Published in final edited form as:

Trends Cardiovasc Med. 2009 April ; 19(3): 69–75. doi:10.1016/j.tcm.2009.05.001.

Utilizing New Tools to Define the Genetic Underpinnings of Risky Traits Associated with Coronary Artery Disease: The SardiNIA Study

James B Strait^{1,2}, Manuela Uda³, Edward G. Lakatta¹, and Samer S. Najjar^{1,*}

¹ Laboratory of Cardiovascular Science, Intramural Research Program, National Institute on Aging, National Institutes of Health, Baltimore, MD

² Laboratory of Genetics, Intramural Research Program, National Institute on Aging, National Institutes of Health, Baltimore, MD

³ Istituto di Neurogenetica e Neurofarmacologia (INN), Consiglio Nazionale delle Ricerche, c/o Cittadella Universitaria di Monseratto, Monserrato, Cagliari, Italy

Abstract

Genome-wide association studies are increasingly being applied to search for novel genes that might underlie cardiovascular diseases. In this article, we briefly review the principles that underlie modern genetic analyses, and provide several illustrations from the SardiNIA Study of genome-wide association studies for cardiovascular risk factor traits.

“Variability is the law of life. As no two faces are the same, so no two bodies are alike, and no two individuals behave alike in the abnormal conditions we know as disease. This is fundamental to the education of the physician”

-Sir William Osler

Introduction

Cardiovascular diseases remain the leading causes of mortality in industrialized countries (Rosamond et al. 2008). Spectacular advances in our understanding of the pathophysiology of these diseases have provided the cornerstone for developing and implementing numerous pivotal preventive and therapeutic strategies, which are credited with remarkable improvements in prognosis and quality of life. However, it appears that the mortality rates from these diseases are no longer declining or, alarmingly, may even be on the rise (Hardoon et al. 2008). Rapid and well publicized advances in the field of genetics (reviewed in (Lusis 2003, Nabel 2003, Damani and Topol 2007, Kullo and Ding 2007)) raise the hope that imminent breakthroughs could be achieved with newfound genetic tools (see glossary of terms). In this article, we briefly review the principles that underlie modern genetic analyses, and provide several examples from our own experience with genome-wide association studies (GWAS) of cardiovascular risk factors in the SardiNIA Study.

*Corresponding Author: Samer S. Najjar, MD, NIA/ASTRA Unit, 5th floor, Harbor Hospital, 3001 South Hanover Street, Baltimore, MD 21225. Email: NajjarSa@mail.nih.gov.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Evolution of Genetic Analysis: From Mendel to Complex Traits

The field of Mendelian-based genetics has traditionally been rooted in the “one gene/one trait hypothesis”. In the earliest clinical genetic studies, investigators focused on the study of rare diseases in which genetic variants found in multiple members of large families were examined to determine whether they played a causative role. Using this approach, researchers succeeded in identifying several single gene mutations that underlie predominantly autosomal dominant diseases and conditions, such as mutations in the LDL-Receptor, and APO-B genes which are associated with hypercholesterolemia (Hobbs et al. 1987, Soria et al. 1989, Kullo and Ding 2007). These exciting discoveries provided great insights into the pathophysiology of the diseases being studied. However, the successes have largely been restricted to rare, monogenic diseases.

Most common heritable conditions are “complex traits”, in that they are likely influenced by the aggregate effects of several genetic variations (or polymorphisms) that are commonly found in the population, are located in multiple distinct genes; and individually contribute only a small effect to the trait. According to the “common disease, common variant hypothesis” (Reich et al. 2001, Bodmer and Bonilla 2008) the polymorphisms or risk alleles that underlie heritable traits accumulate randomly during evolution if they do not individually cause debilitating disease. Individually, each polymorphism contributes a small effect to the trait, and some may even confer an evolutionary advantage to the health of the organism under particular conditions. The phenotypic manifestation of the trait is influenced, in part, by the interaction of these polymorphisms with each other, and is modulated by environmental factors, gene-environment interactions, epigenetic factors, and chance.

Early investigations into the genetics of complex cardiovascular traits focused on “candidate genes”. Selecting these genes requires an a priori understanding of the underlying pathophysiology of the trait. However, there is growing recognition that studies restricted to candidate genes are woefully inadequate to elucidate the genetic underpinning of complex traits, in part, because of our incomplete understanding of all the factors that regulate and influence the manifestation of a given trait. It was hoped that this could be remedied by a broader search for polymorphisms across the entire genome. Because each polymorphism is only responsible for a small proportion of the overall variability in the trait of interest, large populations are needed to gain the statistical power required to identify these polymorphisms. The feasibility of conducting genome-wide studies with sufficient resolution in these large populations, in turn, required progress on three, intertwining fronts- 1) Refinement of genetic markers, 2) utilization of linkage disequilibrium principles, and 3) advances in DNA sequencing technology.

1. Genetic Markers—The major challenges facing genetic investigators are to identify the genes of interest and pinpoint the location of the culprit polymorphisms among the billions of nucleotides that form the human genome. To achieve this, investigators have depended heavily on genetic markers. Initially, these markers were quite crude, and relied on characteristic patterns derived from cleavage of DNA by specific restriction enzymes that generated distinctive patterns known as Restriction Fragment Length Polymorphisms (RFLP). The subsequent use of microsatellite markers (e.g. variable number tandem repeats and short tandem repeats) improved the precision of these markers. More recently, completion of the Human Genome Project along with improvements in sequencing speed and analytic efficiency allowed the ultimate refinement of these markers into single nucleotide polymorphisms (SNPs).

SNPs are specific sites on chromosomal DNA where single nucleotides are known to vary among individuals. These SNPs are identified by a “reference SNP” (abbreviated rs) number. (http://www.ornl.gov/sci/techresources/Human_Genome/faq/snps.shtml). Approximately

12–15 × 10⁶ SNPs have been identified (<http://www.ncbi.nlm.nih.gov/SNP>) to date, and additional SNPs are being identified constantly. On average, SNPs are found every 100–300 base pairs in the human genome, although their distribution is not homogeneous.

While it is possible for an individual SNP to be any one of the four nucleotides (adenine, cytosine, guanine, thymine), each specific location in the genomic code is typically only one of two possible nucleotides. The nucleotide most frequently present in a population is called the “major” allele, while the other one is called the “minor” allele. SNPs can occur in both coding and non-coding regions of the gene. SNPs can be “silent”, or they can result in alterations in gene expression or protein structure/function. In general, polymorphisms are distinguished from mutations in that they are present in >1% of the population.

2. Linkage Disequilibrium and Haplotype Blocks—A basic tenet of genetic analysis is that two alleles located on the same chromosome remain linked together unless they are separated during meiotic recombination (Figure 1A). Therefore, the likelihood that two alleles (or a marker and a gene of interest) would segregate from one another is proportional to their physical distance from one another and the probability that a meiotic crossover site is located between them.

Linkage disequilibrium (LD) denotes a situation in which two SNPs are not inherited independently from each other; instead they remain “linked” to one another with a greater frequency than would be expected from random recombination events. LD is influenced by several factors, including physical distance, genetic linkage, rates of recombination, rate of mutation, random drift, and population structure (reviewed in (Reich et al. 2001, Abecasis et al. 2005) (Figure 1B). At one extreme, two SNPs may fall within a region that does not undergo internal recombination during meiosis. This region is referred to as a haplotype block. Each block is composed of thousands of nucleotides (the number can vary widely), and usually remains as a single unit during meiosis, because recombination preferentially occur at the borders that delimit the block, not within the block. There are approximately 250,000–300,000 haplotype blocks (<http://hapmap.org/whatishapmap.html.en>) in the human genome. These have been characterized and categorized by the “HapMap Project” [<http://www.hapmap.org/thehapmap.html.en>]. An important corollary is that a single genetic marker (“tag SNP”) could be used to represent this haplotype block in genetic studies. The tag SNP and the culprit mutation will co-segregate if they are located in the same haplotype block. The tag SNP can thus serve as a guiding “marker” to the true location of the variant-causing gene or polymorphism.

3. Sequencing Technology—With the rapid increase in the number of identified SNPs, quick and efficient genotyping technology was needed. This became feasible with the advent of high-throughput SNP-based multiplex systems. In one such approach, SNP microarray platforms called “chips” are made from small pieces of glass to which a large number of synthetic, single-stranded DNA oligonucleotide probes are chemically attached, which selectively bind to complementary DNA molecules (adenine to thymine, cytosine to guanine). Since each SNP commonly has two alternative alleles, each SNP will have two probes represented on the chip, one that binds to DNA with the major allele and one with the minor allele (although there can technically be any of four nucleotide variants at a site, only two probes per nucleotide, representing the most common variants are usually placed on the chip). A fluorescent marker is attached to each probe, which lights up when DNA binds to it. This is detected by a microarray reader that examines both the color and the precise location of the fluorescent signal on the microarray. Rapid advances in technology have allowed the development of chips with an ever expanding numbers of SNPs. The costs of the chips have been steadily dropping, which is allowing investigators to genotype an increasing number of subjects.

Linkage Analysis and Genome-Wide Association Studies

The aforementioned advances have allowed the introduction of “genome wide” analysis in which researchers search for phenotype-genotype associations across the entire genome, using one of two strategies: linkage analysis or association studies.

In linkage analysis, genetic markers are sequenced in members of large families to identify regions that are associated with a disease or trait of interest more often than would be expected by chance. These regions of DNA can be quite large and numerous, so further testing is often needed to identify the specific causative genes. Nonetheless, in contrast to the candidate gene approach described above which required *a priori* knowledge of the relationship of the gene to the trait or disease process, linkage scans were the first mapping technique capable of assessing the entire genome, thereby allowing evaluation of all SNPs in an unbiased or “agnostic” fashion (Riley et al. 2000). Linkage analysis has been credited with the discovery of many important genetic relationships (Wang et al. 2003, Helgadóttir et al. 2005) and is still useful in situations where large family units are being studied (Hodge 1993). The limitations of this approach include limited precision, dependence on family structure, and difficulty in identifying genes with small effect sizes (Lusis 2003).

Association studies consist of examining the correlation of a phenotype (a disease or a quantitative trait) with a genotype. When this is performed for a large number of SNPs across the genome, it is termed a genome-wide association study (GWAS) (Hirschhorn and Daly 2005). Unlike linkage analysis, in GWAS large family units are not required (in fact, excessive family structure can bias the results and must be corrected for); and as a corollary, large numbers of individuals can be studied, which improves the likelihood that SNPs with modest effect size can be identified.

SardiNIA Study Population

The SardiNIA Study recruited individuals from a circumscribed region on the island of Sardinia, Italy, who constitute a genetically isolated founder population by virtue of their geographic isolation and ethnic homogeneity. Founder populations are rare populations that arise from a delimited group living in a defined region for many centuries (founders) with minimal admixture from outside populations and reduced environmental variance. The high degree of relatedness in these populations and their environmental homogeneity make it likely that one or a few genes predominate in the causation of a complex phenotype. Thus the study of such populations facilitates detecting and refining the localization of the loci responsible for complex traits, if detailed genealogical information is available and extended multigenerational pedigrees are reconstructed (Varilo et al. 2000). In the SardiNIA study, 6,148 men and women over the age of 14 were recruited from a cluster of four towns in the Lanusei Valley in the Ogliastra province of the island, which has a total population of 11,000 (Pilia et al. 2006). These subjects underwent extensive phenotypic characterization of 98 quantitative traits, including cardiovascular risk factors (LDL, HDL, blood pressure, heart rate, CRP, etc.), anthropometrics (height, weight, waist circumference, BMI, etc), and arterial structure/function (pulse wave velocity, intimal medial thickness, etc).

As a first step, heritability analyses were performed for all the quantitative traits measured in SardiNIA (Pilia et al. 2006, Chen and Abecasis 2007). Heritability analyses evaluate the proportion of phenotypic variance that is shared among related individuals and estimate the contributions of genetic factors (Susser 1985). In the SardiNIA study, it was estimated that genetic effects explained 25% of the variance for 20 measures of cardiovascular risk factors (including systolic blood pressure diastolic blood pressure, pulse wave velocity, and intimal medial thickness), 51% for 5 anthropometric measures (including weight, waist circumference,

body mass index, and height), and 40% for 38 blood tests (including C-reactive protein, LDL, HDL, triglycerides, uric acid) (Pilia et al. 2006).

In SardiNIA, genotyping was performed with the Affymetrix 500K and 10K GeneChip arrays. GWAS in this population have yielded insights into the genetic basis of such complex traits as obesity (Scuteri et al. 2007), uric acid (Li et al. 2007), glucose (Chen et al. 2008), lipids (Willer et al. 2008), height (Sanna et al. 2008), and fetal hemoglobin (Uda et al. 2008). A major aim of the SardiNIA study remains the genetic investigation of aging-related traits. With few recent exceptions (e.g. Tarasov et al. 2009) these analyses are still ongoing, particularly in the context of meta-analyses in large multinational consortia. In the following sections, we will briefly summarize the SardiNIA GWAS experience for five separate CAD risk factors, to illustrate some of the strengths and limitations of GWAS.

Examples of GWAS from the SardiNIA Study

Lipid-Associated Genes in SardiNIA

Lipoproteins (LDL, HDL, and triglycerides) have a well-established role in CAD (Grundy et al. 2004, Rosamond et al. 2008). While smoking, diet, physical activity, and other environmental factors affect the lipid levels of individual patients, family studies suggest that about half of the overall variation in these levels is genetically determined (Pollin et al. 2004, Pilia et al. 2006). Previous genetic studies using candidate genes and linkage analysis identified several genes with rare mutations that have large effects on lipid levels; e.g. LDL receptor (Goldstein et al. 1974), apolipoprotein B-100 (Vega and Grundy 1986). However, the vast majority of individuals afflicted by hypercholesterolemia do not have any of these rare mutations. Recently, several groups, including the Framingham Heart Study (Kathiresan et al. 2007), the SardiNIA/DGI/FUSION Consortium (Willer et al. 2008), and the DGI group (Kathiresan et al. 2008) conducted GWAS that identified SNPs which are associated with lipid levels across the continuum of cholesterol levels.

However, only a subset of these SNPs are believed to be functionally significant, i.e. they are located either in coding regions of genes known or suspected of influencing lipid levels, or in regions known to regulate these genes. Other SNPs were noted to be mapping SNPs in candidate genes, i.e. they are acting as simple markers, because they are non-coding SNPs located within or near genes that are known or suspected of influencing lipid levels. Lastly, some SNPs were found to be in or near newly discovered genes, or genes that were not previously known or suspected of influencing lipid levels.

An example of a SNP that may be functionally significant is rs780094, which is strongly associated with triglyceride levels ($p=6.1 \times 10^{-32}$) (Willer et al. 2008) and is located within a coding region of the glucokinase regulatory (GCKR) gene. It may, therefore affect the level of glucose conversion to triglycerides by affecting the phosphorylation of glucose to glucose 6-phosphate in the liver by glucokinase leading to the formation of glycogen. Whether rs780094 is itself the causative genetic variant responsible for the association with triglyceride levels awaits additional studies that will evaluate the specific effects of this particular SNP on the activity of the GCKR gene and on triglyceride levels.

An example of a “mapping” SNP in a candidate gene, is one in the cholesterol ester transfer protein (CETP) gene (rs3764261), which is strongly associated with HDL levels ($p<2.3 \times 10^{-57}$) (Willer et al. 2008). Although CETP is well known to play a role in cholesterol ester transport, it is not likely that rs3764261 is the causative variant in the CETP gene that is responsible for the association with HDL level, because this SNP is located 2 kilobases upstream of the actual gene. It is possible that this SNP is located within an enhancer or other regulatory region acting from a distance on the CETP gene. However, it is more likely that

rs3764261 is only a marker for an as-yet undiscovered SNP that shares the same haplotype block as rs3764261 within the CETP gene.

Examples of SNPs in new unexpected genes are rs599839 (Willer et al. 2008) and rs646776 (Kathiresan et al. 2008), which are associated with LDL levels ($p=6.11 \times 10^{-33}$ and $p=3 \times 10^{-29}$ respectively). Both of these SNPs are located in a genetic region that was not previously known to be involved in lipid metabolism, specifically in the CELSR2-PSRC1-SORT1 locus on chromosome 1. Importantly, SNPs at this same locus were enriched in a cohort of subjects with known CAD compared to controls. This exciting finding suggests that further elucidation of the relationship between this locus and CAD may provide a potentially novel target for the prevention or treatment of CAD.

Obesity

Obesity is a cardiovascular risk factor whose prevalence is increasing at such an alarming rate that it is feared it may negate the significant strides made in decreasing cardiovascular morbidity and mortality (Rosamond et al. 2008). Although obesity is thought to be related in large part to environmental factors, it is believed to also have a genetic component, as studies have shown that 60–70% of the variation in obesity-related phenotypes is heritable (Comuzzie AG, Allison DB (1998) (Comuzzie and Allison 1998). For example, three independent groups conducting GWAS (Frayling et al. 2007) (Dina et al. 2007) (Scuteri et al. 2007) found a very robust association between obesity and the FTO gene. In the Sardinian population, a cluster of SNPs in several genes (Figure 2), were found to be associated with multiple markers of obesity, with SNPs in FTO exhibiting the most robust relationships [association of rs9930506 with BMI $p = 8.6 \times 10^{-7}$, with hip circumference $p = 3.4 \times 10^{-8}$, and with weight $p = 9.1 \times 10^{-7}$]. This example reinforces the importance of GWAS in uncovering previously unsuspected genes that underlie traits, but also illustrates one of the main limitations of GWAS, namely that the function of the uncovered gene (FTO) is currently unknown. Thus additional experiments are needed to delineate the function of this gene and to determine how it influences obesity, in order to translate this newfound knowledge into clinically useful information. This is no easy task, as illustrated in the next example.

Uric Acid

Elevated serum uric acid (UA) is an independent risk factor for cardiovascular diseases (CVD) in patients with hypertension, heart failure, or diabetes (Alderman and Aiyer 2004), even though the mechanisms underlying this relationship have remained elusive. In the SardiNIA study a GWAS found several SNPs in the GLUT9 gene to be associated with levels of uric acid (Li et al. 2007), including rs6855911 ($p=1.8 \times 10^{-16}$) and rs7442295 ($p=2.57 \times 10^{-12}$). While GLUT9 was known to function as a glucose transporter and to be expressed highly in the kidney, the link to uric acid was unexpected. A series of genetic experiments were undertaken to elucidate the role of these SNPs in connecting GLUT9 to uric acid levels. At the outset, sequence analysis of the coding regions as well as the exon/intron boundaries revealed that rs6855911 and rs7442295 were located in introns and were therefore unlikely to be independently responsible for changes in uric acid levels. Instead, they were suspected to be marker SNPs in high linkage disequilibrium with the culprit SNP, which, in turn would be located in an exonic/coding region. However, sequencing of the GLUT9 gene and putative promoter regions in a subset of the SardiNIA cohort has thus far not revealed the coding SNP that is associated with UA levels. Thus additional studies are needed that would explore novel paradigms. For example, some studies have postulated that the changes in UA levels may occur as a result of alternative splicing that modifies fructose transport which is known to stimulate increased hepatic production of uric acid (Doring et al. 2008, Vitart et al. 2008).

Glucose

Elevated fasting glucose is an early diagnostic sign heralding the presence or impending arrival of diabetes mellitus, and is a risk factor for coronary artery disease. In GWAS of fasting glucose levels in non-diabetic members of the SardiNIA and FUSION studies, a strong association ($p=5.3 \times 10^{-9}$) was found with SNP rs563694 (Chen et al. 2008). This association was strengthened further ($p=6.4 \times 10^{-33}$) in a large meta-analysis that included 24,046 subjects from the SardiNIA, FUSION, METSIM, Caerphilly, BWWHS, and Inter99 studies (Chen et al. 2008). In spite of this very robust statistical association, it is important to point out that this SNP only accounts for approximately 1% of the total variation in fasting glucose. Each copy of the major allele of this SNP is associated with an approximately 0.01–0.16 mM higher levels of fasting glucose. The majority of published GWAS to date have identified SNPs that only exert a modest effect on the trait of interest.

Pulse Wave Velocity

Central arterial stiffening is one of the hallmarks of arterial aging. Carotid-femoral pulse wave velocity (PWV) is the preferred non-invasive measure of central arterial stiffness (Laurent et al. 2006). PWV is increased in patients with hypertension, diabetes, metabolic syndrome, and atherosclerosis. Furthermore, PWV is an independent predictor (Sutton-Tyrrell et al. 2005) of mortality and of incident cardiovascular disease. A GWAS in SardiNIA identified the non-synonymous SNP rs3742207 in the *COL4A1* gene on chromosome 13 to be significantly associated with PWV, and this was successfully replicated in an independent population of European ancestry, the Old Order Amish, leading to an overall $p=5.16 \times 10^{-8}$ (Tarasov et al. 2009). Interestingly, unlike collagen types 1 and 3, which are found abundantly in the tunica media of the arterial wall and are important determinants of the tensile strength of the artery, collagen type 4 is a major structural component of the intimal basement membrane. If the association of rs3742207 with PWV is replicated in other studies, it suggests that previously unrecognized cell-matrix interactions may exert an important, and genetically-based, role in regulating arterial stiffness.

Discussion

While the power of GWAS to identify common genetic variants with small individual effect-sizes has provided a wealth of new information, there are several issues that are worthy of note. First, while large GWAS have provided the statistical power needed to identify the common genetic variants involved in several complex traits, the individual effect of each of these variants is often quite small. Common diseases with a genetic basis usually do not become manifest without the presence and/or interaction of multiple genetic variants as well as the modulating influence of environmental factors. Thus, knowledge of an individual SNP usually cannot predict whether disease will ensue, rather it provides information on the predisposition to disease. Second, statistical methods are only now beginning to address SNP-SNP interactions which are thought to underlie many diseases and traits. While most researchers are eager to look at the interaction of multiple genes in causing disease (i.e. a minor allele for a single SNP may be harmless but a combination of minor alleles at two or more locations or a specific combination of minor/major alleles may lead to disease), the statistical tools and algorithms that are needed are still in the early stages of development (Kotti et al. 2007). Third, in many instances findings from initial linkage analysis failed to replicate in subsequent studies in other populations. The reasons for this are multifactorial, and likely include differing genetic pedigrees, linkage disequilibrium, weak associations, or simple chance (Abecasis et al. 2005). Nonetheless, this experience coupled with the known risks for false positive findings in GWAS due to multiple testing, has led to a growing insistence on replication of SNP association findings in external populations prior to publication.

Currently, we are witnessing a dramatic increase in the number of GWAS being conducted, and in the size of the microarrays that are being utilized. Also, given the modest effect size of individual SNPs, it is becoming increasingly evident that ever larger study populations are needed to attain the statistical power needed to conduct the GWAS of many common diseases and traits. Investigators in the field have responded by forming collaborative consortia composed of many individual study cohorts. Furthermore, funding institutions such as NIH are proactively requesting that groups rapidly make their GWAS data publicly available so it can be used most effectively:

(http://www.ncbi.nlm.nih.gov/entrez/query/Gap/gap_tmpl/about.html). As these results become increasingly available to all investigators, it is hoped that the full potential of the promises of genetic studies can be harnessed and used towards preventing or alleviating the burden of cardiovascular diseases and their risk factors.

Acknowledgments

We thank Monsignore Piseddu, Bishop of Ogliastra; the Mayors of Lanusei, Ilbono, Arzana, and Elini; the head of the local Public Health Unit ASL4; and the residents of the towns for volunteering and cooperation. In addition, we are grateful to the Mayor and the administration in Lanusei for providing and furnishing the clinic site. We thank the team of physicians and nurses, who carried out the physical examinations and the recruitment personnel who enrolled the volunteers.

The SardiNIA ("ProgeNIA") team was supported by Contract NO1-AG-1-2109 from the National Institute on Aging.

This research was supported (in part) by the Intramural Research Program of the NIH, National Institute on Aging.

GLOSSARY

Agnostic approach

A term used to describe the GWAS approach in which the observer does not hold a pre-conceived hypothesis of the final result

Allele

Alternate forms of a gene occupying the same position (locus) on a chromosome

Association analysis

Significant disease-related genes are identified by looking for SNP alleles that are more common in individuals with a disease

Candidate gene analysis

Analysis of a specific gene believed to influence a phenotype due to known biological properties or its location near a region of association or linkage

Coding SNP

a SNP located in a portion of the gene which codes for a specific mRNA

Complex traits

Traits or diseases caused by multiple genetic variations (polymorphisms), each with small individual effect size but causing disease by their sum total effect

Genomewide association study (GWAS)

A study of common genetic variation across the entire human genome designed to identify genetic associations with observable traits. It must include measurement of markers of sufficient density to capture a large proportion of the common variants in the study population in enough individuals to provide sufficient power to detect variants of modest effect

Genomewide significance

A probability threshold required to avoid error produced by multiple testing e.g. Bonferroni-derived correction and False Discovery Rate (FDR)

Haplotype block

A block of DNA which tends to be inherited together, high LD, with sizes ranging from 3 kb to >150 kb

Linkage analysis

A gene hunting technique that looks for causative alleles by their inheritance with a closely linked marker by studying genetics of large, high risk families

Linkage disequilibrium

The tendency for alleles located closely to each other on the same chromosome to be inherited together

Major allele

The more common of two alleles for a given SNP

Mendelian law of independent assortment

Allele pairs segregate independently during gamete formation (typically because they are on separate chromosomes) and that the inheritance pattern of one trait will not affect the inheritance pattern of another

Microsatellite

A class of highly polymorphic, repetitive DNA with repeats that are 2–8 nucleotides in length and frequently used as molecular markers in linkage studies

Minor allele

The less frequent of two alleles for a given SNP

Mutation

Allelic variation present in <1% of the population

Polymorphism

Allelic variation present in >1% of the population

Recombination

The reciprocal exchange of two portions of two homologous chromosomes during gamete formation

Replication

A method to deal with errors of multiple testing in which GWAS findings are duplicated in an additional population with $p < 0.05$

Simple traits

Traits or diseases caused by single, typically rare but highly penetrant genetic variations inherited in Mendelian fashion

Single nucleotide polymorphism

a common single base pair variation occurring in >1% of the population, every 300 base pairs representing the majority of DNA variability and a useful marker for genetic localization

Tag SNP

a readily measured SNP that is in strong linkage disequilibrium with multiple other SNPs so that it can serve as a proxy for these SNPs on large-scale genotyping platforms

References

- Abecasis GR, Ghosh D, Nichols TE. Linkage disequilibrium: ancient history drives the new genetics. *Hum Hered* 2005;59:118–124. [PubMed: 15838181]
- Alderman M, Aiyer KJ. Uric acid: role in cardiovascular disease and effects of losartan. *Curr Med Res Opin* 2004;20:369–379. [PubMed: 15025846]
- Bodmer W, Bonilla C. Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet* 2008;40:695–701. [PubMed: 18509313]
- Chen WM, Erdos MR, Jackson AU, Saxena R, Sanna S, Silver KD, et al. Association studies in Caucasians identify variants in the G6PC2/ABCB11 region regulating fasting glucose levels. *Journal of Clinical Investigation*. 2008In Press
- Chen WM, Abecasis GR. Family-based association tests for genomewide association scans. *Am J Hum Genet* 2007;81:913–926. [PubMed: 17924335]
- Chen WM, Erdos MR, Jackson AU, Saxena R, Sanna S, Silver KD, et al. Variations in the G6PC2/ABCB11 genomic region are associated with fasting glucose levels. *J Clin Invest* 2008;118:2620–2628. [PubMed: 18521185]
- Comuzzie AG, Allison DB. The search for human obesity genes. *Science* 1998;280:1374–1377. [PubMed: 9603720]
- Damani SB, Topol EJ. Future use of genomics in coronary artery disease. *J Am Coll Cardiol* 2007;50:1933–1940. [PubMed: 17996556]
- Dina C, Meyre D, Gallina S, Durand E, Korner A, Jacobson P, et al. Variation in FTO contributes to childhood obesity and severe adult obesity. *Nat Genet* 2007;39:724–726. [PubMed: 17496892]
- Doring A, Gieger C, Mehta D, Gohlke H, Prokisch H, Coassin S, et al. SLC2A9 influences uric acid concentrations with pronounced sex-specific effects. *Nat Genet* 2008;40:430–436. [PubMed: 18327256]
- Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM, Lindgren CM, et al. A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* 2007;316:889–894. [PubMed: 17434869]
- Goldstein JL, Dana SE, Brown MS. Esterification of low density lipoprotein cholesterol in human fibroblasts and its absence in homozygous familial hypercholesterolemia. *Proc Natl Acad Sci U S A* 1974;71:4288–4292. [PubMed: 4373706]
- Grundy SM, Cleeman JI, Merz CN, Brewer HB Jr, Clark LT, Hunninghake DB, et al. Implications of recent clinical trials for the National Cholesterol Education Program Adult Treatment Panel III guidelines. *Circulation* 2004;110:227–239. [PubMed: 15249516]
- Hardoon SL, Whincup PH, Lennon LT, Wannamethee SG, Capewell S, Morris RW. How much of the recent decline in the incidence of myocardial infarction in British men can be explained by changes in cardiovascular risk factors? Evidence from a prospective population-based study. *Circulation* 2008;117:598–604. [PubMed: 18212284]
- Helgadottir A, Gretarsdottir S, St Clair D, Manolescu A, Cheung J, Thorleifsson G, et al. Association between the gene encoding 5-lipoxygenase-activating protein and stroke replicated in a Scottish population. *Am J Hum Genet* 2005;76:505–509. [PubMed: 15640973]
- Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* 2005;6:95–108. [PubMed: 15716906]
- Hobbs HH, Brown MS, Russell DW, Davignon J, Goldstein JL. Deletion in the gene for the low-density-lipoprotein receptor in a majority of French Canadians with familial hypercholesterolemia. *N Engl J Med* 1987;317:734–737. [PubMed: 3627182]
- Hodge SE. Linkage analysis versus association analysis: distinguishing between two models that explain disease-marker associations. *Am J Hum Genet* 1993;53:367–384. [PubMed: 8328455]
- Kathiresan S, Manning AK, Demissie S, D'Agostino RB, Surti A, Guiducci C, et al. A genome-wide association study for blood lipid phenotypes in the Framingham Heart Study. *BMC Med Genet* 2007;8 Suppl 1:S17. [PubMed: 17903299]
- Kathiresan S, Melander O, Guiducci C, Surti A, Burt NP, Rieder MJ, et al. Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nat Genet* 2008;40:189–197. [PubMed: 18193044]

- Kotti S, Bickeboller H, Clerget-Darpoux F. Strategy for detecting susceptibility genes with weak or no marginal effect. *Hum Hered* 2007;63:85–92. [PubMed: 17283437]
- Kullo IJ, Ding K. Mechanisms of disease: The genetic basis of coronary heart disease. *Nat Clin Pract Cardiovasc Med* 2007;4:558–569. [PubMed: 17893684]
- Laurent S, Cockcroft J, Van Bortel L, Boutouyrie P, Giannattasio C, Hayoz D, et al. Expert consensus document on arterial stiffness: methodological issues and clinical applications. *Eur Heart J* 2006;27:2588–2605. [PubMed: 17000623]
- Li S, Sanna S, Maschio A, Busonero F, Usala G, Mulas A, et al. The GLUT9 Gene Is Associated with Serum Uric Acid Levels in Sardinia and Chianti Cohorts. *PLoS Genet* 2007;3:e194. [PubMed: 17997608]
- Lusis AJ. Genetic factors in cardiovascular disease. 10 questions. *Trends Cardiovasc Med* 2003;13:309–316. [PubMed: 14596945]
- Nabel EG. Cardiovascular disease. *N Engl J Med* 2003;349:60–72. [PubMed: 12840094]
- Pilia G, Chen WM, Scuteri A, Orru M, Albai G, Dei M, et al. Heritability of cardiovascular and personality traits in 6,148 Sardinians. *PLoS Genet* 2006;2:e132. [PubMed: 16934002]
- Pollin TI, Hsueh WC, Steinle NI, Snitker S, Shuldiner AR, Mitchell BD. A genome-wide scan of serum lipid levels in the Old Order Amish. *Atherosclerosis* 2004;173:89–96. [PubMed: 15177127]
- Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, et al. Linkage disequilibrium in the human genome. *Nature* 2001;411:199–204. [PubMed: 11346797]
- Riley JH, Allan CJ, Lai E, Roses A. The use of single nucleotide polymorphisms in the isolation of common disease genes. *Pharmacogenomics* 2000;1:39–47. [PubMed: 11258595]
- Rosamond W, Flegal K, Furie K, Go A, Greenlund K, Haase N, et al. Heart disease and stroke statistics--2008 update: a report from the American Heart Association Statistics Committee and Stroke Statistics Subcommittee. *Circulation* 2008;117:e25–146. [PubMed: 18086926]
- Sanna S, Jackson AU, Nagaraja R, Willer CJ, Chen WM, Bonnycastle LL, et al. Common variants in the GDF5-UQC region are associated with variation in human height. *Nat Genet* 2008;40:198–203. [PubMed: 18193045]
- Scuteri A, Sanna S, Chen WM, 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 Genet* 2007;3:e115. [PubMed: 17658951]
- Soria LF, Ludwig EH, Clarke HR, Vega GL, Grundy SM, McCarthy BJ. Association between a specific apolipoprotein B mutation and familial defective apolipoprotein B-100. *Proc Natl Acad Sci U S A* 1989;86:587–591. [PubMed: 2563166]
- Sutton-Tyrrell K, Najjar SS, Boudreau RM, Venkitachalam L, Kupelian V, Simonsick EM, et al. Elevated aortic pulse wave velocity, a marker of arterial stiffness, predicts cardiovascular events in well-functioning older adults. *Circulation* 2005;111:3384–3390. [PubMed: 15967850]
- Uda M, Galanello R, Sanna S, Lettre G, Sankaran VG, Chen W, et al. Genome-wide association study shows BCL11A associated with persistent fetal hemoglobin and amelioration of the phenotype of beta-thalassemia. *Proc Natl Acad Sci U S A* 2008;105:1620–1625. [PubMed: 18245381]
- Varilo T, Laan M, Hovatta I, Wiebe V, Terwilliger JD, Peltonen L. Linkage disequilibrium in isolated populations: Finland and a young sub-population of Kuusamo. *Eur J Hum Genet* 2000;8:604–612. [PubMed: 10951523]
- Vega GL, Grundy SM. In vivo evidence for reduced binding of low density lipoproteins to receptors as a cause of primary moderate hypercholesterolemia. *J Clin Invest* 1986;78:1410–1414. [PubMed: 3771801]
- Vitart V, Rudan I, Hayward C, Gray NK, Floyd J, Palmer CN, et al. SLC2A9 is a newly identified urate transporter influencing serum urate concentration, urate excretion and gout. *Nat Genet* 2008;40:437–442. [PubMed: 18327257]
- Wang L, Fan C, Topol SE, Topol EJ, Wang Q. Mutation of MEF2A in an inherited disorder with features of coronary artery disease. *Science* 2003;302:1578–1581. [PubMed: 14645853]
- Willer CJ, Sanna S, Jackson AU, Scuteri A, Bonnycastle LL, Clarke R, et al. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nat Genet* 2008;40:161–169. [PubMed: 18193043]

A. Recombination

B. Origin of Linkage Disequilibrium

Figure 1.

A. Recombination. During meiosis, chromosomal strands may undergo exchange of blocks of DNA as they cross over with one another. The likelihood that a section of DNA will remain associated with another throughout a population group is related to several factors including population size, genetic drift, population mixture, genetic “geographic” proximity and the presence of recombination “hotspots”.

B. The evolutionary origin of linkage disequilibrium. SNP alleles typically result from the mutation of a single ancestral chromosome. This hypothetical ancestral chromosome (colored entirely black in the figure) contains the mutant allele (indicated by the red star (*)). As this mutant allele is passed down to progeny, each successive generation carrying the mutant allele will also carry a small surrounding stretch of the ancestral chromosome, though the exact length of the conserved stretch will depend on local rates of recombination, gene conversion, and mutation. Alleles within these short stretches are not associated randomly, but instead often appear in the same configuration as in the original mutant chromosome, and are said to be in linkage disequilibrium (Abecasis et al. 2005).

Figure 2.

A. Strength of association between chromosomal regions and BMI. In this “Manhattan Plot” is depicted a summary of association findings, the X-axis lists chromosomal numbers while the $-\log_{10}$ p-values of the associations are displayed on the Y-axis. The location of the FTO gene, which had one of the strongest associations with BMI, is highlighted. Also highlighted are PFKP and CLIC5 whose association with BMI was no longer significant in subsequent replication testing.

B. Summary of the association between SNPs in the region of the FTO gene and BMI. The SNP showing strongest association (rs9930506) is highlighted. Other SNPs are colored according to their degree of disequilibrium with rs9930506 ranging from high (red), to intermediate (green), to low (blue). The X-axis denotes genomic position and transcripts are indicated at the bottom of the graph, with an arrow indicating transcript direction.

C. Pairwise estimates of linkage disequilibrium (LD) between SNPs in the Sardinian population. This graphical representation of likely haplotype blocks is based on the calculation of LD and presented in terms of their genomic position. Multiple colored triangular regions of high LD are depicted as bright red. By correlating the physical location of these blocks with

the most significant association value, the probable location of the gene associated with changes in BMI that could cause the SNP-BMI association can be identified. The grey bar marks a region of linkage disequilibrium containing rs9930506 which lies within the FTO gene and near the KIAA1005 gene. This bar indicates the same genetic location represented in both figures 2B and 2C to facilitate comparison between the two.