene is associated with increased plasma triglyceride and C-reactive protein but lower fasting glucose concentrations.* Diabetes, 2008. **57**(11): p. 3112-21.

191. Sparso T., Andersen G., Nielsen T., Burgdorf K.S., Gjesing A.P., Nielsen A.L., et al. *The GCKR rs780094 polymorphism is associated with elevated fasting serum triacylglycerol, reduced fasting and OGTT-related insulinaemia, and reduced risk of type 2 diabetes.* Diabetologia, 2008. **51**(1): p. 70-5.

192. Vaxillaire M., Cavalcanti-Proenca C., Dechaume A., Tichet J., Marre M., Balkau B., et al. *The common P446L polymorphism in GCKR inversely modulates fasting glucose and triglyceride levels and reduces type 2 diabetes risk in the DESIR prospective general French population.* Diabetes, 2008. **57**(8): p. 2253-7.

193. Zimmet P., Alberti K.G., and Shaw J. *Global and societal implications of the diabetes epidemic.* Nature, 2001. **414**(6865): p. 782-7.

194. Deurenberg P., Deurenberg-Yap M., and Guricci S. *Asians are different from Caucasians and from each other in their body mass index/body fat per cent relationship.* Obes Rev, 2002. **3**(3): p. 141-6.

195. Deurenberg-Yap M., Chew S.K., and Deurenberg P. *Elevated body fat percentage and cardiovascular risks at low body mass index levels among Singaporean Chinese, Malays and Indians.* Obes Rev, 2002. **3**(3): p. 209-15.

196. Frayling T.M., Timpson N.J., Weedon M.N., Zeggini E., Freathy R.M., Lindgren C.M., et al. *A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity.* Science, 2007. **316**(5826): p. 889-94.

197. Sladek R., Rocheleau G., Rung J., Dina C., Shen L., Serre D., et al. *A genome-wide association study identifies novel risk loci for type 2 diabetes.* Nature, 2007. **445**(7130): p. 881-5.

198. Timpson N.J., Lindgren C.M., Weedon M.N., Randall J., Ouwehand W.H., Strachan D.P., et al. *Adiposity-related heterogeneity in patterns of type 2 diabetes susceptibility observed in genome-wide association data.* Diabetes, 2009. **58**(2): p. 505-10.

199. Pascoe L., Frayling T.M., Weedon M.N., Mari A., Tura A., Ferrannini E., et al. *Beta cell glucose sensitivity is decreased by 39% in non-diabetic individuals carrying multiple diabetes-risk alleles compared with those with no risk alleles.* Diabetologia, 2008. **51**(11): p. 1989-92.

200. Ohara-Imaizumi M., Yoshida M., Aoyagi K., Saito T., Okamura T., Takenaka H., et al. *Deletion of CDKAL1 affects mitochondrial ATP generation and first-phase insulin exocytosis.* PLoS One, 2010. **5**(12): p. e15553.

201. Steinthorsdottir V., Thorleifsson G., Reynisdottir I., Benediktsson R., Jonsdottir T., Walters G.B., et al. *A variant in CDKAL1 influences insulin response and risk of type 2 diabetes.* Nature genetics, 2007. **39**(6): p. 770-5.

202. Zhang X. and Yee D. *Tyrosine kinase signalling in breast cancer: insulin-like growth factors and their receptors in breast cancer.* Breast Cancer Res, 2000. **2**(3): p. 170-5.

203. Djavan B., Waldert M., Seitz C., and Marberger M. *Insulin-like growth factors and prostate cancer.* World J Urol, 2001. **19**(4): p. 225-33.

204. Yu H., Spitz M.R., Mistry J., Gu J., Hong W.K., and Wu X. *Plasma levels of insulin-like growth factor-I and lung cancer risk: a case-control analysis.* J Natl Cancer Inst, 1999. **91**(2): p. 151-6.

205. Holzenberger M., Dupont J., Ducos B., Leneuve P., Geloen A., Even P.C., et al. *IGF-1 receptor regulates lifespan and resistance to oxidative stress in mice.* Nature, 2003. **421**(6919): p. 182-7.

206. Scuteri A., Sanna S., Chen W.M., Uda M., Albai G., Strait J., et al. *Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity-related traits.* PLoS genetics, 2007. **3**(7): p. e115.

207. Morris A.P., Lindgren C.M., Zeggini E., Timpson N.J., Frayling T.M., Hattersley A.T., et al. *A powerful approach to sub-phenotype analysis in population-based genetic association studies.* Genetic epidemiology, 2010. **34**(4): p. 335-43.

208. Travers M.E. and McCarthy M.I. *Type 2 diabetes and obesity: genomics and the clinic.* Hum Genet, 2011. **130**(1): p. 41-58.

209. Xu W., Liu X., Sim X., Xu H., Khor C.C., Ong R.T., et al. *A statistical method for region-based meta-analysis of genome-wide association studies in genetically diverse populations.* Eur J Hum Genet, 2012.

210. Teo Y.Y., Ong R.T., Sim X., Tai E.S., and Chia K.S. *Identifying candidate causal variants via trans-population fine-mapping.* Genet Epidemiol, 2010. **34**(7): p. 653-64.

211. Genovese G., Friedman D.J., Ross M.D., Lecordier L., Uzureau P., Freedman B.I., et al. *Association of trypanolytic ApoL1 variants with kidney disease in African Americans.* Science, 2010. **329**(5993): p. 841-5.

212. Manolio T.A., Collins F.S., Cox N.J., Goldstein D.B., Hindorff L.A., Hunter D.J., et al. *Finding the missing heritability of complex diseases.* Nature, 2009. **461**(7265): p. 747-53.

213. Visscher P.M., Hill W.G., and Wray N.R. *Heritability in the genomics era--concepts and misconceptions.* Nature reviews. Genetics, 2008. **9**(4): p. 255-66.

214. Zuk O., Hechter E., Sunyaev S.R., and Lander E.S. *The mystery of missing heritability: Genetic interactions create phantom heritability.* Proceedings of the National Academy of Sciences of the United States of America, 2012. **109**(4): p. 1193-8.

215. Lupski J.R., de Oca-Luna R.M., Slaugenhaupt S., Pentao L., Guzzetta V., Trask B.J., et al. *DNA duplication associated with Charcot-Marie-Tooth disease type 1A.* Cell, 1991. **66**(2): p. 219-32.

216. *Rare chromosomal deletions and duplications increase risk of schizophrenia.* Nature, 2008. **455**(7210): p. 237-41.

217. Stefansson H., Rujescu D., Cichon S., Pietilainen O.P., Ingason A., Steinberg S., et al. *Large recurrent microdeletions associated with schizophrenia.* Nature, 2008. **455**(7210): p. 232-6.

218. Sebat J., Lakshmi B., Malhotra D., Troge J., Lese-Martin C., Walsh T., et al. *Strong association of de novo copy number mutations with autism.* Science, 2007. **316**(5823): p. 445-9.

219. Willer C.J., Speliotes E.K., Loos R.J., Li S., Lindgren C.M., Heid I.M., et al. *Six new loci associated with body mass index highlight a neuronal influence on body weight regulation.* Nat Genet, 2009. **41**(1): p. 25-34.

220. Walters R.G., Jacquemont S., Valsesia A., de Smith A.J., Martinet D., Andersson J., et al. *A new highly penetrant form of obesity due to deletions on chromosome 16p11.2.* Nature, 2010. **463**(7281): p. 671-5.

221. Craddock N., Hurles M.E., Cardin N., Pearson R.D., Plagnol V., Robson S., et al. *Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls.* Nature, 2010. **464**(7289): p. 713-20.

222. Xu H., Poh W.T., Sim X., Twee-Hee Ong R., Suo C., Tay W.T., et al. *SgD-CNV, a database for common and rare copy number variants in three Asian populations.* Hum Mutat, 2011.

223. Alkan C., Coe B.P., and Eichler E.E. *Genome structural variation discovery and genotyping.* Nat Rev Genet, 2011. **12**(5): p. 363-76.

224. Cohen J.C., Kiss R.S., Pertsemlidis A., Marcel Y.L., McPherson R., and Hobbs H.H. *Multiple rare alleles contribute to low plasma levels of HDL cholesterol.* Science, 2004. **305**(5685): p. 869-72.

225. Wang J., Cao H., Ban M.R., Kennedy B.A., Zhu S., Anand S., et al. *Resequencing genomic DNA of patients with severe hypertriglyceridemia (MIM 144650).* Arterioscler Thromb Vasc Biol, 2007. **27**(11): p. 2450-5.

226. Romeo S., Pennacchio L.A., Fu Y., Boerwinkle E., Tybjaerg-Hansen A., Hobbs H.H., et al. *Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL.* Nat Genet, 2007. **39**(4): p. 513-6.

227. Johansen C.T., Wang J., Lanktree M.B., Cao H., McIntyre A.D., Ban M.R., et al. *Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia.* Nat Genet, 2010. **42**(8): p. 684-7.

228. Cirulli E.T. and Goldstein D.B. *Uncovering the roles of rare variants in common disease through whole-genome sequencing.* Nat Rev Genet, 2010. **11**(6): p. 415-25.

229. Sanna S., Li B., Mulas A., Sidore C., Kang H.M., Jackson A.U., et al. *Fine mapping of five Loci associated with low-density lipoprotein cholesterol detects variants that double the explained heritability.* PLoS Genet, 2011. **7**(7): p. e1002198.

230. Dickson S.P., Wang K., Krantz I., Hakonarson H., and Goldstein D.B. *Rare variants create synthetic genome-wide associations.* PLoS Biol, 2010. **8**(1): p. e1000294.

231. Thomas D. *Gene--environment-wide association studies: emerging approaches.* Nat Rev Genet, 2010. **11**(4): p. 259-72.

232. Liu L., Li Y., and Tollefsbol T.O. *Gene-environment interactions and epigenetic basis of human diseases.* Curr Issues Mol Biol, 2008. **10**(1-2): p. 25-36.

233. Gallou-Kabani C. and Junien C. *Nutritional epigenomics of metabolic syndrome: new perspective against the epidemic.* Diabetes, 2005. **54**(7): p. 1899-906.

234. Smith F.M., Garfield A.S., and Ward A. *Regulation of growth and metabolism by imprinted genes.* Cytogenet Genome Res, 2006. **113**(1-4): p. 279-91.

235. Hunt K.A., Zhernakova A., Turner G., Heap G.A., Franke L., Bruinenberg M., et al. *Newly identified genetic risk variants for celiac disease related to the immune response.* Nat Genet, 2008. **40**(4): p. 395-402.

236. Soranzo N., Spector T.D., Mangino M., Kuhnel B., Rendon A., Teumer A., et al. *A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium.* Nat Genet, 2009. **41**(11): p. 1182-90.

237. Ganesh S.K., Zakai N.A., van Rooij F.J., Soranzo N., Smith A.V., Nalls M.A., et al. *Multiple loci influence erythrocyte phenotypes in the CHARGE Consortium.* Nat Genet, 2009. **41**(11): p. 1191-8.

238. Gudbjartsson D.F., Bjornsdottir U.S., Halapi E., Helgadottir A., Sulem P., Jonsdottir G.M., et al. *Sequence variants affecting eosinophil numbers associate with asthma and myocardial infarction.* Nat Genet, 2009. **41**(3): p. 342-7.

239. Ikram M.K., Sim X., Jensen R.A., Cotch M.F., Hewitt A.W., Ikram M.A., et al. *Four novel Loci (19q13, 6q24, 12q24, and 5q14) influence the microcirculation in vivo.* PLoS Genet, 2010. **6**(10): p. e1001184.

240. Fitau J., Boulday G., Coulon F., Quillard T., and Charreau B. *The adaptor molecule Lnk negatively regulates tumor necrosis factor-alpha-dependent VCAM-1 expression in endothelial cells through inhibition of the ERK1 and -2 pathways.* J Biol Chem, 2006. **281**(29): p. 20148-59.