13

Approaches for Dissection of the Genetic Basis of Complex Disease Development in Humans

Nicole J. Lake, Kiymet Bozaoglu,
Abdul W. Khan and Jeremy B. M. Jowett
Baker IDI Heart and Diabetes Institute, Melbourne, Australia

1. Introduction

When genome-wide association (GWA) studies emerged as an approach for dissecting the genetic architecture of complex disease, there were great hopes for its potential to unlock the mysteries of complex human disease, a full dissection of which had largely evaded numerous genetic linkage analyses and candidate gene studies. Based on the hypothesis that common variants influenced much of common disease development, GWA focus was initially on common allelic variants with a minor allele frequency of at least 5%. Six years and thousands of these associated loci down the track, GWA studies have seemingly reached a fork in the road. It has become apparent that the influence of common variants on complex disease development can only explain a modest proportion of phenotypic variation. Furthermore difficulty in finding the causal variant for an associated locus has slowed anticipated translation towards the clinic. Consequently there has been wide discussion surrounding the limitations of GWA studies and value of a continued focus on common variants. A key debate therefore has arisen at this nexus that compares the Common Disease-Common Variant against the emerging Common Disease-Rare Variant hypothesis that is now gathering support.

In this chapter we discuss whether further progress can be made with the GWA approach and contrast its potential against a future with increasingly affordable whole-genome sequencing. We also explore the various sequencing strategies that seek to make progress within tight research budgets. Success in elucidating the genetic architecture of complex human disease has the potential to affect many by laying the foundations for understanding the complex mechanisms of disease facilitating translation into new drug treatments and the development of biomarkers for assessment of onset of disease.

2. Background

Complex human diseases represent a serious global health concern. Development of complex common diseases including cardiovascular disease (CVD) and diabetes relies on an intricate network of interactions between genes and environment. The continuum over

which a complex disease can be defined is itself influenced by continuous traits such as CVD being influenced by weight and cholesterol levels. It is thus not unexpected that approaches to decode such complex genetic puzzles have been slow in their progress.

GWA studies survey the association of single nucleotide polymorphisms (SNPs) with a disease or trait to identify genetic loci. Commercial chips use a 'tag' SNP for each region of linkage disequilibrium to capture variation across genomes by high-throughput genotyping technologies (Manolio et al. 2009; Witte 2010). The large numbers of samples that can be processed relatively inexpensively provides GWA study designs with excellent statistical power to detect modest effect sizes that have been integral to its success. The utility of GWA produced data has enabled efficient replication studies and consequent reduction of false-positive findings. The advent of GWA was an important advance on the existing approaches used for dissecting human disease and the first approach tailored for complex disease research. Prior to this, linkage analysis used comparatively widely spaced polymorphic markers and determined their co-segregation with the disease or trait in affected families. While this proved effective for isolating rare, highly penetrant variants associated with Mendelian diseases, it struggled to provide disease gene identification due to broad localisation intervals. Additionally the variable heritability of many complex phenotypes and challenges in accurate trait measurement may have contributed to the shortcomings of those linkage studies (Hirschhorn et al. 2005). On the other hand, candidate gene studies offered higher resolution but ignored most of the genome. The reliance on a correct hypothesis for the candidate gene's involvement in disease risk can also result in false-positive associations (Witte 2010) or true associations that were difficult to replicate (Manolio et al. 2009).

The Common Disease-Common Variant (CDCV) hypothesis provided inspiration and optimism that GWA studies were a feasible approach for characterising the genetic architecture of complex common diseases. Largely brought to attention by Lander and Reich (Reich et al. 2001), the CDCV had some attractive qualities. The well established 'Out of Africa' model illustrated proliferation of the global population from a small group of founders, a dynamic which could have allowed mutations to become common (Iyengar et al. 2007). CDCV supporters (Lander 1996; Collins et al. 1998) explained that these common variants that were once neutral, or even beneficial, had become harmful in the context of dramatic environmental changes (Kryukov et al. 2007). As the HapMap project progressed, efficiently characterising patterns of linkage disequilibrium, testing of the CDCV hypothesis became a possibility using the emerging platforms and GWA study design. Free from the constriction of a functional hypothesis and candidate gene focus, GWA studies have revealed numerous loci previously not implicated in complex disease and further allowed investigation of the genetic basis of commonality between multiple complex phenotypes (Frazer et al. 2009). However despite these successes, the validity of the CDCV hypothesis is under challenge (Frazer et al. 2009) and the translation of the findings to useful mechanistic knowledge has been problematic.

3. Genome-Wide Association study designs

A GWA study can be undertaken using several different designs. The traditional GWA study approach has been a case-control design, but cohort studies and family-based approaches also have merit. Each offers unique advantages and drawbacks. Cost, time and intentions for data application are key influences over the optimal choice of design.

3.1 Case-control

Traditionally, disease associated loci were illustrated by comparison of population-based affected 'case' and unaffected 'control' groups. Control groups can be shared between studies without introducing bias (2007; Witte 2010). Case-control designs can thus be organised easily and at minimum costs for large group sizes. The ability to survey unrelated individuals is a strength of GWA studies for ease of ascertainment of study participants. However it also raises issues such as confounding effects of population stratification (Thomas et al. 2002; Wacholder et al. 2002) where the genotyped variants may have different frequencies among ethnically defined strata of a population. The differences in allele frequencies between these strata can create false-positive associations (Dickson et al. 2010). Initially there were significant concerns for the impact of population stratification on the GWA approach (Thomas et al. 2002), however recent findings (Wacholder et al. 2002; 2007; Goldstein 2009; Hindorff et al. 2009) have largely quelled this discussion by showing the confounding effects of stratification were overestimated. This is not to say however that it does not exist, but rather that careful consideration of population samples and genotyping of specific sets of variants can reduce bias and provide the ability to detect and adjust for it if present.

3.2 Cohort

Cohort design is distinct from case-control GWA study design where a sample of the population, the cohort, is randomly ascertained and traits relevant for the disease of interest measured (Manolio 2009). A primary advantage of cohort design is that it allows avoidance of dichotomising phenotypes that reduces the statistical power of GWA (McCarthy et al. 2008). It is thus particularly useful for analysis of quantitative disease-related traits such as body mass index and fasting blood glucose levels. Additionally, environmental exposures and other covariates can easily be investigated in this design (Manolio 2009), allowing dissection of the relative influences of environmental and genetic factors.

3.3 Family-based

Although they require more effort to ascertain, phenotype and sample, family-based studies offer many advantages. Using related groups reduces the likelihood of population stratification (Hirschhorn et al. 2005; Lasky-Su et al. 2010; Witte 2010), while the increased commonality of alleles among families reduces complexity associated with genetic heterogeneity. The ability to perform joint linkage and association analysis is another asset of a family-based approach (Visscher et al. 2008). Relative to unrelated populations, a family study can offer increased efficiency and statistical power per individual genotyped by enabling additional pairings such as sibling and parent-child. General belief that family studies offer substantially reduced power compared to unrelated samples (Hirschhorn et al. 2005; McCarthy et al. 2008; Lasky-Su et al. 2010; Witte 2010) stems from misinterpretation of an earlier publication by Risch and Merikangas (Risch et al. 1996) and as explained by Blangero (Blangero 2004). Their findings were based on comparison of an ill-advised family design with an optimal association design (Blangero 2004). For common complex diseases, the statistical power of a more efficient family design can be shown to be similar to GWA of unrelated individuals (Figure 1) (Laird et al. 2006). Indeed it has been argued that the advantages offered by a related group outweigh the small statistical power lost (Visscher et

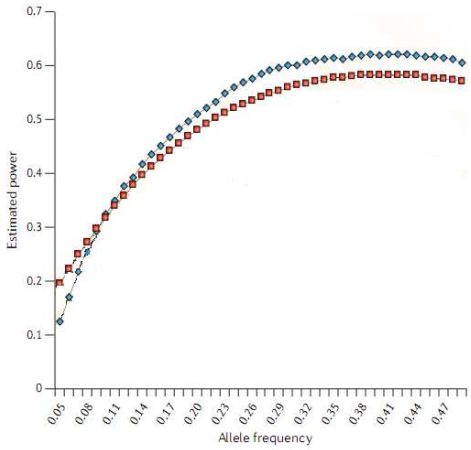


Fig. 1. The statistical power of GWA for blue case-control compared to red family trio (affected child and parents). This is modeled for a disease with 14% prevalence, similar to CVD in Australia. It should also be noted that a family design is more powerful for low frequency variants. Adapted from(Laird et al. 2006)

al. 2008). Additionally family designs allow investigation of the influence of parent of origin effects of an allele which is being increasingly recognised as an important factor in assessing disease risk (Kong et al. 2009; Manolio et al. 2009; Agopian et al. 2011). The finding that variants from parental origin are likely to confer increased risk for complex diseases (Kong et al. 2009) supports this approach. Despite favour for one or the other (Bell et al. 1997; Visscher et al. 2008), complementation of both related and unrelated groups will be important for a robust approach.

4. Limitations of Genome-Wide Association studies

GWA studies have been an unrefuted success in finding trait/disease-associated SNPs (TASs) (Hindorff et al. 2009) and candidate genes, however they have produced disappointment in establishing causal variants and disease mediating genes, that has arisen in part due to the unexpected complexity of gene transcriptional regulation and variable linkage disequilibrium structure. Furthermore, for the overwhelming majority of complex diseases studied, the collective effects of associated variants can only explain a small proportion of the trait heritability (Manolio et al. 2009). There are several possible explanations for this so called "missing heritability" that are further explored in this section.

4.1 GWA only detects associated loci and not the causal variant directly

In most cases GWA studies do not directly identify the causal variant. Once a region has been characterised as harbouring a causal locus, detailed study of all genetic variants in linkage disequilibrium with the TAS is needed to discover the causal variant. This is a difficult and time consuming process, and so a candidate gene approach relying on previously known functions of neighboring genes and their relationship to the associated trait is often used to provide direction. If the TAS lies within a gene coding sequence,

causative gene selection can be straightforward, however examples of associated SNPs within gene introns regulating another more distant gene show that selection based on proximity can by myopic (Jowett et al. 2010; Ragvin et al. 2010). Furthermore it is difficult to assign TASs that lie within an intergenic gene desert. Given that approximately 88% of TASs are located in intronic and intergenic regions (Hindorff et al. 2009) the process of identifying the causal gene is problematic. Table 1 lists how SNP location in relation to gene structure may affect a trait influencing gene.

Once the causal variant is identified, understanding how it contributes to the disease phenotype also remains challenging, particularly for non-coding polymorphisms. On account of SNP selection on high density SNP chips, GWA studies have driven investigation into intronic and intergenic regulation of gene expression. Increased understanding about the role of non-coding RNA has also illustrated a means for intergenic variants to influence gene expression. Characterisation of intergenic SNPs within long non-coding RNA has begun (Pasmant et al. 2011), and may be expected to continue.

SNP location		Potential mechanisms
Coding region	*Promoter region*	Affect transcription factor binding[1]
	UTR	Affect enhancer activity, binding of translational regulators[2]
	Exon	Missense mutation[3]
	Non-coding RNA target site	Affect miRNA regulation[4]
	Non-coding RNA sequence	Influences miRNA production[5]
Intron	*Splice junction*	Influence alternative splicing[6]
	Outside splice junction	Influence transcriptional activity[6]
Intergenic	*Intergenic non-coding DNA*	Influences gene regulation[7]
	Non-coding RNA sequence	Influence long non-coding RNA target regulation and heterochromatin formation[8]

Table 1. Potential biological consequences of a genetic variant in relation to its location. Examples of these have been described in the following; [1](Hindorff et al. 2009), [2](He et al. 2009), [3](Rutter 2010), [4](Zhang et al. 2011), [5](Ryan et al. 2010), [6](Cooper 2010), [7](Jia et al. 2009), [8](Pasmant et al. 2011).

4.2 GWA cannot discriminate against true and 'indirect' associations

Recently Dickson et al (Dickson et al. 2010) found that the strength of an association signal can be strongly affected when there is a large difference in allele frequency between the genotyped common tagging variant and causal variant. This concept of a TAS representing a diluted signal from a neighboring causal variant was named a 'synthetic association' (Dickson et al. 2010). The hypothesis proposes that when one or more rare causal variants are enriched in a haplotype with a particular allele of the common variant, the association may be falsely assigned. This raises the possibility that a common variant association may instead represent several rare variants of potentially larger effect, and consequently the true

strength of the association may have been underestimated. Furthermore these rare variants may act over large distances up to megabases to create synthetic associations (Dickson et al. 2010). This suggests that the region typically selected for re-sequencing (~500kb) may have been underestimated, and thus causal variants missed. The synthetic association hypothesis is supported by the Crohn's disease associated loci NOD2 which features three rare coding variants that drive a strong association signal at nearby common variants (Wang et al. 2010).

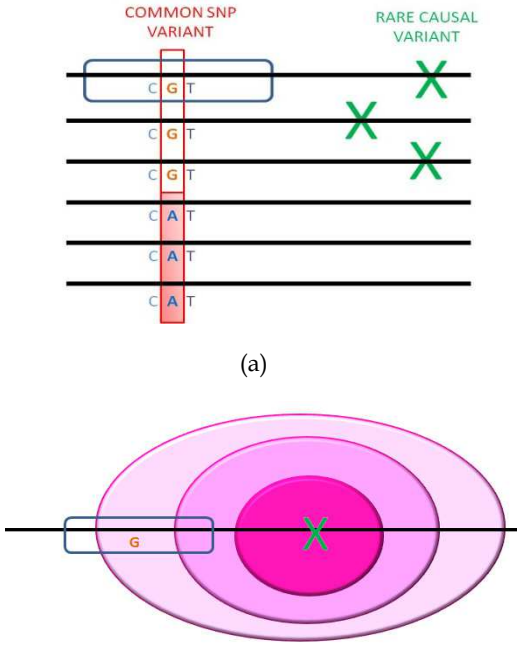


Fig. 2. (a) Synthetic Associations occur when causal rare variants are inherited more commonly with one tag SNP allele (e.g. "G") than the other ("A") allele (black lines represent different individuals). This rare variant may be outside the region of linkage disequilibrium defined by the common SNP (shown by the blue box). Adapted from (Cirulli et al. 2010). (b) Because the rare variant is too infrequent to be detected by GWAS, the true strength of the association signal will be missed. A diluted version of this signal will instead be represented by the common variant (Dickson et al. 2010).

However, debate has arisen over the validity of the synthetic association paradigm and its applicability to all common TAS (Orozco et al. 2010; Anderson et al. 2011; Wray et al. 2011). Since rare variants are likely to have arisen relatively recently, they are unlikely to be shared by divergent ethnic groups. Consequently, common variants that reflect a signal driven by rare alleles, or synthetic associations, are also expected to be largely population specific (Orozco et al. 2010; Anderson et al. 2011). A locus that harbours a cluster of low frequency variants that confer high risk of disease should be amenable to linkage mapping (Orozco et al. 2010; Anderson et al. 2011). Therefore, associations that are sensitive to linkage analysis but not ethnically replicated potentially represent synthetic associations. In light of

widespread trans-ethnic associations and discrepancies between linkage and GWA loci, it has been argued that synthetic associations are unlikely to explain the majority of GWA signals (Orozco et al. 2010; Waters et al. 2010; Anderson et al. 2011). The hypothesis remains to be further evaluated with the aid of high-throughput sequencing efforts (Orozco et al. 2010). If synthetic associations prove common, commentary about the large role of regulatory variants (Hardy et al. 2009; Hindorff et al. 2009) in complex phenotypes may be a little premature as non-coding common variants may instead represent a causal variant within a coding region megabases away.

4.3 GWA is limited to common variants

The limitation of GWA to detecting only common variants is central to literary debate over missing heritability sources. In the context of the CDCV hypothesis this was initially a logical approach, however large missing heritability for almost all complex diseases (Table 2) suggest that the CDCV may not be sustainable and that the influence of common variants on complex phenotypes has been overestimated. Despite the impressive number of variant sites genotyped by high density SNP chips, a full genome wide survey remains incomplete, with up to 20% of common SNPs not sufficiently tagged (Frazer et al. 2009). This has arisen mainly due to the remarkable and perhaps originally underestimated genetic diversity of the human species. Another confounder has been the observed complex patterns of linkage disequilibrium (LD), where so called "LD blocks" tend to fragment and fail when lower frequency variants are considered. The result is that for the most part, lower frequency variants are missed by GWA studies. Inability to detect these low frequency variants has been suggested widely in the literature as contributing significantly to missing heritability (Hirschhorn et al. 2005; Hardy et al. 2009; Manolio et al. 2009; Manolio et al. 2009; Schork et al. 2009; Cirulli et al. 2010; Dickson et al. 2010; Park et al. 2010). In this context the potential for rare variants with moderate effect sizes to influence common disease has been recognised (Manolio et al. 2009; Schork et al. 2009; Carvajal-Carmona 2010). Furthermore this paradigm has been suggested as more consistent with evolutionary genetics where less harmful variants with smaller effect sizes remain under less pressure from the action of negative selection and are therefore more likely to be common (Schork et al. 2009).

	Associated loci	Heritability	Heritability explained
Type 1 diabetes (T1D)	45 [1]	0.8 [2]	31% [3]
Type 2 diabetes	46 [4, 5]	0.42 [2]	10% [6]
Crohn's disease	71 [7]	0.55 [2]	23% [7]
Height	51 [8]	0.8 [9]	5% [9]

Table 2. Despite large numbers of associated loci identified, many complex diseases and traits have only a small proportion of heritability explained. Heritability is the proportion of phenotype explained by genetics.[1] (Burren et al. 2011), [2](So et al. 2011), [3](Clayton 2009), [4](Kooner et al. 2011), [5](Parra et al. 2011), [6](Imamura et al. 2011), [7](Fransen et al. 2011), [8](Zhao et al. 2010), [9](Yang et al. 2010).

4.4 Missing heritability

Despite numerous GWA studies and the identification of dozens of trait associated variants for each complex disease, a large proportion of heritability remains unexplained for almost all complex phenotypes examined (Table 2). The study of height illustrates the gap between discovered and estimated heritability. Extensive GWA has revealed 51 loci significantly associated height, a complex trait with high heritability of 0.80, yet these variants combined only explained 5% of the heritability (Yang et al. 2010; Zhao et al. 2010). The small contribution of these common variants to complex phenotypic variance has also limited utility of these markers for disease predication (Hirschhorn 2009). If the proportion of heritability explained is regarded as a measure of the success of the GWA approach, then its performance in elucidating complex disease genetics may appear underwhelming. However against the backdrop of many additional contributing factors such as rare variants, copy number variants, epigenetic modifications and epistasis, the achievement of the GWA approach can be considered solid progress.

4.4.1 Copy number variants

Copy number variants (CNVs) including insertions and deletions ranging from a few bases up to megabases in size have been highlighted as having potential to contribute to complex phenotypes (Frazer et al. 2009; Manolio et al. 2009; Cirulli et al. 2010). A broad study across 8 complex diseases by the Welcome Trust Case Control Consortium using existing datasets to tag CNVs found that for those CNVs typable on the employed platforms, they were unlikely to play a major role in the determination of the genetic basis of common diseases (Craddock et al. 2010). A key caveat as highlighted in the study was what could be measured by the SNP chip platforms. As noted by Alkan et al (Alkan et al. 2011), these platforms fail to detect CNV's less than 500 bp in size. These small CNVs thought to arise during recombination events and consequently are both difficult to detect and difficult to genotype. However, new sequencing data is emerging showing that this class of small CNV is not only pervasive but also likely to contribute significantly to disease risk. Therefore it is too early to dismiss the potential contribution of CNVs to complex disease development.

4.4.2 Epigenetics

Heritable epigenetic modifications to DNA and histones may be another potential contributor to complex disease processes (Bell et al. 2011). Epigenetic marks include post-translational modification of histones, non-coding RNA and DNA methylation (Rakyan et al. 2011). Epigenetic reprogramming of zygotes is used in arguing against meiotic transmission of epigenetic states, however several observations suggest that retention is possible (Rideout et al. 2001; Rakyan et al. 2003; Blewitt et al. 2006; Bell et al. 2011). Results from the study of monozygote twins show that they have more similar epigenetic profiles relative to dizygotic twins (Fraga et al. 2005). A complicating factor for investigation of epigenetic changes is that they are often tissue specific. Unlike GWA studies where most tissues are suitable for identifying germ line genetic variation, easily accessible samples such as blood may not efficiently reflect robust epigenetic alterations in other tissues that carry an impact on disease development. Since many tissue types that are important to disease development are effectively inaccessible due to unacceptable risks to the patient, such as brain, the development of comprehensive epigenetic profiles may be limited. Furthermore,

intra tissue heterogeneity presents an additional complication where a tissue, such as adipose, may be infiltrated with other cell types such as macrophages as observed in states of obesity. Reliable solutions that resolve these confounding issues are yet to be developed.

4.4.3 Environment

Phenotypic variation is traditionally considered as a result of two sources of variation; genetic and environmental (Richards 2006). While the magnitude of environmental effects on complex disease is disputed (Hardy et al. 2009; Rappaport et al. 2010), their influence on common disease development is generally accepted (Bell et al. 1997; Manolio 2009; Manolio et al. 2009; Cirulli et al. 2010). It has been proposed that the smaller the odds ratio, the more likely that environmental factors predominate (Bodmer et al. 2008) and hence a need for a more comprehensive measure of environmental exposure and its interaction with genetic variants is required (Hemminki et al. 2006; Murcray et al. 2009; Eichler et al. 2010; Rappaport et al. 2010). However if a study does record environmental factors they are usually by participant self report and can be biased by inaccurate and variable recall, leading some researchers to dismiss their inclusion as they generate too much noise. It is therefore not surprising that gene-environment interactions are currently poorly understood.

The novel characterisation of the 'exposome' by Rappaport and Smith (Rappaport et al. 2010) provides a basis to explore gene-environment interactions. The exposome represents all exposures that impact the internal chemical environment of humans (Rappaport et al. 2010). GWA studies have the potential to identify SNPs that demonstrate heterogeneity between subgroups defined by an environmental exposure (Murcray et al. 2009) and could potentially reveal loci that might otherwise have been dismissed. A good starting point for maximising GWA information about gene-environment interactions is to perform environmental-wide association studies as undertaken by Patel et al (Patel et al. 2010). This approach using exposome biomarkers can be used to propose novel environmental factors for further study, such as γ-tocopherol which was positively associated with type 2 diabetes risk (Patel et al. 2010). Notably γ-tocopherol is the most abundant component of Vitamin E in the US diet, and represents up to 50% of total Vitamin E in insulin-target tissues (Patel et al. 2010) making it a plausible candidate. A GWA study with γ-tocopherol may reveal candidates for gene-environment interaction points.

Evidence has shown that the environment can trigger heritable epigenetic changes (Jirtle et al. 2007; Ng et al. 2010) however this requires more investigation to determine the extent of influence on gene expression. Indeed it has been argued that GWA studies might be misleading without consideration of environmental factors (Bell et al. 1997). Additionally the observation that some environmental risk factors can have a 'heritable' component, such as a heritability estimate of 0.6 for regular tobacco use (Kendler et al. 2000), suggests that the environmental component of complex disease can be "contaminated" with heritability arising from behavioral and risk taking psychological traits (Petronis 2010).

4.4.4 Is the metric wrong?

Realisation that epigenetic variation can contribute to complex phenotypic variation (Richards 2006; Petronis 2010; Bell et al. 2011) has important implications for heritability estimates. Discussion around the potential for human transgenerational epigenetic

inheritance (Richards 2006; Petronis 2010; Bell et al. 2011; Rakyan et al. 2011) has suggested that the genetic contribution to complex phenotypes may have overestimated. Therefore attributing heritability estimates solely to genetic variants may be short sighted if epigenetic marks are significantly heritable, as Slatkin suggests (Slatkin 2009). Heritability can differ temporally and between environments (Heath et al. 1998; Turkheimer et al. 2003; Visscher et al. 2008) implying that dissecting phenotypic variation is not as simple as currently treated. Daily variance in some traits such as blood glucose and difficulty in standardising measurement have also contributed to variable estimates of trait heritability. However on balance, the estimates today are not likely to be sufficiently far from the actual so as to remove all of the missing heritability for complex diseases.

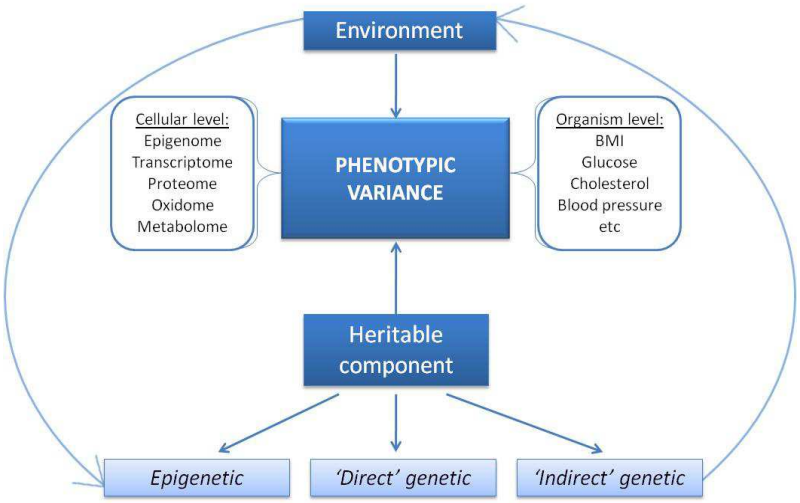


Fig. 3. Heritability estimates represent the genetic variation contributing to phenotypic variation and may be considered as divisible into 3 components; epigenetic, 'direct' genetic and 'indirect' genetic. Epigenetics and 'direct' genetics influence the phenotype directly, whereas 'indirect' genetics represent the heritable portion of behaviours that contribute to environmental influences, such as heritability of regular tobacco use. The environment may induce epigenetic changes suggesting that the two traditional components of phenotypic variance, namely the environment and the heritable portion, may not be wholly distinct.

4.4.5 Epistasis

The extent of missing heritability encouraged revision on the factors that influence genetic effects on disease development. Epistasis in genetics is defined as the interaction between genes; a phenomenon where one gene influences the expression of another. For example the E4 allele of the apolipoprotein E (ApoE) gene is associated with increased serum cholesterol levels, but only when the individual possesses the A2A2 genotype of the low-density lipoprotein receptor (LDLR) gene (Tyler et al. 2009). ApoE's phenotypic effect thus depends on the LDLR locus. Epistatic effects amplify the complexity between genotype and phenotype, suggesting that the combination of genotypes inherited may be instrumental to

disease risk (Moore et al. 2009), and it has been argued that epistasis is likely to be a ubiquitous component of the genetic architecture of common complex disease (Moore 2003; Moore et al. 2009). Humans have evolved robust biological systems that use network redundancy to confer resistance to genetic or environmental fluctuations (Moore et al. 2009). Disease could be perceived as an accumulation of parallel pathway 'breakdowns', whereby an important pathway and all of its 'backups' become dysfunctional. In this context, the particular combination of alleles inherited will determine disease phenotype. This concept could explain failure to replicate significant SNP associations and the low effect sizes of each. When considered in a multilocus model, a variant may exert a larger effect than estimated with a single locus approach (Wang et al. 2005). Indeed a "genometype" as proposed by Moore (Moore 2009) would ideally be considered as a basis for calculating clinical risk.

The computational burden associated with statistical analysis of epistasis on a genome-wide scale is unfortunately a limitation in dissecting complex disease genetics. To consider all possible pair-wise interactions alone, approximately 500 billion SNP-pairs would have to be examined (Zhang et al. 2011). Development of efficient two-locus epistasis tests has been challenging as reflected in a recent study by Bell et al (Bell et al. 2011). A two-locus analysis on type 2 diabetes GWA data identified 79 significant pair-wise results, but all included a TCF7L2 SNP which was the most significant in a single-locus analysis. It was realised that the strong TCF7L2 signal was driving the two-locus significance results, ultimately obscuring multilocus findings and signaling a hurdle for such analysis. The complex computation for investigation of genome wide epistasis therefore remains a significant challenge.

5. Approaches to addressing GWA limitations

Despite its shortcomings, GWA studies have produced data that is rich in information. Further optimisation and re-analysis may yield additional insights from this approach while identification of disease genes within the loci is expected to improve current understanding of complex disease development.

5.1 Transcriptome analysis enables prioritisation of candidate variants

Considering that most trait associated loci feature variants that do not affect protein sequence, it is anticipated that many of these loci have a regulatory role on gene expression (Heap et al. 2010), although as discussed above this assumption may be somewhat premature. Genomic regions that contribute to phenotypic variation by influencing transcription, mRNA stability and splicing are termed expression quantitative trait loci or eQTLs (Heap et al. 2010). Integration of expression profiles with GWA data can reveal eQTLs by identifying transcripts that correlate with the TASs (Figure 4). This approach allows empirical ranking of juxtaposed candidate disease-mediating genes that can be taken forward into *in vitro* and *in vivo* functional biological assays to determine the role and mechanism in the disease process. Translation of these eQTLs into enhanced biological knowledge of disease has been seriously impeded by the dearth of large scale human datasets that carry both genetic and transcript data from the same individual that facilitate identification of the disease mediating gene(s) (McCarthy et al. 2008). However some datasets do exist and a demonstration of this approach has been reported where variants in an intronic region of FTO were found to be strongly correlated with gene expression of a more distant gene RBL2 (Jowett et al. 2010). Similarly, approaches relying on conservation of synteny and nucleotide sequence among vertebrate

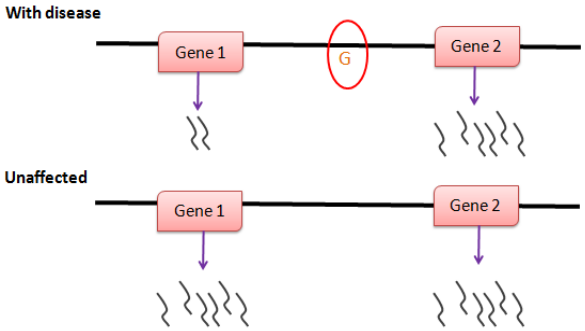


Fig. 4. GWA has found the circled SNP to be associated with a disease. It can be difficult though to assign this SNP to a candidate gene. Transcriptome analysis can make this easier by comparing the expression levels of flanking genes to each other, and between affected and unaffected individual. The decreased expression of gene 1 is associated with the disease, suggesting that gene 1 is the better candidate gene.

species have highlighted a second gene IRX3 flanking FTO that may contribute to disease processes (Ragvin et al. 2010). Further function validation is required to resolve the relative contribution of these genes at this locus.

The study of eQTLs has shown that they generally explain a greater proportion of the phenotype than risk alleles (Freedman et al. 2011), suggesting that consideration of expression data may uncover missing heritability. It has also reinforced that eQTL associations can be tissue-specific (Freedman et al. 2011), differ between *in vivo* and *in vitro* systems and also between cell types (Goring et al. 2007; Farber et al. 2009; Hardy et al. 2009). Evaluation of expression profiles using a single platform for liver, adipose tissue and blood showed a 30% overlap in the three tissues (Nica et al. 2008). This observation suggests that there is value in using any available transcriptome data set for analysis, but also highlights that the majority of expression patterns might be tissue-specific (Nica et al. 2008). The present lack of a comprehensive data set detailing expression profiles of normal tissues calls for caution in interpreting significant associations derived from tissue–specific expression data. This issue has encouraged the Genotype-Tissue Expression project, an idea which was pre-empted in the literature (Goring et al. 2007; Nica et al. 2008). Allelic-specific expression can be used to characterise an eQTL in an individual heterozygous for the risk allele and offers a means of avoiding sources of error that can confound assessment of expression profiles across multiple individuals (Heap et al. 2010). Gene expression can also differ temporally, especially for developmentally important genes. Model organisms such as the zebrafish have utility in investigating the effects of non-coding elements on gene expression in a particular context (Ragvin et al. 2010). Defining the appropriate tissue and time point to glean human expression data from is not clear, and is an area earmarked for further development (Freedman et al. 2011).

5.2 A wider region should be considered for fine mapping

Identification of synthetic associations discussed above has shown that a wider region encompassing the TAS should be considered during fine mapping. If rare variants can act

over megabase distances to drive an association signal (Dickson et al. 2010), then the average region size selected for investigation (~500kb) may be too narrow to capture causal variants. This approach may increase the explained heritability for complex phenotypes as the combined influence of multiple rare variants could exhibit a larger effect collectively than the synthetically associated common variant (Dickson et al. 2010) where the effect size has eroded due to distance and LD. Following up on non-reproducible common associations and loci that have known linkage peaks within a few megabases might increase the likelihood of finding a causal variant.

5.3 Integrate with Epigenetic-Wide Association studies

Recent technology advances have allowed contemplation of epigenome-wide association (EWA) studies (Rakyan et al. 2011), where up to 450,000 DNA methylation sites can be measured on a single chip. Although the primary focus of the analysis of these results will be on DNA methylation, there is evidence to suggest that information might be obtained on other epigenetic marks such as histone modifications (Bernstein et al. 2007) and can thus have a broader application. Development of epigenetic maps by the International Human Epigenome Consortium (Rakyan et al. 2011) will empower EWA synonymous to the utility of the HapMap in GWA studies. Additional insights might arise where loci harbouring genetic variation influence the methylation state of the region (methQTLs) (Rakyan et al. 2011). Therefore there is potential utility in combining data from GWA and EWA studies as illustrated by a recent study where a haplotype-specific DNA methylation locus for a type 2 diabetes risk variant was identified (Bell et al. 2010).

While EWA holds great promise for revealing genetic architecture of complex disease, it also presents unique challenges. A significant difference from GWA is that epigenetic association with a phenotype can be causal as well as consequential (Rakyan et al. 2011). Although this adds another layer of complexity to EWA analysis, it can be dissected by appropriate study design. Randomly ascertained longitudinal cohort studies will be integral to determining the temporal origins of the reported epigenetic association and elucidating whether its origin is pre or post onset of the phenotype (Rakyan et al. 2011). Furthermore with thorough detailing of environmental changes, this study design may reveal an environmental cue for epigenetic change. However this presents a double-edged sword, with environmental factors such as smoking able to confound EWA (unlike GWA) and inflate effect size estimates (Rakyan et al. 2011). Additionally ascertainment, phenotyping and sample collection of a large cohort as would be required for such a study is not a trivial undertaking. Looking at this issue from another perspective, it presents a means to investigate environmental backgrounds which increase epigenetic risk variants. Such information has scope for application in a clinical setting, and could be useful in encouraging behavioral changes in patients with a 'vulnerable' epigenetic locus.

5.4 Reduce statistical rigour

The statistical rigour applied to GWA study data analysis has been suggested as potential explanation for the observed missing heritability (Yang et al. 2010). Variants with weak but relevant effect sizes which fall under stringent thresholds resulting from adjustment for multiple testing may potentially be excluded from analysis to reduce the false positive rate. While commonly each TAS is evaluated individually, Yang et al considered a model that

evaluates the variance explained by a group of TASs collectively (Yang et al. 2010) using a software program that estimated the variance explained by the group on a chromosome or on the whole genome (Yang et al. 2010). Application of this approach to human height GWA studies revealed that as much as 40% of missing heritability may be found by such accumulation of the variance explained (Yang et al. 2010). The biological plausibility of considering ~300,000 loci to influence this trait is debatable. However the principal has been demonstrated and further functional studies are required to empirically determine appropriate statistical rigour that may differ from one phenotype to the next. Application to predict disease onset found improvement with the whole genome method (Makowsky et al. 2011), however genotyping thousands of variants is unlikely to be sufficiently cost effective and carry a substantially large clinical benefit to warrant its widespread use.

5.5 Increase study size

Even a relatively large GWA study (5000 cases and 5000 controls) has relatively low power to detect an association; 0.9% at a P-value of 10^{-7} (Kraft et al. 2009). Therefore increasingly large cohorts have been studied in addition to the combining existing sets into the meta-analysis studies (Zeggini et al. 2008; Barrett et al. 2009; Park et al. 2010). While this has been productive to an extent, increasing size also increases the potential for increased genetic heterogeneity and thereby raises the background noise level. Additionally since such large cohorts are impractical to collect from a single clinic or site, they must be undertaken at multiple participating centers that increase the chance for differences in phenotyping accuracy. The point at which additional loci might be discovered against increasing amount of background noise might have been reached as demonstrated by a recent GWA study for central obesity and fat distribution using waist and waist-hip measurements. The cohort consisted of ~100,000 participants, but the results yielded only three associated loci, with each explaining a mere fraction (0.05%, 0.04% and 0.02%; combined total 0.11%) of the total trait variance (Lindgren et al. 2009).

An alternate to increasing study size is to increase diversity, that is to undertake similarly sized GWA studies in a range of diverse ethnic groups. It has been established that some SNPs are population specific (2010) and that any given variant may show a different effect size between differing ethnic groups (Dickson et al. 2010). Since many GWA studies had an initial focus on European populations, examining non-European groups may reveal additional loci. Such studies are now starting to appear and have successfully identified new sets of loci (Kooner et al. 2011; Parra et al. 2011). While it is apparent that a causal variant may not be consistently associated among different populations due to unique genetic and environmental factors, a population-specific variant should also not be discounted as having limited clinical application. This is because of the commonality of intra- and inter-cellular molecular pathways in humans. That is any loci leading to identification of a disease gene will reveal a pathway that will be present across all members of the species, and as such is expected to have broad application.

The observed missing heritability suggests that many inherited factors remain to be discovered. For example, if further pursuing common variants, a crude statistical estimate suggested that there may be up to 800 additional variants yet to be discovered for type 1 diabetes (So et al. 2011). However Goldstein argues that analytical approaches based on discovery of common variants have been exhausted and that if any loci are left to discover, it will not be worth the time and cost of ever larger studies to detect them (Goldstein 2009).

This stance is supported by Park et al who proposed that additional loci discovered in larger studies will generally have smaller effects (Park et al. 2010) and therefore may be limiting for detecting large effect variants. Such estimates have provided additional momentum to a paradigm shift that analysis of complex disease should be redirected toward rare variants with potentially large effects.

6. Case study: Type 2 diabetes

The pursuit for genetic factors that underlie Type 2 Diabetes (T2D) encapsulates the challenges and rewards of GWA studies. Almost 350 million people suffer from diabetes mellitus, with approximately 90% of those having developed T2D specifically (Danaei et al. 2011). Characterised by insulin resistance and abnormal beta-cell function (Imamura et al. 2011), the genetic contribution of T2D was well established by family and twin studies (Groop et al. 1996; Poulsen et al. 1999). Current heritability estimates for T2D predict ~40% of the phenotype is explained by genetics (So et al. 2011).

With the incidence and consequent economic impact of this common complex disease projected to increase considerably (Colagiuri 2010), there has been a strong focus on identification of the genetic basis of T2D development. Although GWA has facilitated the identification of over 45 genetic susceptibility variants (Kooner et al. 2011), the promise of clinical application remains unfulfilled.

6.1 The GWA study boom

Prior to the advent of GWA studies, linkage analysis and candidate gene studies were used to discover loci associated with T2D. Despite extensive efforts spanning a decade (Frayling 2007) progress was slow. Notable achievements included identification of PPARG and KCNJ11 as candidate genes (Altshuler et al. 2000; Gloyn et al. 2003), and TCF7L2 as the common variant driving linkage at a T2D risk gene region (Grant et al. 2006). Additionally linkage analysis identified several hundred loci, of which over 50 were replicated by 5 or more independent studies (Lillioja et al. 2009). As GWA studies emerged from 2007, a sudden increase of T2D associated loci followed, reflecting the potential of GWA to provide genetic clues of disease etiology (Table 3).

6.2 The role of epigenetic modification

As has been observed with almost all complex diseases, only ~10% of the heritability can be explained despite over 45 associated loci reported (Imamura et al. 2011). As discussed above, there are several possibilities that may explain the missing heritability including rare variants, copy number polymorphisms and inherited epigenetic polymorphisms. Epigenetic modification as a result of environmental influences and in particular the fetal environment has been highlighted as having potential to predispose to T2D (Ling et al. 2009; Liguori et al. 2010). The Dutch Hunger Winter Study revealed an association between prenatal exposure to famine and decreased glucose tolerance (Ravelli et al. 1998). Effects on methylation were also investigated with all but one famine exposed individual showing significant hypomethylation of the known T2D candidate gene insulin-like growth factor 2 (IGF2) relative to their unaffected sibling (Heijmans et al. 2008). This example highlights the vulnerability of the fetal genome to the prenatal environment, and the ability of epigenetic changes to persist between generations.

Candidate genes	Gene product	Effect size odds ratio
KCNQ1	*Potassium voltage-gated channel*	1.43
TCF7L2	*Transcription factor*	1.37
DUSP9	*Dual specificity phosphatase*	1.27
UBE2E2	*Ubiquitin-conjugating enzyme*	1.19
IRS1	*Insulin receptor substrate*	1.19
IGF2BP2	*Insulin-like growth factor mRNA binding protein*	1.17
FTO	*Dioxygenase that repairs alkylated DNA and RNA by oxidative demethylation*	1.15
KCNJ11	*Potassium inwardly-rectifying channel*	1.15
THADA	*Thyroid adenoma associated protein*	1.15
PPARG	*Peroxisome proliferator-activated receptor*	1.14
HHEX-KIF11-IDE	*HHEX: homeobox family transcription factor, KIF11: kinesin-like protein, IDE: insulin-degrading zinc metallopeptidase*	1.13

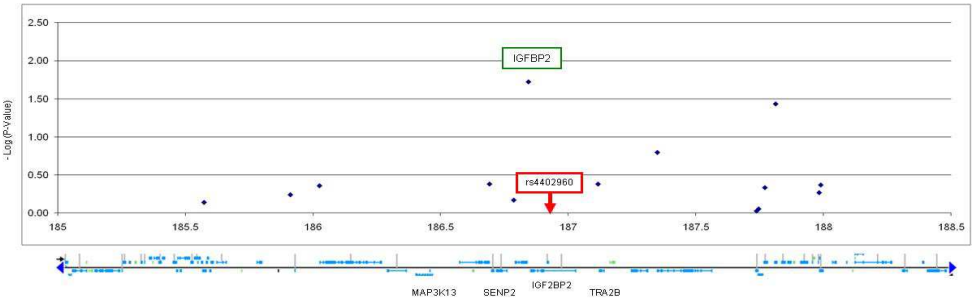
Table 3. Top ranking candidate genes for type 2 diabetes as per estimated effect size. Odds ratio's as listed in (Imamura et al. 2011).

6.3 The slow walk from association to mechanism

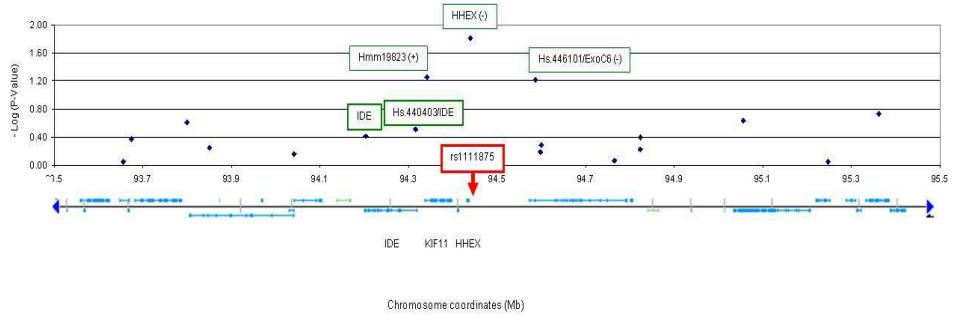
While T2D GWA studies have been successful for revealing associated loci, it has proved difficult to understand the mechanisms through which these loci mediate their biological consequences. Here we examine this aspect of translation of GWA identified loci to useful biological knowledge using several examples to illustrate.

6.3.1 IGF2BP2

An intronic locus within translation regulator IGF2 mRNA binding protein IGF2BP2 was significantly associated with T2D, implicating IGFBP2 and supporting the candidacy of IGF2 and its regulation in T2D. Previously unpublished data using transcriptome analysis combined with genetic variation data in a large human cohort showed a nominally significant association between IGFBP2 and the intronic TAS supporting the hypothesis that this locus mediates its effects through regulating the expression of IGFBP2 (Figure 5). The methodology for this analysis was as reported for the FTO loci by Jowett and colleagues (Jowett et al. 2010). Investigation of top T2D GWA candidates in the Dutch Hunger Winter cohort showed that IGF2BP2 variants were most strongly associated with increased risk of T2D, impaired glucose tolerance and area under the curves (AUC) for oral glucose tolerance testing (OGTT) (van Hoek et al. 2009). The presence of the IGF2BP2 risk allele in those exposed to famine was nominally significantly associated with lower AUC for glucose. That two factors which confer risk to T2D are together associated with increased glucose uptake may seem contradictory, but in light of a similar observation for another risk allele (van Hoek et al. 2009) this can be explained. IGF2BP2 variant may instead represent a protective allele, acting to repair the detrimental consequences of fetal malnutrition on glucose tolerance in later life (van Hoek et al. 2009). IGF2BP2's mode of action in T2D development is still unclear, with recent research revealing its capacity to both initiate and repress



(a)



(b)

Fig. 5-a. Association of the rs4402960 SNP with the expression levels of genes within 5Mb of the SNP coordinate. Strong association of the IGFBP2-residing SNP with IGFBP2 is observed, suggesting that this locus mediates its effects through IGFBP2 regulation of expression. 5-b. Association of the rs1111875 SNP with the expression levels of genes within 5Mb of the SNP coordinate. Strong association of the SNP with HHEX is observed, suggesting that this loci mediates its effects through HHEX regulation of expression. The intron/exon structure of the genes in this region is illustrated by the lines and blocks beneath the x-axis (thin line, intron; block, exon). The negative log of the P-value is plotted against the coordinates in megabase pairs (Mb). The methodology for this analysis was as reported for the FTO loci by Jowett and colleagues (Jowett et al., 2010).

translation of targets (Boudoukha et al. 2010; Dai et al. 2011). This reported association between IGF2 hypomethylation and IGF2BP2 risk variant suggests scope for an increased risk in an individual that possesses both marks. Investigation of IGF2 expression levels in the Dutch Winter Cohort may provide insight into IGF2BP2's role by reporting a condition that exacerbates the variants risk effects.

6.3.2 HHEX-KIF11-IDE

The HHEX-IDE-KIF11 locus encompasses an extended region of high linkage disequilibrium spanning three genes, two of which were considered plausible candidates; the homeobox HHEX transcription factor involved in pancreatic development (Prokopenko et al. 2008) and

the insulin degrading enzyme IDE. Two different approaches suggested that the locus might regulate HHEX expression as the mediator of its biological consequences (Ragvin et al. 2010) (Jowett unpublished). As described above, relying on a large human cohort with transcriptomic and genetic variation data, we observed that HHEX was the most strongly associated transcript with the associated variant. IDE was not associated, while KIF11 was not within detectable limits of expression and could not be commented on further. In the second approach it was observed that the TAS for this locus was within a highly conserved noncoding element (HCNE) exhibiting synteny between a number of vertebrate species including mouse, chicken, frogs and zebrafish (Ragvin et al. 2010). Subsequent reporter gene studies of the enhancer region found that expression was specific to the pancreatic islet cells in a zebrafish model (Ragvin et al. 2010). However there is also evidence for a role for IDE in T2D development (Farris et al. 2003; Pivovarova et al. 2009), suggesting potential for an associated locus to modulate expression of more than one gene in the region.

6.3.3 CDKN2A-CDKN2B

With the intergenic SNP lying 125kb upstream of the closest protein-coding gene, the CDKN2A-CDKN2B locus also demonstrates a difficult scenario for assigning a candidate gene. These genes encode cyclin-dependent kinase inhibitors known to have an important role in β-cell function and regeneration (Bao et al. 2011), and thus are plausible biological candidates for T2D. This SNP is not lost in the 'gene desert', but rather it maps to a long non-coding RNA ANRIL, which lies anti-sense to CDKN2B and has been shown to have a pivotal role in regulating CDKN2A/B expression in mice (Pasmant et al. 2011). ANRIL achieves this by interacting with polycomb repressive complexes to form heterochromatin at the locus, ultimately leading to its repression (Aguilo et al. 2011). Given that over expression of the CDKN2A orthologue in a rodent model induces a T2D phenotype (Krishnamurthy et al. 2006), the regulatory action of ANRIL is likely important in T2D development. As more is learnt about the abundance and function of non-coding RNAs in general, it may not be unexpected that this mode of regulation might apply in more GWA loci where there are no obvious candidate genes (Pasmant et al. 2011).

6.3.4 TCF7L2

The TCF7L2 intronic SNP association is highly replicated and carries one of the largest effect sizes of all GWA loci identified to date. Possessing two risk alleles increases T2D risk by 80% (Florez et al. 2006). The loss of TCF7L2 has been observed to induce impaired β-cell function and apoptosis (Shu et al. 2008; da Silva Xavier et al. 2009), providing an explanation of its effect on T2D development. Paradoxically, decreased protein levels in diabetic islets were correlated with increased TCF7L2 transcript levels (Le Bacquer et al. 2011). Correlation of the risk allele with a 'more open' chromatin state and enhanced access for transcriptional machinery (Gaulton et al. 2010) may provide a possible explanation for observed increase in mRNA (Billings et al. 2010), but how reduced protein levels result from increased mRNA expression remains to be explained. Reports that different isoforms may exert opposing effects on β-cell survival (Le Bacquer et al. 2011) may contribute to an explanation of this paradox. Future experiments to measure the relative expression levels of the pro and anti-apoptotic splice variants in islet cells from people with diabetes may be informative. TCF7L2 illustrates that understanding a strong candidate gene's role in disease pathogenesis can be difficult, even despite extensive effort.

As further mechanistic work progresses providing biological information of which genes mediate the consequences of genetic variation and how this happens, links between these genes may become apparent thus revealing potential insight into networks and pathways that control disease mechanism. For example, IGF2 has been shown to compete with insulin as an IDE substrate (Misbin et al. 1983; Ding et al. 1992), and thus IDE is likely to act in the same pathway as IGF2BP2. Additionally, HHEX and TCF7L2 are both targets of the WNT signaling cascade (Frayling 2007) and may operate collectively on the output of this pathway. Such commonality of pathways between T2D associated loci highlight the potential for a broader understanding of the etiological basis of an association, and provide a rationale for direct testing for evidence of genetic interaction.

Limitations	Type 2 diabetes example	Approach
A SNP may be regulating a gene other than the obvious candidate	*FTO intronic SNP*	Use transcriptomics to determine a variant's mechanism of action
Difficulty in establishing the mechanism of action for a strong candidate gene	*TCF7L2*	Characterise alternative spice forms
Only detects loci associated, and not the causal variant, and as such a region may harbour more than one plausible candidate	*HHEX-IDE region*	Consider one TAS at a time; use model systems of highly conserved non-coding elements
Associated SNP is a large distance from closest gene	*CDKN2A-CDKN2B locus*	Look beyond protein-encoding genes
Large missing heritability	*Only ~10% explained*	Extend analysis to consider rare variants, epigenetic and epistatic effects , collect large trans-ethnic samples

Table 4. Analysis of GWA data has been challenging. Approaches to overcoming these limitations for type 2 diabetes loci may be useful for other TASs.

6.4 The rewards of GWA studies

Despite the numerous loci identified by GWA studies, translation of this data into novel therapeutics and risk prediction tools has been slow. The primary cause has been difficulty in establishing the disease mediating gene, where in many cases annotation by proximity can be misleading. Even with a clear understanding of causality, a gene product may not necessarily represent a chemically tractable drug target. Characterisation of T2D locus SLC30A8 represents a potential exception, where the associated SNP encoded a missense mutation (Arg → Trp) in the C-terminus of the protein (Rutter 2010). SLC30A8 encodes a zinc transporter involved in insulin storage and secretion (Imamura et al. 2011) and has also been implicated in type 1 diabetes (Wenzlau et al. 2007). A person with two copies of the risk allele experience a 53% increased risk of developing diabetes (Rutter 2010) most likely due to reduced transporter activity. A drug that enhances the activity of this transporter may potentially increase insulin secretion and thus lower blood glucose levels (Imamura et al. 2011), however chemical agonists are harder to develop then antagonists, thus the prospect of new drugs from this research remains confined to the future.

The use of T2D risk alleles in disease prediction to date has also found limited utility in clinical practice. Adding a genetic risk score to established prediction models based on clinical traits such as BMI and family history showed only limited improvement in predictive power (Lyssenko et al. 2008) and is not sufficient to warrant the additional cost of genotyping. Other studies (de Miguel-Yanes et al. 2011) have suggested that such clinical applications might carry relatively greater benefit for younger people since they may not have developed age-dependent clinical risk factors such as hypertension and dyslipidemia. Discussion about the potential for a genetic risk score alone to influence healthy patients to make lifestyle changes (Grant et al. 2009) suggests that using raw information may be an attractive alternative to producing numerical scores. Clearly there is some way to go before results from GWA studies can be used to their full fruition, but the progress in risk loci characterisation gives hope that novel insights will be gained with potential benefits eventually feeding into clinical practice.

7. Future direction: Common Disease-Rare Variant hypothesis

In light of the large proportion of unexplained heritability for most common complex diseases, new tactics have been employed to advance its genetic characterisation. The Common Disease-Rare Variant (CDRV) hypothesis has become the most attractive theory for explaining the failures of GWA studies (Nielsen 2010). This model proposes that rare variation is the major contributor to complex diseases, with genes or genomic regions potentially harbouring many different rare variants (Schork et al. 2009). In contrast to the CDCV, rare variants are expected to have arisen relatively recently, or have become rare due to negative selection (Schork et al. 2009). Therefore the Multiregional evolution hypothesis that describes multiple founder groups contributing to the origin of modern humans is supportive of the CDRV rather than the Out of Africa model (Iyengar et al. 2007). Candidate genes proposed by GWA for type 1 diabetes (Nejentsev et al. 2009), hypertriglyceridemia (Johansen et al. 2010) and Crohn's disease (Momozawa et al. 2011) have been found to harbour rare missense mutations, highlighting the plausibility of the CDRV theory. Particularly alluring is that less frequent variation is more likely to be functional (Schork et al. 2009). Indeed the majority of rare (<1%) missense mutations present in humans are somewhat deleterious in nature (Kryukov et al. 2007), providing hope that rare variants will carry larger effect sizes and hopefully more obvious in mechanism than the often cryptic common variation.

7.1 New approaches

Despite claims that expectations for finding missing heritability of complex disease in rare variants may be overoptimistic (Carvajal-Carmona 2010), recent studies have highlighted the promise of the study of rare variants in complex disease (Bowden et al. 2010; Holm et al. 2011). DNA sequencing is the most efficient method for the identification of rare variants and while only a few years ago this would be impractical due to costs and time for large sample numbers, the advent of next generation sequencing technologies has bought this approach into the realm of possibility (Ng et al. 2009; Cirulli et al. 2010; Nielsen 2010). To capture all genetic variation, whole-genome sequencing (WGS) would be the preferred approach and has already found utility in characterisation of rare monogenic disease (Choi et al. 2009). While WGS costs are becoming more affordable, they remain impractical for large sample sizes, prompting development of sequencing approaches that will reduce costs

such as target enrichment and sample multiplexing. Cohort design for rare variants also forms an important consideration, with large extended pedigree designs offering the most efficient method for generating sufficient copies of the rare variant to undertake meaningful statistical analysis. These strategies (Figure 6) will represent the next stage of genetic dissection for complex diseases.

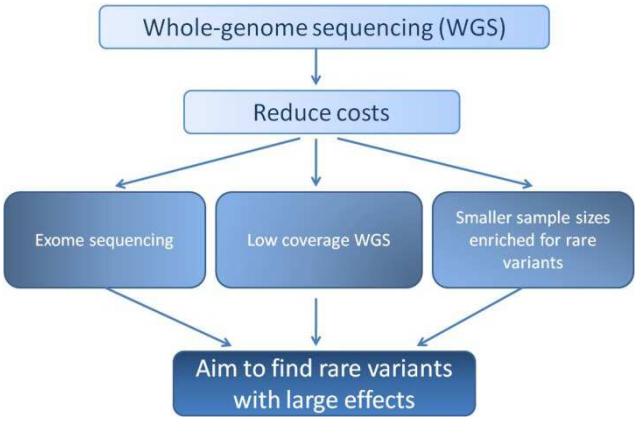


Fig. 6. Until WGS is affordable, alternatives such as exome sequencing, low-coverage WGS and smaller sample sizes reduce costs for a sequencing approach.

7.1.1 Exome sequencing

The exome (all genomic coding sequence) represents approximately 2.3% of the genome (Lehne et al. 2011) and is the genomic region most likely to harbour variants of large effect. Consistent with this has been the experience from GWA studies where stronger association signals for coding variants than non coding variants have been reported (Lehne et al. 2011). Given its smaller size it is also less expensive and faster to sequence than the whole genome allowing a large number of samples to be analysed and increasing statistical power for evaluation of effects. Successful identification of rare causal variants by this approach has been recently reported for several Mendelian diseases (Ng et al. 2010; Ng et al. 2010) and also for complex disease traits (Bowden et al. 2010) and disorders (Sanders 2011). However focusing solely on the exome ultimately ignores the majority of the genome where regulatory variation is located. Sequencing exomes in blocks that include flanking regions can increase genome coverage, with 40kb flanks equating to 34% of the genome (Lehne et al. 2011). The size of the flanking region considered will be cost-limited, but a focus on annotated regulatory elements could be used in a gene specific manner and customised to limit the amount of sequence required and reduce costs (Lehne et al. 2011). Ultimately exome sequencing is considered a temporary measure for making advances before the cost of WGS becomes affordable.

7.1.2 Low-coverage whole-genome sequencing

Another approach for reducing costs is to sequence genomes at reduced coverage, with only 250 samples required to find 99% of synonymous variants (2010). Although this technique

still allows a genome to be sequenced several times, there may be incomplete genome coverage and an expected increase in error rate (Nielsen 2010). It has been noted that marginal increase in error rate of only a few percent can equate to substantially reduced statistical power (Nielsen 2010). Computational biology approaches, such as imputation, uses the data from sequenced individuals to infer the identity of missing nucleotides in a new data set, and has been shown to increase the statistical power of GWA studies (Nielsen 2010). Imputation accuracy decreases parallel to the variant's frequency, with an error rate of 35% reported for minor alleles observed only twice in a low-coverage sample (2010). However the increased risk of pursuing a typing error variant is of concern, especially for cohorts of unrelated individuals. Family cohort based designs are superior in this regard as genotyping errors will be more transparent through precedence of Mendelian inheritance. Nevertheless, low-coverage WGS has been demonstrated as being effective for detecting rare variants influencing complex diseases (Holm et al. 2011).

7.1.3 Smaller sample sizes

The advent of sequencing approaches for the identification of rare variants has placed a limit on the number of samples that can be analysed using current technologies, due to available resources. As such, study designs that required smaller sample sizes have been favoured (Cirulli et al. 2010; Zeggini 2011). These include extreme trait design, isolated populations and cohorts of families (Cirulli et al. 2010; Zeggini 2011). Isolated populations and family studies offer not only increased allele frequencies of rare or even private variants, but also the potential for reduced phenotypic, environmental and genetic heterogeneity (Cirulli et al. 2010; Zeggini 2011).

7.1.4 Successful application of sequencing approaches

The 1000 Genomes Project uses WGS, exome sequencing and low-coverage WGS to catalog 95% of variants with frequency >1% and coding variants with frequencies as low as 0.1% (Durbin et al. 2010). An exciting recent publication by Holm et al (Holm et al. 2011) shows successful application of this dataset and imputation for a project studying the complex cardiac disorder sick sinus syndrome in the isolated Icelandic cohort. They identified a rare variant (MAF 0.004) in MYH6 that was strongly associated with disease (OR 12.5). The success of Holm et al (Holm et al. 2011) can be in part attributed to their isolated population study design and its reduced genetic heterogeneity. The variant described by Holm et al may not be present in other populations (Zeggini 2011), limiting its utility in variant specific tests of prediction. This is consistent with the low-coverage 1000 Genomes pilot that highlighted the potential for rare variants to be ethnically unique with most of the ~9 million new SNPs discovered being population specific (Via et al. 2010) with non-synonymous variants more susceptible to this trend than synonymous SNPs (2010). Gene centric as opposed to variant centric measurements may be necessary to allow utility of risk assessments across diverse population groups.

7.1.5 Modern linkage analysis

A recent study of plasma levels of the adipocytokine adiponectin levels (Bowden et al. 2010) illustrated the usefulness of reliance on linkage analysis for finding families enriched with

large effect variants. A key finding from this paper was that only a few families may contribute significantly to the originally observed linkage signal (Bowden et al. 2010). These families were examined with exome and direct sequencing to determine the causal variant for adiponectin levels, and a rare variant (MAF 0.011) identified that explained 17% of the phenotypic variation in the whole sample, and 63% in carrier families. Association methods enabled detection of this variant in other families. This success validates Blangero's argument that linkage analysis was unjustly discarded as a method for dissecting complex disease development (Blangero 2004), and that a combination of both sequencing and family linkage data will be an efficient method for the study of rare variants in common complex disease (Bowden et al. 2010; Cirulli et al. 2010).

8. Conclusion

The era of GWA studies has taught us a much about the influence of common variants on complex diseases, and revealed the influence of many previously unknown loci. As the disease mediating genes are in turn identified, it is anticipated that novel insights into the molecular pathways affecting disease risk will arise. However it has become clear that the contribution of common variants to common complex diseases has been overestimated, as substantial missing heritability for many complex phenotypes shows that only small proportion of overall phenotypic variation can be accounted for the vast majority of traits studied. Systems biology approaches combining large scale genetic and transcriptomic datasets appear key to the identification of the disease genes and the dearth of such well characterised large cohorts has no doubt contributed to the slow pace of translation.

The study of complex diseases field now looks towards rare variants and emerging next-generation sequencing technology for the next major steps forward, with family-based approaches carrying several advantages over competing designs. Until WGS becomes economically viable, affordable alternative approaches will provide the platform for advancement including exome targeted sequencing and imputation. In due course, it is anticipated that technological advances such as those under development by Pacific Biosciences will evolve to enable WGS for large sample sizes at high statistical power, allowing comprehensive characterisation of human genetic variation.

Although drug focused outcomes of genetic studies are in their infancy, there is excitement for future success (Hirschhorn 2009; Manolio et al. 2009; Cirulli et al. 2010). It is acceptable to have a proportion of heritability unexplained if the variants found can initiate development of an effective, low cost drug (Manolio et al. 2009), therefore a variant's usefulness might be based on a cost benefit ratio for treatment advancement rather than the proportion of heritability explained. For now the future of genetic analysis of complex disease looks bright as new technologies will allow great insights into the genome that were inconceivable just a decade ago.

9. References

(2007). "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls." *Nature* 447(7145): 661-678.

(2010). "A map of human genome variation from population-scale sequencing." *Nature* 467(7319): 1061-1073.

Agopian, A. J., et al. (2011). "MI-GWAS: a SAS platform for the analysis of inherited and maternal genetic effects in genome-wide association studies using log-linear models." *BMC Bioinformatics* 12: 117.

Aguilo, F., et al. (2011). "Long Noncoding RNA, Polycomb, and the Ghosts Haunting INK4b-ARF-INK4a Expression." *Cancer Res* 71(16): 5365-5369.

Alkan, C., et al. (2011). "Genome structural variation discovery and genotyping." *Nat Rev Genet* 12(5): 363-376.

Altshuler, D., et al. (2000). "The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes." *Nat Genet* 26(1): 76-80.

Anderson, C. A., et al. (2011). "Synthetic associations are unlikely to account for many common disease genome-wide association signals." *PLoS Biol* 9(1): e1000580.

Bao, X. Y., et al. (2011). "Association between Type 2 Diabetes and CDKN2A/B: a meta-analysis study." *Mol Biol Rep*.

Barrett, J. C., et al. (2009). "Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes." *Nat Genet* 41(6): 703-707.

Bell, C. G., et al. (2010). "Integrated genetic and epigenetic analysis identifies haplotype-specific methylation in the FTO type 2 diabetes and obesity susceptibility locus." *PLoS One* 5(11): e14040.

Bell, D. A., et al. (1997). "Genetic analysis of complex disease." *Science* 275(5304): 1327-1328; author reply 1329-1330.

Bell, J. T., et al. (2011). "A twin approach to unraveling epigenetics." *Trends Genet* 27(3): 116-125.

Bell, J. T., et al. (2011). "Genome-wide association scan allowing for epistasis in type 2 diabetes." *Ann Hum Genet* 75(1): 10-19.

Bernstein, B. E., et al. (2007). "The mammalian epigenome." *Cell* 128(4): 669-681.

Billings, L. K., et al. (2010). "The genetics of type 2 diabetes: what have we learned from GWAS?" *Ann N Y Acad Sci* 1212: 59-77.

Blangero, J. (2004). "Localization and identification of human quantitative trait loci: king harvest has surely come." *Curr Opin Genet Dev* 14(3): 233-240.

Blewitt, M. E., et al. (2006). "Dynamic reprogramming of DNA methylation at an epigenetically sensitive allele in mice." *PLoS Genet* 2(4): e49.

Bodmer, W., et al. (2008). "Common and rare variants in multifactorial susceptibility to common diseases." *Nat Genet* 40(6): 695-701.

Boudoukha, S., et al. (2010). "Role of the RNA-binding protein IMP-2 in muscle cell motility." *Mol Cell Biol* 30(24): 5710-5725.

Bowden, D. W., et al. (2010). "Molecular basis of a linkage peak: exome sequencing and family-based analysis identify a rare genetic variant in the ADIPOQ gene in the IRAS Family Study." *Hum Mol Genet* 19(20): 4112-4120.

Burren, O. S., et al. (2011). "T1DBase: update 2011, organization and presentation of large-scale data sets for type 1 diabetes research." *Nucleic Acids Res* 39(Database issue): D997-1001.

Carvajal-Carmona, L. G. (2010). "Challenges in the identification and use of rare disease-associated predisposition variants." *Curr Opin Genet Dev*.

Choi, M., et al. (2009). "Genetic diagnosis by whole exome capture and massively parallel DNA sequencing." *Proc Natl Acad Sci U S A* 106(45): 19096-19101.

Cirulli, E. T., et al. (2010). "Uncovering the roles of rare variants in common disease through whole-genome sequencing." *Nat Rev Genet* 11(6): 415-425.

Clayton, D. G. (2009). "Prediction and interaction in complex disease genetics: experience in type 1 diabetes." *PLoS Genet* 5(7): e1000540.

Colagiuri, R., Brown, J., and Dain, K (2010). A call to action on diabetes. I. D. Federation. Brussels.

Collins, F. S., et al. (1998). "A DNA polymorphism discovery resource for research on human genetic variation." *Genome Res* 8(12): 1229-1231.

Cooper, D. N. (2010). "Functional intronic polymorphisms: Buried treasure awaiting discovery within our genes." *Hum Genomics* 4(5): 284-288.

Craddock, N., et al. (2010). "Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls." *Nature* 464(7289): 713-720.

da Silva Xavier, G., et al. (2009). "TCF7L2 regulates late events in insulin secretion from pancreatic islet beta-cells." *Diabetes* 58(4): 894-905.

Dai, N., et al. (2011). "mTOR phosphorylates IMP2 to promote IGF2 mRNA translation by internal ribosomal entry." *Genes Dev* 25(11): 1159-1172.

Danaei, G., et al. (2011). "National, regional, and global trends in fasting plasma glucose and diabetes prevalence since 1980: systematic analysis of health examination surveys and epidemiological studies with 370 country-years and 2.7 million participants." *Lancet* 378(9785): 31-40.

de Miguel-Yanes, J. M., et al. (2011). "Genetic risk reclassification for type 2 diabetes by age below or above 50 years using 40 type 2 diabetes risk single nucleotide polymorphisms." *Diabetes Care* 34(1): 121-125.

Dickson, S. P., et al. (2010). "Rare variants create synthetic genome-wide associations." *PLoS Biol* 8(1): e1000294.

Ding, L., et al. (1992). "Comparison of the enzymatic and biochemical properties of human insulin-degrading enzyme and Escherichia coli protease III." *J Biol Chem* 267(4): 2414-2420.

Durbin, R. M., et al. (2010). "A map of human genome variation from population-scale sequencing." *Nature* 467(7319): 1061-1073.

Eichler, E. E., et al. (2010). "Missing heritability and strategies for finding the underlying causes of complex disease." *Nat Rev Genet* 11(6): 446-450.

Farber, C. R., et al. (2009). "Future of osteoporosis genetics: enhancing genome-wide association studies." *J Bone Miner Res* 24(12): 1937-1942.

Farris, W., et al. (2003). "Insulin-degrading enzyme regulates the levels of insulin, amyloid beta-protein, and the beta-amyloid precursor protein intracellular domain in vivo." *Proc Natl Acad Sci U S A* 100(7): 4162-4167.

Florez, J. C., et al. (2006). "TCF7L2 polymorphisms and progression to diabetes in the Diabetes Prevention Program." *N Engl J Med* 355(3): 241-250.

Fraga, M. F., et al. (2005). "Epigenetic differences arise during the lifetime of monozygotic twins." *Proc Natl Acad Sci U S A* 102(30): 10604-10609.

Fransen, K., et al. (2011). "The quest for genetic risk factors for Crohn's disease in the post-GWAS era." *Genome Med* 3(2): 13.

Frayling, T. M. (2007). "Genome-wide association studies provide new insights into type 2 diabetes aetiology." *Nat Rev Genet* 8(9): 657-662.

Frazer, K. A., et al. (2009). "Human genetic variation and its contribution to complex traits." *Nat Rev Genet* 10(4): 241-251.

Freedman, M. L., et al. (2011). "Principles for the post-GWAS functional characterization of cancer risk loci." *Nat Genet* 43(6): 513-518.

Gaulton, K. J., et al. (2010). "A map of open chromatin in human pancreatic islets." *Nat Genet* 42(3): 255-259.

Gloyn, A. L., et al. (2003). "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* 52(2): 568-572.

Goldstein, D. B. (2009). "Common genetic variation and human traits." *N Engl J Med* 360(17): 1696-1698.

Goring, H. H., et al. (2007). "Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes." *Nat Genet* 39(10): 1208-1216.

Grant, R. W., et al. (2009). "The clinical application of genetic testing in type 2 diabetes: a patient and physician survey." *Diabetologia* 52(11): 2299-2305.

Grant, S. F., et al. (2006). "Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes." *Nat Genet* 38(3): 320-323.

Groop, L., et al. (1996). "Metabolic consequences of a family history of NIDDM (the Botnia study): evidence for sex-specific parental effects." *Diabetes* 45(11): 1585-1593.

Hardy, J., et al. (2009). "Genomewide association studies and human disease." *N Engl J Med* 360(17): 1759-1768.

He, M., et al. (2009). "Functional SNPs in HSPA1A gene predict risk of coronary heart disease." *PLoS One* 4(3): e4851.

Heap, G. A., et al. (2010). "Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing." *Hum Mol Genet* 19(1): 122-134.

Heath, A. C., et al. (1998). "Interaction of marital status and genetic risk for symptoms of depression." *Twin Res* 1(3): 119-122.

Heijmans, B. T., et al. (2008). "Persistent epigenetic differences associated with prenatal exposure to famine in humans." *Proc Natl Acad Sci U S A* 105(44): 17046-17049.

Hemminki, K., et al. (2006). "The balance between heritable and environmental aetiology of human disease." *Nat Rev Genet* 7(12): 958-965.

Hindorff, L. A., et al. (2009). "Potential etiologic and functional implications of genome-wide association loci for human diseases and traits." *Proc Natl Acad Sci U S A* 106(23): 9362-9367.

Hirschhorn, J. N. (2009). "Genomewide association studies--illuminating biologic pathways." *N Engl J Med* 360(17): 1699-1701.

Hirschhorn, J. N., et al. (2005). "Genome-wide association studies for common diseases and complex traits." *Nat Rev Genet* 6(2): 95-108.

Holm, H., et al. (2011). "A rare variant in MYH6 is associated with high risk of sick sinus syndrome." *Nat Genet* 43(4): 316-320.

Imamura, M., et al. (2011). "Genetics of type 2 diabetes: the GWAS era and future perspectives [Review]." *Endocr J*.

Iyengar, S. K., et al. (2007). "The genetic basis of complex traits: rare variants or "common gene, common disease"?" *Methods Mol Biol* 376: 71-84.

Jia, L., et al. (2009). "Functional enhancers at the gene-poor 8q24 cancer-linked locus." *PLoS Genet* 5(8): e1000597.

Jirtle, R. L., et al. (2007). "Environmental epigenomics and disease susceptibility." *Nat Rev Genet* 8(4): 253-262.

Johansen, C. T., et al. (2010). "Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia." *Nat Genet* 42(8): 684-687.

Jowett, J. B., et al. (2010). "Genetic variation at the FTO locus influences RBL2 gene expression." *Diabetes* 59(3): 726-732.

Kendler, K. S., et al. (2000). "Tobacco consumption in Swedish twins reared apart and reared together." *Arch Gen Psychiatry* 57(9): 886-892.

Kong, A., et al. (2009). "Parental origin of sequence variants associated with complex diseases." *Nature* 462(7275): 868-874.

Kooner, J. S., et al. (2011). "Genome-wide association study in individuals of South Asian ancestry identifies six new type 2 diabetes susceptibility loci." *Nat Genet*.

Kraft, P., et al. (2009). "Genetic risk prediction--are we there yet?" *N Engl J Med* 360(17): 1701-1703.

Krishnamurthy, J., et al. (2006). "p16INK4a induces an age-dependent decline in islet regenerative potential." *Nature* 443(7110): 453-457.

Kryukov, G. V., et al. (2007). "Most rare missense alleles are deleterious in humans: implications for complex disease and association studies." *Am J Hum Genet* 80(4): 727-739.

Laird, N. M., et al. (2006). "Family-based designs in the age of large-scale gene-association studies." *Nat Rev Genet* 7(5): 385-394.

Lander, E. S. (1996). "The new genomics: global views of biology." *Science* 274(5287): 536-539.

Lasky-Su, J., et al. (2010). "On genome-wide association studies for family-based designs: an integrative analysis approach combining ascertained family samples with unselected controls." *Am J Hum Genet* 86(4): 573-580.

Le Bacquer, O., et al. (2011). "TCF7L2 splice variants have distinct effects on beta-cell turnover and function." *Hum Mol Genet* 20(10): 1906-1915.

Lehne, B., et al. (2011). "Exome localization of complex disease association signals." *BMC Genomics* 12: 92.

Liguori, A., et al. (2010). "Epigenetic changes predisposing to type 2 diabetes in intrauterine growth retardation." *Frontiers in Endocrinology* 1.

Lillioja, S., et al. (2009). "Agreement among type 2 diabetes linkage studies but a poor correlation with results from genome-wide association studies." *Diabetologia* 52(6): 1061-1074.

Lindgren, C. M., et al. (2009). "Genome-wide association scan meta-analysis identifies three Loci influencing adiposity and fat distribution." *PLoS Genet* 5(6): e1000508.

Ling, C., et al. (2009). "Epigenetics: a molecular link between environmental factors and type 2 diabetes." *Diabetes* 58(12): 2718-2725.

Lyssenko, V., et al. (2008). "Clinical risk factors, DNA variants, and the development of type 2 diabetes." *N Engl J Med* 359(21): 2220-2232.

Makowsky, R., et al. (2011). "Beyond missing heritability: prediction of complex traits." *PLoS Genet* 7(4): e1002051.

Manolio, T. A. (2009). "Cohort studies and the genetics of complex disease." *Nat Genet* 41(1): 5-6.

Manolio, T. A., et al. (2009). "The HapMap and genome-wide association studies in diagnosis and therapy." *Annu Rev Med* 60: 443-456.
Manolio, T. A., et al. (2009). "Finding the missing heritability of complex diseases." *Nature* 461(7265): 747-753.
McCarthy, M. I., et al. (2008). "Genome-wide association studies for complex traits: consensus, uncertainty and challenges." *Nat Rev Genet* 9(5): 356-369.
Misbin, R. I., et al. (1983). "Inhibition of insulin degradation by insulin-like growth factors." *Endocrinology* 113(4): 1525-1527.
Momozawa, Y., et al. (2011). "Resequencing of positional candidates identifies low frequency IL23R coding variants protecting against inflammatory bowel disease." *Nat Genet* 43(1): 43-47.
Moore, J. H. (2003). "The ubiquitous nature of epistasis in determining susceptibility to common human diseases." *Hum Hered* 56(1-3): 73-82.
Moore, J. H. (2009). "From genotypes to genometypes: putting the genome back in genome-wide association studies." *Eur J Hum Genet* 17(10): 1205-1206.
Moore, J. H., et al. (2009). "Epistasis and its implications for personal genetics." *Am J Hum Genet* 85(3): 309-320.
Murcray, C. E., et al. (2009). "Gene-environment interaction in genome-wide association studies." *Am J Epidemiol* 169(2): 219-226.
Nejentsev, S., et al. (2009). "Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes." *Science* 324(5925): 387-389.
Ng, S. B., et al. (2010). "Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome." *Nat Genet* 42(9): 790-793.
Ng, S. B., et al. (2010). "Exome sequencing identifies the cause of a mendelian disorder." *Nat Genet* 42(1): 30-35.
Ng, S. B., et al. (2009). "Targeted capture and massively parallel sequencing of 12 human exomes." *Nature* 461(7261): 272-276.
Ng, S. F., et al. (2010). "Chronic high-fat diet in fathers programs beta-cell dysfunction in female rat offspring." *Nature* 467(7318): 963-966.
Nica, A. C., et al. (2008). "Using gene expression to investigate the genetic basis of complex disorders." *Hum Mol Genet* 17(R2): R129-134.
Nielsen, R. (2010). "Genomics: In search of rare human variants." *Nature* 467(7319): 1050-1051.
Orozco, G., et al. (2010). "Synthetic associations in the context of genome-wide association scan signals." *Hum Mol Genet* 19(R2): R137-144.
Park, J. H., et al. (2010). "Estimation of effect size distribution from genome-wide association studies and implications for future discoveries." *Nat Genet* 42(7): 570-575.
Parra, E. J., et al. (2011). "Genome-wide association study of type 2 diabetes in a sample from Mexico City and a meta-analysis of a Mexican-American sample from Starr County, Texas." *Diabetologia* 54(8): 2038-2046.
Pasmant, E., et al. (2011). "ANRIL, a long, noncoding RNA, is an unexpected major hotspot in GWAS." *FASEB J* 25(2): 444-448.
Patel, C. J., et al. (2010). "An Environment-Wide Association Study (EWAS) on type 2 diabetes mellitus." *PLoS One* 5(5): e10746.
Petronis, A. (2010). "Epigenetics as a unifying principle in the aetiology of complex traits and diseases." *Nature* 465(7299): 721-727.

Pivovarova, O., et al. (2009). "Glucose inhibits the insulin-induced activation of the insulin-degrading enzyme in HepG2 cells." *Diabetologia* 52(8): 1656-1664.

Poulsen, P., et al. (1999). "Heritability of type II (non-insulin-dependent) diabetes mellitus and abnormal glucose tolerance--a population-based twin study." *Diabetologia* 42(2): 139-145.

Prokopenko, I., et al. (2008). "Type 2 diabetes: new genes, new understanding." *Trends Genet* 24(12): 613-621.

Ragvin, A., et al. (2010). "Long-range gene regulation links genomic type 2 diabetes and obesity risk regions to HHEX, SOX4, and IRX3." *Proc Natl Acad Sci U S A* 107(2): 775-780.

Rakyan, V. K., et al. (2003). "Transgenerational inheritance of epigenetic states at the murine Axin(Fu) allele occurs after maternal and paternal transmission." *Proc Natl Acad Sci U S A* 100(5): 2538-2543.

Rakyan, V. K., et al. (2011). "Epigenome-wide association studies for common human diseases." *Nat Rev Genet* 12(8): 529-541.

Rappaport, S. M., et al. (2010). "Epidemiology. Environment and disease risks." *Science* 330(6003): 460-461.

Ravelli, A. C., et al. (1998). "Glucose tolerance in adults after prenatal exposure to famine." *Lancet* 351(9097): 173-177.

Reich, D. E., et al. (2001). "On the allelic spectrum of human disease." *Trends Genet* 17(9): 502-510.

Richards, E. J. (2006). "Inherited epigenetic variation--revisiting soft inheritance." *Nat Rev Genet* 7(5): 395-401.

Rideout, W. M., 3rd, et al. (2001). "Nuclear cloning and epigenetic reprogramming of the genome." *Science* 293(5532): 1093-1098.

Risch, N., et al. (1996). "The future of genetic studies of complex human diseases." *Science* 273(5281): 1516-1517.

Rutter, G. A. (2010). "Think zinc: New roles for zinc in the control of insulin secretion." *Islets* 2(1): 49-50.

Ryan, B. M., et al. (2010). "Genetic variation in microRNA networks: the implications for cancer research." *Nat Rev Cancer* 10(6): 389-402.

Sanders, S. S. (2011). "Whole-exome sequencing: a powerful technique for identifying novel genes of complex disorders." *Clin Genet* 79(2): 132-133.

Schork, N. J., et al. (2009). "Common vs. rare allele hypotheses for complex diseases." *Curr Opin Genet Dev* 19(3): 212-219.

Shu, L., et al. (2008). "Transcription factor 7-like 2 regulates beta-cell survival and function in human pancreatic islets." *Diabetes* 57(3): 645-653.

Slatkin, M. (2009). "Epigenetic inheritance and the missing heritability problem." *Genetics* 182(3): 845-850.

So, H. C., et al. (2011). "Evaluating the heritability explained by known susceptibility variants: a survey of ten complex diseases." *Genet Epidemiol* 35(5): 310-317.

So, H. C., et al. (2011). "Evaluating the heritability explained by known susceptibility variants: a survey of ten complex diseases." *Genet Epidemiol*.

Thomas, D. C., et al. (2002). "Point: population stratification: a problem for case-control studies of candidate-gene associations?" *Cancer Epidemiol Biomarkers Prev* 11(6): 505-512.

Turkheimer, E., et al. (2003). "Socioeconomic status modifies heritability of IQ in young children." *Psychol Sci* 14(6): 623-628.

Tyler, A. L., et al. (2009). "Shadows of complexity: what biological networks reveal about epistasis and pleiotropy." *Bioessays* 31(2): 220-227.

van Hoek, M., et al. (2009). "Genetic variant in the IGF2BP2 gene may interact with fetal malnutrition to affect glucose metabolism." *Diabetes* 58(6): 1440-1444.

Via, M., et al. (2010). "The 1000 Genomes Project: new opportunities for research and social challenges." *Genome Med* 2(1): 3.

Visscher, P. M., et al. (2008). "Genome-wide association studies of quantitative traits with related individuals: little (power) lost but much to be gained." *Eur J Hum Genet* 16(3): 387-390.

Visscher, P. M., et al. (2008). "Heritability in the genomics era--concepts and misconceptions." *Nat Rev Genet* 9(4): 255-266.

Wacholder, S., et al. (2002). "Counterpoint: bias from population stratification is not a major threat to the validity of conclusions from epidemiological studies of common polymorphisms and cancer." *Cancer Epidemiol Biomarkers Prev* 11(6): 513-520.

Wang, K., et al. (2010). "Interpretation of association signals and identification of causal variants from genome-wide association studies." *Am J Hum Genet* 86(5): 730-742.

Wang, W. Y., et al. (2005). "Genome-wide association studies: theoretical and practical concerns." *Nat Rev Genet* 6(2): 109-118.

Waters, K. M., et al. (2010). "Consistent association of type 2 diabetes risk variants found in europeans in diverse racial and ethnic groups." *PLoS Genet* 6(8).

Wenzlau, J. M., et al. (2007). "The cation efflux transporter ZnT8 (Slc30A8) is a major autoantigen in human type 1 diabetes." *Proc Natl Acad Sci U S A* 104(43): 17040-17045.

Witte, J. S. (2010). "Genome-wide association studies and beyond." *Annu Rev Public Health* 31: 9-20 24 p following 20.

Wray, N. R., et al. (2011). "Synthetic associations created by rare variants do not explain most GWAS results." *PLoS Biol* 9(1): e1000579.

Yang, J., et al. (2010). "Common SNPs explain a large proportion of the heritability for human height." *Nat Genet* 42(7): 565-569.

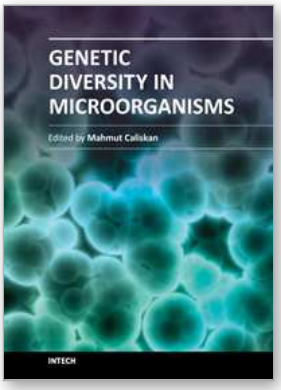
Zeggini, E. (2011). "Next-generation association studies for complex traits." *Nat Genet* 43(4): 287-288.

Zeggini, E., et al. (2008). "Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes." *Nat Genet* 40(5): 638-645.

Zhang, L., et al. (2011). "Functional SNP in the microRNA-367 binding site in the 3'UTR of the calcium channel ryanodine receptor gene 3 (RYR3) affects breast cancer risk and calcification." *Proc Natl Acad Sci U S A* 108(33): 13653-13658.

Zhang, X., et al. (2011). "Tools for efficient epistasis detection in genome-wide association study." *Source Code Biol Med* 6(1): 1.

Zhao, J., et al. (2010). "The role of height-associated loci identified in genome wide association studies in the determination of pediatric stature." *BMC Med Genet* 11: 96.



Genetic Diversity in Microorganisms

Edited by Prof. Mahmut Caliskan

ISBN 978-953-51-0064-5
Hard cover, 382 pages
Publisher InTech
Published online 24, February, 2012
Published in print edition February, 2012

Genetic Diversity in Microorganisms presents chapters revealing the magnitude of genetic diversity of microorganisms living in different environmental conditions. The complexity and diversity of microbial populations is by far the highest among all living organisms. The diversity of microbial communities and their ecologic roles are being explored in soil, water, on plants and in animals, and in extreme environments such as the arctic deep-sea vents or high saline lakes. The increasing availability of PCR-based molecular markers allows the detailed analyses and evaluation of genetic diversity in microorganisms. The purpose of the book is to provide a glimpse into the dynamic process of genetic diversity of microorganisms by presenting the thoughts of scientists who are engaged in the generation of new ideas and techniques employed for the assessment of genetic diversity, often from very different perspectives. The book should prove useful to students, researchers, and experts in the area of microbial phylogeny, genetic diversity, and molecular biology.

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821