1 Introduction

1.1 The burden of cardiovascular disease in modern society

Over the past few decades, improved sanitation and medical advances have led to a considerable decrease in mortality from infectious diseases. At the same time, chronic conditions such as cardiovascular disease (CVD) became the principal cause of mortality in the developed world (Kuller 1976). Although mortality from CVD has been decreasing, it is still the number one cause of mortality among chronic diseases. CVD refers to all the diseases of the heart and circulation system, including coronary heart disease (CHD), stroke, angina, heart attack, congenital heart disease. CHD and stroke are the two most common forms of CVD and both are mainly caused by atherosclerosis, a condition where arteries become narrowed by a gradual build-up of fatty material (i.e., atheroma) within artery walls. When the arteries become too narrow and there is inadequate oxygen-rich blood delivered to the heart, it causes angina, manifested by a pain or discomfort in the chest. When an atheroma or part of it in the arteries breaks away, it causes clotting in the circulation and cutting off the supply of oxygen-rich blood to heart muscle, leading to myocardial infarction (MI), commonly known as heart attack. When the blood clot blocks an artery that carries blood to the brain, it causes an ischaemic stroke. Another form of stroke is haemorrhagic stroke, caused by the rupture of a blood vessel in the brain.

Based on the World Health Organization’s report of global status on non-communicable diseases (year 2010), an estimated 17.3 million people died from CVD in 2008, representing 30% of all global deaths. It was projected that that this number would reach 23.3 million by 2030, making CVD remain to be the single leading cause of death over the next decade. For the two most common forms of CVD, CHD and stroke accounted for an estimated 7.3 million and 6.2 million of the total death respectively. Over 80% of CVD deaths take place in low- and middle-income countries. CVD is responsible for 10% of Disability-adjusted life years (DALYs) lost in low- and middle-income countries and 18% in high-income countries. DALYs is used more often to estimate the total burden of a disease, as opposed to simply count the number of resulting deaths.

1.2 Established and emerging risk factors for CVD

The term "risk factor" was first coined in Dr. Kannel's 1961 report of the association between circulating low-density lipoprotein cholesterol (LDL) and CVD (Kannel et al. 1961). Risk prediction is mainly used for disease prevention, defined as actions directed to avoid illness and promoting health to reduce the need for secondary and tertiary health care. Risk factors are important for assessing disease risk and therefore for disease prevention, while intermediate phenotypes usually reflect disease progression and are important markers for disease intervention and treatment. Risk factors were usually first identified through epidemiological studies. For example, the Framingham Heart Study (FHS) used a prospective design and identified age, male sex, smoking status, diabetes mellitus, hypertension, and serum cholesterol level as the most important risk factors for developing CVD (Dawber et al. 1959, Kannel et al. 1964). The INTERHEART study is based on a case-control design and reported a longer list of factors that account for most of the MI risk in 52 countries (Yusuf et al. 2004). There are more than 100 risk factors reported for association with CVD (Brotman et al. 2005). The criteria for being an established CVD risk factor include: a significant independent impact on the risk of CVD, a high prevalence in many populations, and a reduced level of CVD by the treatment and control of the risk factor. LDL is the first established risk factor for CVD. The decrease in mortality from CVD since 1980s was closely associated with lowering underlying risk factors especially LDL, which accounted for more than one-third of the observed decrease in mortality from CHD (Hunink et al. 1997).

Classical CVD risk factors include dyslipidemia (Kannel et al. 1961, Anderson et al. 1987), hypertension (Kannel et al. 1980), obesity (Lavie and Milani 2003), smoking (Service. 1983, Lavie and Milani 2003, Yusuf et al. 2004, Teo et al. 2006), alcohol drinking (Stampfer et al. 1988, Rimm et al. 1991), and physical inactivity (Pate et al. 1995). New risk factors include inflammatory markers especially C-reactive protein (CRP) (Koenig et al. 2004, Cushman et al. 2005), heamostasis markers such as figrinogen (Kannel et al. 1987), white blood cell count (WBC) (Kannel et al. 1992), homocysteine (Selhub et al. 1995), lipoprotein (a) (Bostom et al. 1996, Helfand et al. 2009), and uric acid (Kim et al. 2010) (**Figure 1.1**). CRP and WBC will be described in detail in later chapters. Risk factors initiated the atherosclerotic process and continued to be present throughout the cardiovascular disease continuum (CVDC). The concept of CVDC was originally described by Dzau and colleagues in 1991 (Dzau and Braunwald 1991), later on validated by clinical evidence of improved patient outcomes (Dzau et al. 2006). In CVDC, a chain of events are precipitated by several risk factors, which eventually cause end-stage heart failure and death if untreated (**Figure 1.2)**. Most CVD could be prevented by addressing modifiable risk factors such as smoking, unhealthy diet and physical inactivity, hypertension, and dyslipidemia.

Risk factors have been used to estimate the onset of both non-fatal and fatal cardiovascular events through the calculation of a risk score. Among them are the Framingham risk score (Wilson et al. 1998), the Joint British Societies risk charts (British Cardiac et al. 2005), the ASSIGN score (Tunstall-Pedoe et al. 2006), the Systematic COronary Risk Evaluation (SCORE) risk charts (Graham et al. 2007), and the Reynolds Risk Score (Ridker et al. 2007). There are differences among these scoring approaches. For example, the Framingham risk score is based on data from a single community, while the SCORE risk charts were based on data from 12 European countries. These epidemiologic risk profiling did not address the fact that risks can differ between regions and countries due to different life styles, life expectancy and genetic predisposition. Therefore, these risk prediction algorithms need to evolve over time. An updated Framingham risk score in 2008 predicted risk for more CVD outcomes including cerebrovascular events, peripheral artery disease and heart failure (D'Agostino et al. 2008), compared to the one first developed in 1998. Type-2 diabetes (T2D) was dropped from the updated Framingham risk score because it was considered to be a disease outcome itself, with similar risk factors as that for CVD. These risk scores are used to determine who should be offered preventive drugs such as those lowering blood pressure or cholesterol levels. Individuals with <10%, 10-20%, and >20% CVD risks are considered low, intermediate, and high risk respectively.

The term “biomarker”, as used in the title of this thesis, focuses more on the biologically measurable risk factors. It is meant to distinguish from lifestyle related risk factors such as smoking, drinking, and nutrition. The term biomarker was established as a medical subject heading term in 1989, meaning “measurable and quantifiable biological parameter (e.g. specific enzyme concentration, specific hormone concentration, specific gene phenotype distribution in a population, presence of biological substance) which serves as index for health- and physiology-related assessments, such as disease risk, psychiatric disorders, environmental exposure and its effects, disease diagnosis, metabolic processes, substance abuse, pregnancy, cell line development, epidemiologic studies, etc.” In 2001, an updated definition of biomarker is given by the US National Institutes of Health 2001, as “a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacological responses to a therapeutic intervention" (Biomarkers Definitions Working 2001). This definition made the term biomarker more inclusive. In this thesis, the studied cardiovascular biomarkers are all biological molecules existing in circulatory system.

Figure 1.1 Established and new/emerging risk factors for CVD

This figure is adopted from (Badimon and Vilahur 2012) as is.

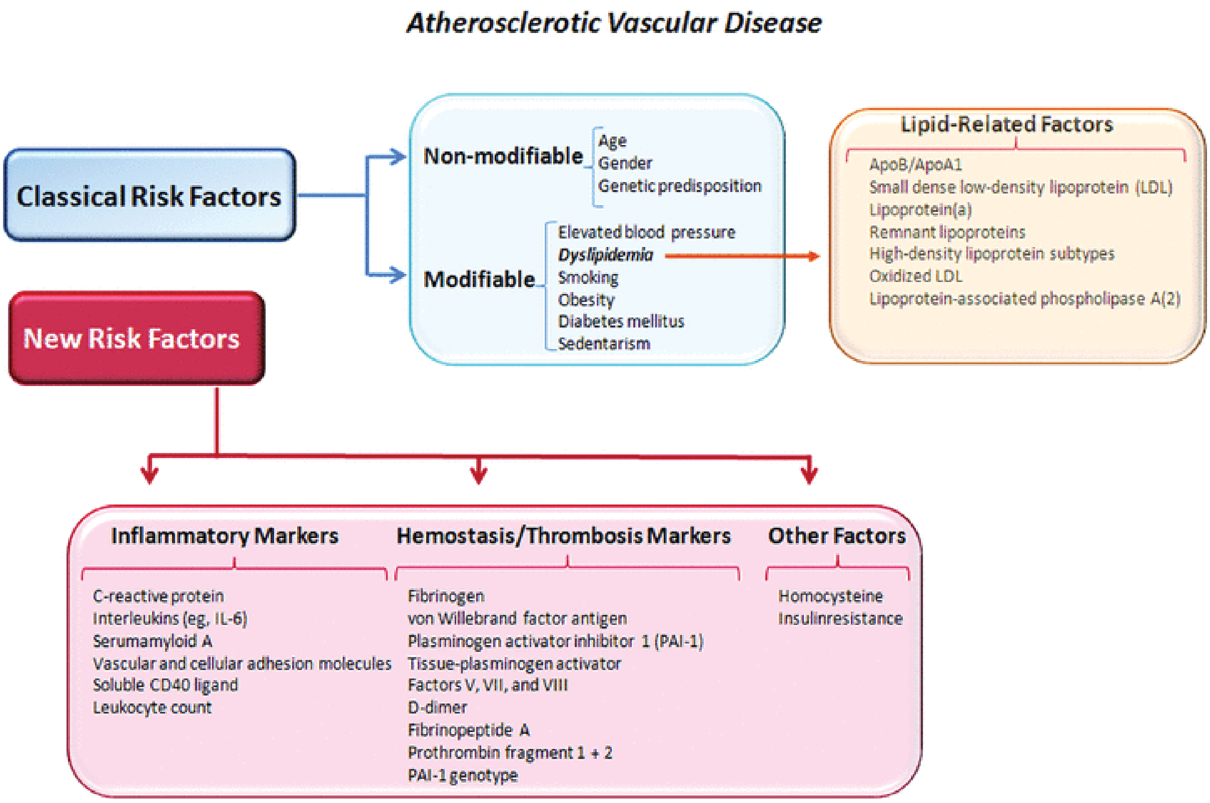
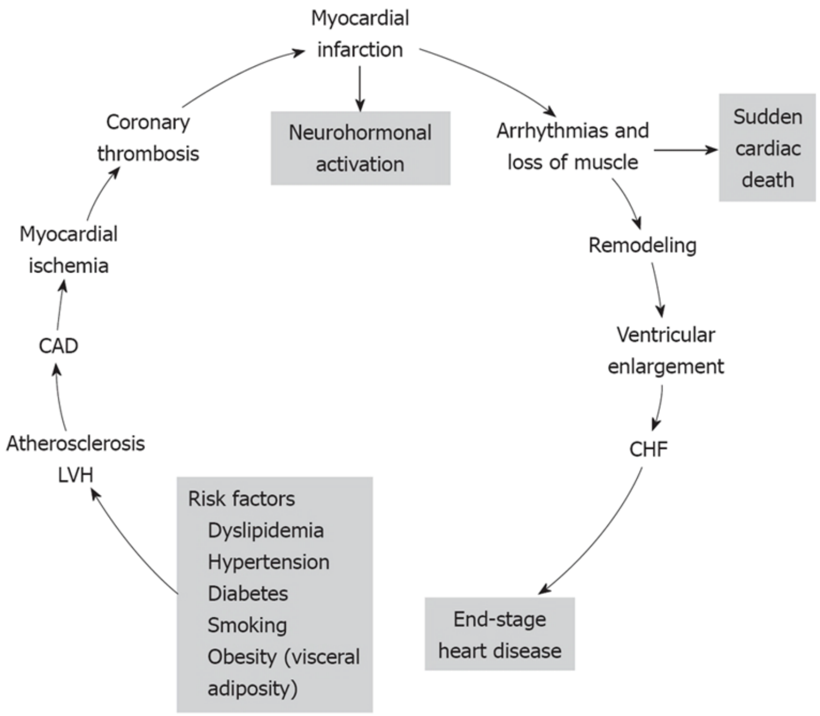


Figure 1.2 The cardiovascular disease continuum

This figure was adapted from Dzau et.al as it (Dzau and Braunwald 1991). LVH indicates left ventricular hypertrophy. CHF indicates congestive heart failure. The major risk factors leading to CVDC are listed at the bottom. All these risk factors, with the exception of smoking, constitute the metabolic syndrome.



1.3 The allelic architecture of complex traits

Population genetics is the study of the distributions and changes of allele frequency in a population, while the population is subject to evolutionary processes. Study areas of population genetics include recombinations, Mendelian inheritance, genetic linkage and linkage disequilibrium (LD), population stratification, etc. Allelic architecture refers to the number and frequencies of susceptibility alleles underlying complex diseases. Diseases with high prevalence in the general population such as T2D and CHD are polygenic, i.e., determined by multiple genetic variants, together with lifestyle and environmental factors. This is also the case for complex, quantitative risk factors. Although there is distinct difference of allelic architecture between high prevalent complex diseases and low prevalent Mendelian diseases, these two are not completely disconnected. Recently, a study linked complex diseases to unique collections of Mendelian loci by showing that common variants associated with complex diseases are enriched in the genes with Mendelian patterns of inheritance (Blair et al. 2013).

Genetic research on complex traits began with surveying candidate variants or regions of the genome, followed by analysis analyses that scan the whole genome with limited resolution, and then genome-wide association studies (GWAS) over the past ~10 years. Due to the nature of "hypothesis driven", candidate gene studies used a very liberal *P* value (such as *P*<0.05) threshold to claim significance, which could lead to a high level of reported false positives (Masicampo and Lalande 2012). Actually, less than 5% of associations identified in candidate gene studies were replicated in larger GWAS (Ioannidis et al., 2011). Linkage analysis is suitable for detecting rare and highly penetrant variants causative for rare diseases with classical Mendelian patterns of inheritance. Early success example of linkage studies included the identification of causal mutations for cystic fibrosis (Kerem et al. 1989) and Huntington disease (MacDonald et al. 1992). In general, linkage analysis is not suitable for detecting common alleles of unusually large effects for complex diseases, but there are a few exceptions, including the successful discoveries of the *INS* locus in T1D (Bell et al. 1984) and the *ApoE* locus in early onset Alzheimer's disease (St George-Hyslop et al. 1987, Goate et al. 1991). The LOD score (logarithm (base 10) of odds) is a statistical test often used for linkage analysis (Morton 1955). It compares the likelihood of obtaining the test data if the two loci are indeed linked, to the likelihood of observing the same data purely by chance. A LOD score of 3.3 or higher has been shown to correspond to a statistical significance level of

0.05. There are two main algorithms used to calculate LOD score: the Elston–Stewart algorithm (Elston and Stewart 1971), and the Lander– Green algorithm (Lander and Green 1987). The major difference is whether the recursion took place over individuals in a pedigree (computing increases linearly with pedigree size but exponentially with the number of loci) or over loci (computing increases linearly with the number of loci but exponentially with pedigree size). The Elston–Stewart algorithm is applicable to very large pedigrees while the Lander–Green algorithm can accommodate thousands of markers on a chromosome.

Before GWAS approach was widely used, there were two theories for explaining genetic underpinning of complex diseases with high prevalence: common disease common variant (CDCV) and common disease rare variant (CDRV). The CDCV theory hypothesised that a small number of common variants could explain a large proportion of phenotypic variation for common traits (Lander 1996, Reich and Lander 2001, Pritchard and Cox 2002, Botstein and Risch 2003). This CDCV theory has been well supported by GWAS where many common variants are identified for association with common diseases and complex traits (Hindorff et al. 2009). However, common variants did not explain common variation fully (Manolio et al. 2009), and this led to a slightly modified version of CDCV - the infinitesimal model. The infinitesimal model highlighted the role of a much larger number of common variants with much smaller effects. This model was also supported by GWAS especially large scale meta-analysis with adequate power for both diseases traits (International Schizophrenia et al. 2009) and quantitative traits (Yang et al. 2011). In contrast to CDCV and infinitesimal model, the CDRV theory hypothesized that a large number of rare variants with large effects could explain a large proportion of heritability (Cirulli and Goldstein 2010). It is worth noting that very rare variants would not be common enough to explain large variance or reach genome-wide significance even if they are causal and have large effects in a small proportion of studied samples. Statistical simulations have shown that CDCV and CDRV are not necessarily mutually exclusive, with both rare and common variants underlying a polygenic genetic architecture for complex traits (Hemani et al. 2013). Other models such as the broad sense heritability model (Eichler et al. 2010) looked beyond genetic variants by considering the combined effects of genotypic, environmental and epigenetic interactions.

1.4 Genome-wide association studies (GWAS)

The completion of the human genome project (Lander et al. 2001, Venter et al. 2001) and the rapid improvement of technologies for ascertaining and analysing the human genome set the stage for GWAS, which has changed the landscape of genetic study on complex diseases. In 2005, only a few dozen loci were reported for association with a handful of complex diseases. By the end of 2011, the NHGRI GWAS catalogue has reported over 2,000 association signals for over 200 complex traits. Actually, the idea of GWAS was not new, proposed as early as in 1996, when association testing was found to have greater power than linkage analysis especially for detecting variants with modest effect sizes (Risch and Merikangas 1996). Risch and colleagues suggested that creating high-density genome-wide polymorphism maps would allow well-powered association testing across all genes. Although the concept and analytic methods for GWAS were ready at that time, it was only implemented around 2005 when genome-wide SNP array were commercialized and were affordable for research projects with large sample size (Syvanen 2005). The genetic polymorphism selection by major vendors was mainly based on data generated from the International HapMap project (International HapMap et al. 2007, International HapMap et al. 2010). For the two biggest vendors, Affymetrix used a strategy of randomly selected SNPs while Illumina used tagging methods that maximize coverage in European populations (Barrett and Cardon 2006). The early versions of SNP arrays usually include less than 1 million common variants, which could be imputed to up to 3 million variants discovered from the HapMap project. When a common set of haplotype variants are analysed by most individual cohorts, results could be cross-examined and meta-analysed in large collaborative consortia.

Compared to candidate gene studies and linkage analysis, GWAS scan the whole genome in a systematic manner for detecting genetic variants susceptible to diseases and quantitative traits (Hirschhorn and Daly 2005). Since GWAS became available, large advances have been made. One of the early successes of GWAS was the identification of the *Complement Factor H* gene as a major risk factor for age-related macular degeneration (AMD) (Haines et al. 2005, Klein et al. 2005), in studies of relatively small sample size (~100 cases) and employing a sparse SNP array (~110K). These studies not only identified strongly associated genetic variants, but also proved that common variants included in genome-wide SNP array could tag underlying causal variants, a key assumption for GWAS.

Follow-up resequencing studies revealed a functional polymorphism that is in high linkage disequilibrium (LD) with the discovered GWAS signal. However, the AMD genetic variants identified in these two studies are rare examples where common variants (MAF >5%) have large effects (OR > 4). In general, the identification of genetic variants linked to complex traits would require many more samples and variants to tag the whole genome and survive the large number of multiple testing. In 2007, a landmark GWAS study with ~17,000 subjects typed on half a million variant SNP array (Wellcome Trust Case Control Consortium 2007) identified 24 independent association signals for seven common diseases. This first WTCCC study was the largest set of GWAS of its time, costing a total of $9 million. It identified 21 loci, of which 14 were novel. All these associations has been confirmed in later meta-analyses. Later on, many other studies conducted extensive replication for suggestive signals coming from this WTCCC study and identified many more novel loci, for type 1 diabetes (Todd et al. 2007), type 2 diabetes (Zeggini et al. 2007), rheumatoid arthritis (Thomson et al. 2007, Barton et al. 2008), and Crohn's disease (Parkes et al. 2007). This in a way established the importance of performing independent replication for modern GWAS. This study also provided a first strong indication of differences in allelic architecture for different traits, with many more associations detected for autoimmune diseases as opposed to hypertension or CAD. Besides novel findings, a number of novel techniques and protocols used in this study became standards in GWAS since then, for example, systematic assessing and adjusting for population stratification, and using the HapMap reference panel for genotype imputation. This study also characterised other types of genomic variations including copy number variants (CNV) and large insertions and deletions. The second landmark genomic study from the WTCCC concluded that most common CNVs are well tagged by common SNPs and are unlikely to discover novel findings for common human diseases (Wellcome Trust Case Control et al. 2010). However, rare CNV and large deletions have been reported for association with other categories of complex diseases including autism and schizophrenia (International Schizophrenia 2008, Glessner et al. 2009).

The subsequent widespread implementation of imputation analysis based on common reference maps (HapMap2 mainly) has been instrumental in the completion of powered meta-analyses of GWAS studies, allowing reaching sample sizes necessary for robust genetic discoveries. As of September 2014, more than 2,000 robust associations with complex traits have been reported (Hindorff et al. 2009), which revealed important biological pathways and defined novel therapeutic hypotheses (Visscher et al. 2012). For example, GWAS on T2D have played an important role in shifting research focus away from insulin resistance towards insulin production (McCarthy and Zeggini 2009) and led to the identification of many new drug targets (Wolfs et al. 2009). Another example is the discovery of *BCL11A* as a major modifier of disease severity in haemoglobinopathies (Akinsheye et al. 2011), which led to the development of new treatment options for sickle cell disease and beta-thalassemia (Bauer and Orkin 2011).

1.5 GWAS studies of CVD events and cardiovascular biomarkers

The heritability for CHD and stroke was established to be 50% (Fischer et al. 2005) and 32% (Bak et al. 2002) respectively. Although the prevalence of the metabolic syndrome has greatly increased in the past decades due to lifestyle changes, a large portion of the phenotypic variation in cardio-metabolic traits between individuals is still due to genetic variation (van Dongen et al. 2013). GWAS have been widely used to study both end points and intermediate phenotypes of CVD. As mentioned above, the first WTCCC study studied CAD and hypertension together with five other diseases. It reported one locus for coronary CAD but none for hypertension (Wellcome Trust Case Control Consortium 2007). Over the past few years, collaborative efforts have made it possible to conduct large meta-analysis of GWAS with the sample size up to tens of times of the original WTCCC study. Two published large meta-analysis on CAD reported a total of 46 genetic loci for association with CAD (Schunkert et al. 2011, Consortium et al. 2013). The 2013 study reported that 12 and 5 of these 46 CAD loci show significant associations with lipids and BP respectively. It further reported that the four most significant pathways mapping to networks comprising 85% of these putative genes are linked to lipid metabolism and inflammation, underscoring the causal role of lipids and inflammation in the genetic aetiology of CAD. The latest efforts on CAD GWAS used a similar sample size as that in the 2013 study (60,801 cases and 123,504 controls vs. 63,746 CAD cases and 130,681 controls), but used the 1000GP data as imputation reference panel so that it interrogated 6.7 million common (MAF>0.05) and 2.7 million low frequency (0.005<MAF<0.05) (CARDIoGRAMplusC4D Consortium 2015). In addition to confirming most known CAD loci, this study identified 10 novel loci, eight additive and two recessive. However, this study suggested a lack of evidence of low frequency variants with larger effects and no evidence of synthetic association and suggested that the genetic susceptibility of CAD is largely determined by common SNPs of small effect size.

It was proven challenging that the CAD loci discovered from GWAS could add improvement for risk prediction (Buijsse et al. 2011, Companioni et al. 2011) as compared to other phenotypes such as AMD (Seddon et al. 2009). In general, using genetic loci for risk prediction has unique advantages because genetics do not change over an individual's lifetime and are not affected by other risk factors. Therefore, risk prediction can be carried out much further in advance. In the past 15 years, interest has grown on predicting CVD risk at longer-term (for example, 30-year or lifetime). Genetic information shall benefit such efforts to improve communication of risk, and motivate risk-factor modification especially in young patients (Wong 2014). Also, Mendelian Randomization (MR) studies using genetic variants as instrumental variables could resolve epidemiological problems of establishing causality, which established the causal role for LDL to CVD (Linsel-Nitschke et al. 2008), but not for high-density lipoprotein cholesterol (HDL) (Voight et al. 2012). This approach could also be used to perform retrospective drug trials, for example, the establishment of IL6R as a drug target for CVD (Interleukin-6 Receptor Mendelian Randomisation Analysis et al. 2012).

As stated above, CVD risk factors are critical for the initiation and progression of CVD events. From the point view of genetic research, quantitatively measured risk factors are also preferred to dichotomous CVD events due to increased power and an often more interpretable outcome. For example, assays for LDL levels are precise and standardized around the world, but the diagnosis and clinical criteria for CHD might differ significantly. The beta statistics of a particular variant indicates a unit change in LDL level per allele, but such a statistic for disease outcome would be less intuitive for interpretation. Once genetic variants for quantitative variants are discovered, they could provide clinical insights to the associated diseases (Teslovich et al. 2010). Compared to the disease end points, meta-analyses for quantitative traits have identified many more loci and explained much larger proportion of phenotypic variance. A GWAS meta-analysis for plasma lipids identified 95 loci that explain ~12% of phenotypic variance for high density lipoprotein (HDL), LDL, and total cholesterol (TC). The large sample size is proving powerful for identifying genetic variants with small effect size. Compared to the first WTCCC study that included ~2,000 cases and ~3,000 controls for studying hypertension and discovered no associated locus, the largest GWAS on BP included more than 200,000 samples identified a total of 29 loci (16 novel) for association with BP. A genetic risk score based on these 29 variants are associated with hypertension, left ventricular wall thickness, stroke and CAD (Ehret et al. 2011). This effectively demonstrates the value of using quantitative risk factors for genetic study of CVD events.

1.6 Rare variants and the motivation for whole genome sequencing (WGS)

Common variants identified by GWAS have proven highly informative to identify novel biological processes underlying common disease (Hindorff et al. 2009). But GWAS is only well powered to detect associations that are well covered by common tag SNPs. Populations with different LD to the HapMap populations, or meta-analyses across populations with different patterns of LD, can confound the tag SNP approach (Teo et al. 2010). Also, low frequency variants are not well tagged by common SNPs (International HapMap et al. 2010). So far, common variants discovered from first generation GWAS explained only a small proportion of phenotypic variance for most common traits and there is a lack of proven added predictive value in clinical usage by including GWAS signals on top of risk factors already known. The missing heritability theory (Manolio et al. 2009) hypothesized that GWAS might have missed variants that have large effects but too low frequency to be detected by SNP array. This is also supported by the evolution theory that alleles susceptible to diseases and their risks are likely to be deleterious and could not reach high frequency due to purifying selection (Pritchard 2001, Goldstein et al. 2013). Although it is debatable on whether, and how much, synthetic associations from variants could explain common variants effects, it was already shown that rare copy number variants contribute to several complex neurodevelopmental disorders (International Schizophrenia 2008, Glessner et al. 2009). The variants with low to rare frequency (shown in light blue in **Figure 1.3**) could be where a large proportion of missing heritability resides. This is a key underlying reasoning for the new generation of population genetic studies where sequencing technologies are used for discovering low frequency (defined here as MAF between 1-5%) and rare variants (defined here as MAF <1%). Sequencing could identify low frequency and rare SNPs, various types of structural variations, as well as more common variants (~ 10-15%) that are not well tagged by SNP arrays (Flannick et al. 2012). Sequencing studies could also potentially discover causal functional variants that could not be well interrogated on SNP array or imputation (Cirulli and Goldstein 2010).

The desire to study low frequency and rare variants in a genome-wide fashion was met by fast development in sequencing technologies. In 2004, the 454 pyrosequencing method pioneered the field by allowing hundreds of thousands of sequencing reactions to be carried out in parallel (Langaee and Ronaghi 2005). In 2006, the Solexa reversible termination sequencing method was commercialized by Illumina. In 2007, the Oligonucleotide Ligation and Detection (SOLiD) technology was introduced by ABI (now Life Tech). By 2007, it was possible to sequence over 500Mb a day on a single machine (Mardis 2008), and that was when the 1000 Genomes Project (1000GP) was founded to perform low-coverage (2-4X) sequencing on up to 2,500 human genomes. Since 2008, more sequencing technologies are developed, including Ion torrent, pacific biosciences, Illumina's MiSeq (Quail et al. 2012). In January 2010, Illumina unveiled the HiSeq 2000 sequencing system. It initially generated two billion paired-end reads and 200Gb of quality filtered data in a single run, which allows researchers to obtain 30-fold coverage of two human genomes in a single run. This is the sequencing technology adopted by the UK10K project, which is funded by the Wellcome Trust in March 2010.

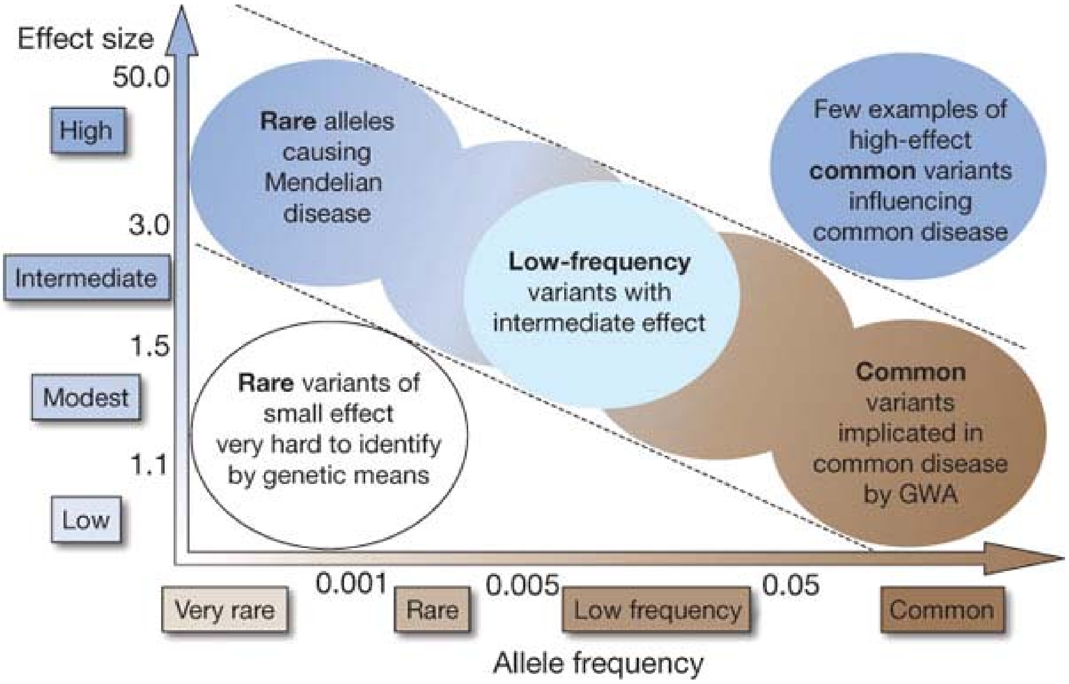
While WGS is still prohibitively expensive for large population based studies, the development of sequence capture technology enabled sequencing of the whole exome (Albert et al. 2007), which covers ~1.5% of the human genome (Lander et al. 2001). Compared to WGS, whole exome sequencing (WES) studies have been conducted at an even greater scale over the past several years, due to cost efficiency as well as data analysis efficiency where genomic boundaries and annotations could be defined straightforward and therefore the results are easier to be interpreted. WES became the dominant method for discovering causal variants for Mendelian diseases (Bamshad et al. 2011), while WGS should discover a lot more biologically relevant variants for common complex traits. This is consistent with findings from the ENCODE project that most variants that control protein biochemistry are non-coding and are not within exons (Pennisi 2012). Currently, most WGS technology sequence the whole genome in low depth, sometimes complemented by high-depth sequencing of the whole exome (Abecasis et al. 2012).

Finally, the increased availability of whole-genome and whole-exome sequencing data is bringing linkage analysis once again to the forefront of genetic research, owing to the development of powerful methods to detect rare variants and the use of family-based data.”

In association studies, population stratification can lead to an increased number of false-positive results if not properly accounted for. However, this is not a problem in linkage analysis because the family structure instead of the population genotype frequencies dictates a proband’s genotypes. Given that large and complete pedigree is usually hard to get for genetic studies, it is preferable to combine positive aspects of linkage and association analysis by using family-based rather than population-based control individuals. Although the transmission disequilibrium test (TDT) tests have already used such family-based controls .it is only powerful when there is both linkage and association. The TDT test was recently extended (the rare variant-TDT (RV-TDT)) to WGS data, with several rare variant association tests methods implemented (He et al. 2014). Linkage analysis not only effectively adjusts for population stratifications, but also provides statistical evidence for disease aetiology. Over the past couple of years, linkage analysis coupled with WGS have identified many new disease susceptibility genes, with a sample size that is much smaller that would be needed for a population based genome-wide scan. In the future, linkage analysis of WGS data is expected to be even more widely used (Yan et al. 2013, Santos-Cortez et al. 2014).

Figure 1.3 The allelic spectrum of human disease predisposition

This figure is copied as is from Maniolio et al. 2009 (Manolio et al. 2009). It illustrates the relationship between frequency and effect size for genetic variants contributing to human disease, from common to rare. The focus of WGS based studies aim to low-frequency to rare alleles with modest effect sizes, as shown by the light blue circle in the figure.


Effect size
50.0
High
3.0
Intermediate
1.5
Modest
1.1
Low
Rare alleles causing Mendelian disease
Few examples of high-effect common variants influencing common disease
Low-frequency variants with intermediate effect
Rare variants of small effect very hard to identify by genetic means
Common variants implicated in common disease by GWA
0.001
0.005
0.05
Very rare
Rare
Low frequency
Common
Allele frequency

1.7 The UK10K Project

In 2010, the Wellcome Trust found the largest WGS study at the time - the UK10K project, with a £10.5 million funding support. The UK10K project aims to better understand the link between low frequency and rare genetic variants and their impact on health and diseases (The UK10K Consortium 2015). The full UK10K project conducted sequencing for ~10,000 samples: the cohort arm (referred as UK10K-Cohorts) conducted WGS for ~4,000 population based samples; the disease arm conducted high-depth WES for ~6,000 affected individuals. For the ~4,000 samples included in the cohort arm, ~2,000 each are from two well established population studies in UK: TwinsUK (Spector and Williams 2006) and The Avon Longitudinal Study of Parents and Children (ALSPAC) (Golding et al. 2001). TwinsUK is a general population throughout UK (Moayyeri et al. 2012) while ALSPAC is a population-based birth cohort study that recruited more than 13,000 pregnant women resident in Bristol (formerly Avon) UK. For both cohorts, study participants were selected to maximise phenotypic coverage, previous genome-wide array genotyping, coverage with other "-omic" datasets (transcriptomic, metabolomic) and consent to WGS, but were otherwise representative of the original population samples.

Using low-depth WGS in UK10K-Cohorts is a cost-effective approach when high-depth WGS is still prohibitively expensive for thousands of samples. For example, it was shown that sequencing 3,000 individuals at low-depth (4X) provides similar power to sequencing of >2,000 individuals at high depth (30X) for disease-associated variants with frequency >0.2%, but the low-depth approach only requires ∼20% of the sequencing resources (Li et al. 2011). An average sequencing depth of 7X in the UK10K-Cohorts project enables the identification of almost all accessible SNPs, Insertion/Deletion polymorphism (InDel) and other structural variants down to MAF of 0.1% (Le and Durbin 2011). This is one magnitude higher resolution compared to the 1000 genome project (1000GP) that fully characterize variants down to MAF of 1% (Abecasis et al. 2012). The low-depth sequencing was proven sensitive for detecting rare variants, which detected more than 70% of singletons and more than 90% of doubletons that are discovered in the UK10K high-depth (80X) WES arm. The UK10K WGS approach also discovered a lot of rare variants that could be potentially characteristic of the UK population. Roughly, only 10% of singletons discovered in UK10K WGS were previously discovered by 1000GP (The UK10K Consortium 2015).

Besides the ~4,000 samples directly sequenced, the two cohorts in UK10K cohort arm (TwinsUK and ALSPAC) have an additional ~10,000 samples with genome-wide SNP array data, which could be imputed into the full set of variants discovered from WGS. All variants with MAF down to 0.1% should be imputable, where minor alleles occur more than five times in the study sample and the definition of a shared haplotype between study sample and reference sample is possible. A total of 64 biomedically relevant traits (60 quantitative traits and four binary traits) were measured in these two cohorts and were analysed in UK10K-Cohorts, 31 of which exist in both cohorts and are their initial association results were presented in the UK10K flagship paper (The UK10K Consortium 2015). The sample size for each of the 64 traits is listed in **Table 1.1**. My PhD thesis concentrate on a total of 13 CVD related biomarkers, including four lipid traits (HDL, LDL, TC, TG), one inflammatory biomarker (CRP), and eight haematological traits (Hemoglobin (HGB), Mean corpuscular hemoglobin (MCH), Mean corpuscular hemoglobin concentration (MCHC), Mean corpuscular volume (MCV), Packed cell volume (PCV), Platelet counts (PLT), Red blood cell counts (RBC), White blood cell counts (WBC)).

The large number of traits measured on the same individuals in the UK10K-Cohorts provided a good opportunity to learn about the general allelic architecture especially rare variants architecture of those traits. Since single marker association tests are typically underpowered for rare variants (MAF <1%), the UK10K-Cohort projects adopted an integrative framework of variance component method and burden tests implemented in sequence kernel association test (SKAT) and SKAT optimized (SKAT-O) (Wu et al. 2011, Liu and Leal 2012) . The details of these association tests will be described in Chapter 2.

Table 1.1 List of traits in UK10K-Cohorts

The 64 traits were grouped into categories based on biomedical relevance. WGS means those samples sequenced, GWA means those samples with SNP-array data, imputed to the WGS reference panel.

Category	Name	TwinsUK WGS	ALSPAC WGS	Total WGS	TwinsUK GWA	ALSPAC GWA	Total GWAS	Total
Obesity /anthropometry	**BMI**	1747	1791	3538	2330	4101	6431	9969
	Height	1747	1794	3541	2331	4103	6434	9975
	Weight	1747	1812	3559	2330	4132	6462	10021
	Hip circumference	1266	1808	3074	1623	4115	5738	8812
	Waist circumference	1265	1807	3072	1624	4121	5745	8817
	Waist hip ratio	1265	1806	3071	1620	4116	5736	8807
	Total fat mass	1716	1683	3399	2095	3815	5910	9309
	Total lean mass	1716	1683	3399	2095	3815	5910	9309
	Trunk fat mass	1514	1683	3197	547	3815	4362	7559
	Forearm length	-	1760	1760	-	4367	4367	6127
	Head circumference	-	1762	1762	-	4388	4388	6150
	Leg length	-	1764	1764	-	4386	4386	6150
	Sitting height	-	1764	1764	-	4387	4387	6151
	Upperarm length	-	1762	1762	-	4369	4369	6131
	Adiponectin	864	1461	2325	737	2772	3509	5834
	Leptin	958	1459	2417	663	2765	3428	5845
Diabetes Biochemistry	**Glucose**	1701	1224	2925	2202	1701	3903	6828
	HOMA-B	1669	1219	2888	1671	1697	3368	6256
	HOMA-IR	1577	1219	2796	1659	1695	3354	6150
	Insulin	1676	1220	2896	1927	1693	3620	6516
Heart function	**Heart rate (ECG+pulse)**	1385	1590	2975	939	2932	3871	6846
CVD hypertension	**DBP**	1536	1773	3309	1457	4046	5503	8812
	SBP	1536	1773	3309	1457	4046	5503	8812
CVD Biochemistry	**HDL**	1713	1497	3210	1896	2820	4716	7926
	LDL	1696	1495	3191	1870	2815	4685	7876
	TC	1711	1495	3206	1895	2817	4712	7918
	TG	1705	1497	3202	1882	2820	4702	7904
	VLDL	1700	1497	3197	1874	2820	4694	7891
	Apolipoprotein A1	1449	1465	2914	995	2772	3767	6681
	Apolipoprotein B	1443	1468	2911	989	2765	3754	6665
	Homocysteine	1279	93	1372	799	184	983	2355
	CRP	879	1167	2046	1017	2226	3243	5289
Blood Biochemistry	**HGB**	1553	1524	3077	1056	2882	3938	7015
	MCH	1549	-	1549	1061	-	1061	2610
	MCHC	942	-	942	947	-	947	1889
	MCV	1548	-	1548	1058	-	1058	2606
	PCV	1555	-	1555	1062	-	1062	2617
	PLT	1553	-	1553	1070	-	1070	2623
	RBC	1561	-	1561	1062	-	1062	2623
	WBC	1551	-	1551	1065	-	1065	2616
	Interleukin 6	-	1480	1480	-	2779	2779	4259
Liver Function	**Albumin**	1713	-	1713	1700	-	1700	3413
	Alkaline phosphatase	1702	-	1702	1636	-	1636	3338
	Bilirubin	1702	-	1702	1637	-	1637	3339
	Gamma glutamyl transpeptidase	1699	-	1699	1594	-	1594	3293

Table 1.1 List of traits in UK10K-Cohorts (*continued*)

Category	Name	TwinsUK WGS	ALSPAC WGS	Total WGS	TwinsUK GWA	ALSPAC GWA	Total GWAS	Total
Renal Function	**Bicarbonate**	1714	-	1714	1676	-	1676	3390
	Creatinine	1707	-	1707	1629	-	1629	3336
	Phosphate	1392	-	1392	1691	-	1691	3083
	Sodium	1683	-	1683	1677	-	1677	3360
	Urea	1697	-	1697	1617	-	1617	3314
	Uric acid	1305	-	1305	1588	-	1588	2893
Lung Function	**FEV/FVC ratio**	1676	1604	3280	1892	3521	5413	8693
	Forced Expiratory Capacity	1679	1606	3285	1896	3522	5418	8703
	Forced Expiratory Volume	1681	1606	3287	1896	3522	5418	8705
Birth	**Birth weight**	-	1691	1691	-	5327	5327	7018
	Birth length	-	1137	1137	-	3470	3470	4607
	Gestational age	-	1712	1712	-	5390	5390	7102
	Ponderal index	-	1122	1122	-	3421	3421	4543
	Placental weight	-	703	703	-	2166	2166	2869
Dynamic	**Grip strength**	1514	1682	3196	901	3465	4366	7562
	Ever broken bone*	-	1756	1756	-	3657	3657	5413
	Eye preference*	-	1671	1671	-	4158	4158	5829
	Handedness tasks*	-	1700	1700	-	3972	3972	5672
	Handedness drawing*	-	1676	1676	-	3875	3875	5551

* binary traits

1.8 This thesis

In this chapter, I have reviewed the research on complex disease genetics in general, and the genetics of cardiovascular biomarkers in particular. I also laid out the motivation for WGS based studies and gave a description of the UK10K project. My main hypothesis is that applying WGS to deeply phenotyped population samples is capable of discovering rare but highly penetrant genetic variants. The main research aim is to utilize large-scale WGS data and WGS imputed data to identify novel genetic variants that contribute to CVD related traits. As it is still not clear whether some of the selected biomarkers are direct mediators of the disease or merely markers of disease manifestation, I hope to identify highly penetrant genetic determinants of these biomarkers that can, in the future, be used to assess genetic risk and causal effects. I have contributed to the whole UK10K-Cohorts study and will elaborate on some general lessons learned from this study in the general discussion section. In the following chapters, I describe methods and results for WGS based imputation (chapter 3) and the deep analysis of 13 CVD biomarkers (chapters 4-6). Specifically, I seek to evaluate the following three broad aspects: 1. what are the characteristics of phasing and imputation with WGS data? 2. what novel analytic methods could be applied to a large scale WGS based association study on a rich of phenotypes? 3. can I identify novel and potentially stronger effect genetic variants that are associated with the chosen CVD traits?