Human Complex Trait Genetics: Lifting the Lid of the Genomics Toolbox - from Pathways to Prediction

Suzanne J. Rowe[1] and Albert Tenesa*[,1,2]

[1]*The Roslin Institute, The University of Edinburgh, Easter Bush Campus, Midlothian, EH25 9RG, Scotland, UK*

[2]*Institute of Genetics and Molecular Medicine, 4th Floor, MRC Human Genetics Unit, Western General Hospital, Crewe Road South, Edinburgh, EH4 2XU, UK*

Abstract: During the initial stages of the genome revolution human genetics was hugely successful in discovering the underlying genes for monogenic diseases. Over 3,000 monogenic diseases have been discovered with simple patterns of inheritance. The unravelling and identification of the genetic variants underlying complex or multifactorial traits, however, is proving much more elusive. There have been over 1,000 significant variants found for many quantitative and binary traits yet they explain very little of the estimated genetic variance or heritability evident from family analysis. There are many hypotheses as to why this might be the case. This apparent lack of information is holding back the clinical application of genetics and shedding doubt on whether more of the same will reveal where the remainder of the variation lies. Here we explore the current state of play, the types of variants we can detect and how they are currently exploited. Finally we look at the future challenges we must face to persuade the human genome to yield its secrets.

Received on: June 05, 2011 - Revised on: September 09, 2011 - Accepted on: October 05, 2011

Keywords: Association complex human genetics genome-wide genomics GWAS prediction.

INTRODUCTION

Complex traits are determined by the interplay of multiple genetic and environmental factors. Understanding the interaction of nature and nurture in the development of common human disease and continuous traits is the main interest of complex trait human geneticists. Unlike monogenic traits, complex trait variation is not entirely explained by one or a small number of genes but results from a complex mixture of inherited and environmental factors. The proportion of phenotypic variation explained by inherited genetic factors is known as the heritability [1], and this measures the degree of resemblance among relatives [2]. Twin and other family-based studies have shown that genetic variation explains a large proportion of the phenotypic variation observed in humans. Heritability estimates vary widely among traits, 40% of variation of most complex traits can be explained by inherited factors increasing to over 70% for some diseases such as schizophrenia [3].

Genetic variants, however, often predispose us to, rather than cause disease, and even more complications arise as they predispose us in combination with other variants and environmental influences. Penetrance is usually defined as the probability of disease given the genotype of a person, however for complex disease, it makes more sense to define penetrance as a function of total genetic load where each genetic variant is weighted according to the risk it confers (Fig. **1**). Furthermore phenotypic variation comes in many layers of biological complexity at the cellular, tissue and organism level that are likely to be involved in the pathogenesis of disease. Since disease is in itself complex, it is useful to study, and in many cases clinically treat, some intermediate disease endpoints that can be measured before the onset of the disease. For instance, high blood pressure and LDL cholesterol are treatable intermediate disease endpoints for myocardial infarction or stroke. There are multiple sources of genetic variation linked to phenotypic variation, which include multiple layers of biological complexity that lead to disease. Understanding how these are controlled by DNA alterations that are transmitted from parents to their children is one of the fundamental challenges of modern biology.

COMPLEX TRAIT MAPPING

The use of statistical methods to pinpoint the regions of the genome that harbour the DNA changes and genes that control complex traits is known as mapping. Early linkage studies exploited repetition in the genome by using cutting enzymes to identify markers that might be co-inherited with causal variants thus providing clues as to the locations in the genome affecting traits of interest. The highly polymorphic and co-dominant microsatellite markers were technically easier to work with and quickly replaced restriction fragment length polymorphisms (RFLP) in linkage studies aiming to map complex trait loci (CTL).

There are two broad approaches to CTL mapping in human family-based linkage studies. Either sampled pairs of sibs, or large sets of relatives from extended or nuclear families are analysed. The methodology relies on using marker information to measure how related individuals are at

*Address correspondence to this author at the Roslin Institute, The University of Edinburgh, Easter Bush, Roslin, Midlothian, EH25 9RG, Scotland, UK; Tel: +44 (0)131 6519226; Fax: +44 (0)131 6519105;
E-mail: albert.tenesa@ed.ac.uk

1875-5488/12 $58.00+.00

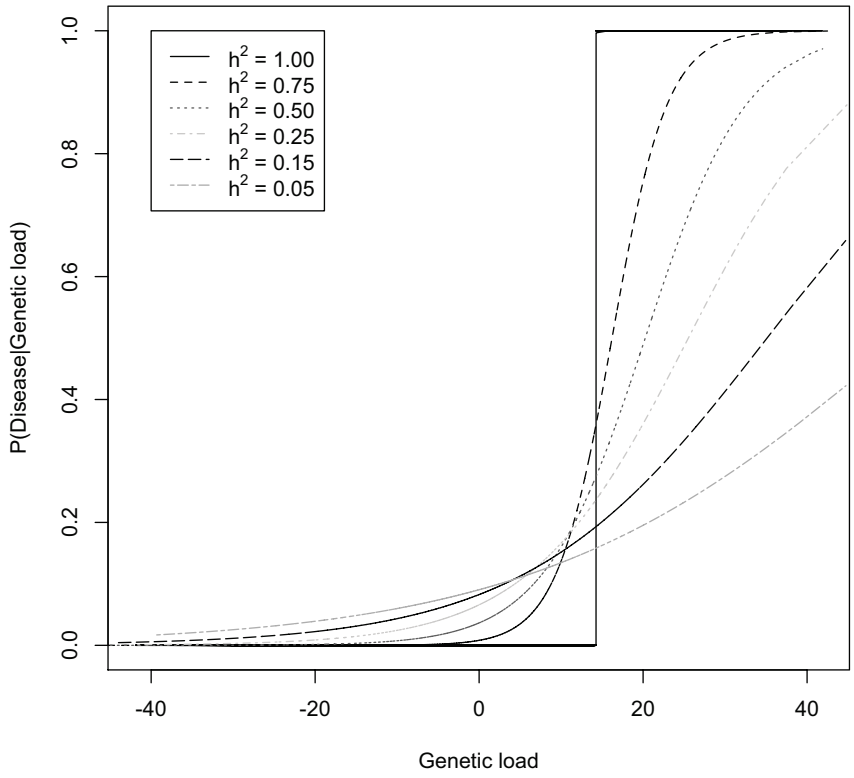

Fig. (1). Penetrance, the probability of disease given genetic load, for diseases with the same prevalence (0.1) and varying heritability (h^2).

the putative CTL. This measure of identity is termed identity by descent (IBD) where individuals share copies of a gene inherited from a common ancestor. Phenotypic similarity between two individuals is modelled as a function of their relatedness at each test position. The most widely used non-parametric method for linkage analysis of quantitative traits is sib pair analysis based on the regression method of Haseman and Elston [4]. The squared difference between the trait values for a pair of relatives is regressed against the proportion of marker alleles IBD. A negative co-efficient reflects a tendency for individuals to be more similar with respect to the trait as they share a greater proportion of the alleles IBD thus implying linkage between trait and marker. Different allelic and causal variant combinations may be seen across families, therefore, data is analysed within families and information combined across multiple families. An advantage of microsatellite markers is that they are highly polymorphic and with phased data, models can easily be extended to incorporate interactions such as dominance, epistasis and imprinting. Methods for tracking the transmission of alleles from parents dependent on affected status – and that exploit linkage disequilibrium (LD) within families and at the population level – include the transmission disequilibrium test (TDT) [5] and the family based association test (FBAT) [6] These methods have been further adapted for quantitative traits(QTDT) [7]. Family based linkage studies provided a powerful approach for identifying rare variants of large effect segregating within individual families and proved very successful for simple or monogenic traits. For complex traits, however, the approach has yielded little fruit and few of the reported findings have been successfully replicated. Lack of reproducible associations raised the question, as it does now with GWAS, of whether the approach had reached its limits. In a seminal paper, Risch and Merikangas [8] argued that modest gene effects were likely to explain the lack of reproducible linkage results and that a new approach would be required to overcome the next hurdle of human complex traits genetics.

The common disease/common variant (CD/CV) [9] hypothesis states that common complex diseases are underpinned by a large number of common variants of small effect segregating in the population, rather than a small number of rare variants of large effect. In order to fine map the CTL found in linkage studies and in particular to look for variants assumed to be common in the population, association mapping was born. The aim is to look for over representation of marker genotypes associated at a population level with a particular disease or phenotype. This necessitates a very dense marker map with adjacent markers in high LD in order to track ancestral haplotypes after many generations of recombination. It is estimated that there is at least 1 single nucleotide polymorphism (SNP) every 100-300 base pairs with greater than 1% minor allele frequency (MAF) [10] making them ideally suited as markers to tag causal variants.

The use of association for fine mapping candidate regions from linkage studies quickly gave way to more general or 'genome-wide' association studies (GWAS). One of the greatest benefits of GWAS is that it is 'agnostic' or based on no prior assumptions. Generally a simple regression analysis is used to systematically test each biallelic SNP across the genome for association with a trait or disease. Many generations of recombination creates

smaller regions of LD which with a dense-enough marker coverage provides a much higher resolution than linkage and the potential to tag common causal variants. Care must be taken to ensure that the associated loci are not spurious associations due to, for example, population substructure or admixture [11]. The use of hundreds of thousands of markers also necessitates very strict significance criteria making it difficult to detect all but the largest effects.

To facilitate association mapping the HapMap project was developed, the second phase of which identified more than 3.1 million SNPs, from 270 individuals from 4 populations [10]. Phase 3 offered a further 1.6 million SNPs and expansion to a total of 11 populations and 1184 individuals [12]. The abundance of these SNPs means that most common SNPs are in high LD with neighbouring SNPs with an average minimum r^2 squared value of between 0.9 and 0.96 depending on population providing excellent coverage of common variation across the genome. The reference panels of the Hapmap project are routinely used to statistically estimate (i.e. impute) genotypes within marker intervals of sparser data sets [13-16]. This aids the meta-analysis of multiple cohorts genotyped with different genotyping arrays.

SUCCESS FROM GENOME-WIDE ASSOCIATION STUDIES

The initial maelstrom of GWAS began in 2005/2006. Fig. (**2**) shows that since 2006 the number of published SNPs exceeding genome-wide significance from GWAS has risen linearly. The National human genome Research Institute (NGHRi) database of published results contains 5118 entries to date affecting over 500 traits (http://www.genome.gov/gwastudies) [17]. Fig. (**3**) shows the top 30 diseases with the greatest number of entries. There have been more than 90 cancer susceptibility loci identified [18], over 180 loci for height [19], 39 for type 2 diabetes [20] and 71 for Crohn's disease [21]. With commercial arrays it is now commonplace for published studies to involve analyses using over 500,000 SNPs.

Despite these successes much of the heritability or genetic variance estimated to exist remains unaccounted for [22-24] (Table **1**). Complex traits are often associated with high heritability yet there is mounting empirical evidence from GWAS results that there are few common variants of large effect. Fig. (**2**) shows the median odds ratio for the NHGRi database is only 1.1. Height is often used as an example as although 180 loci have been found above the genome wide significance level, these variants only explain

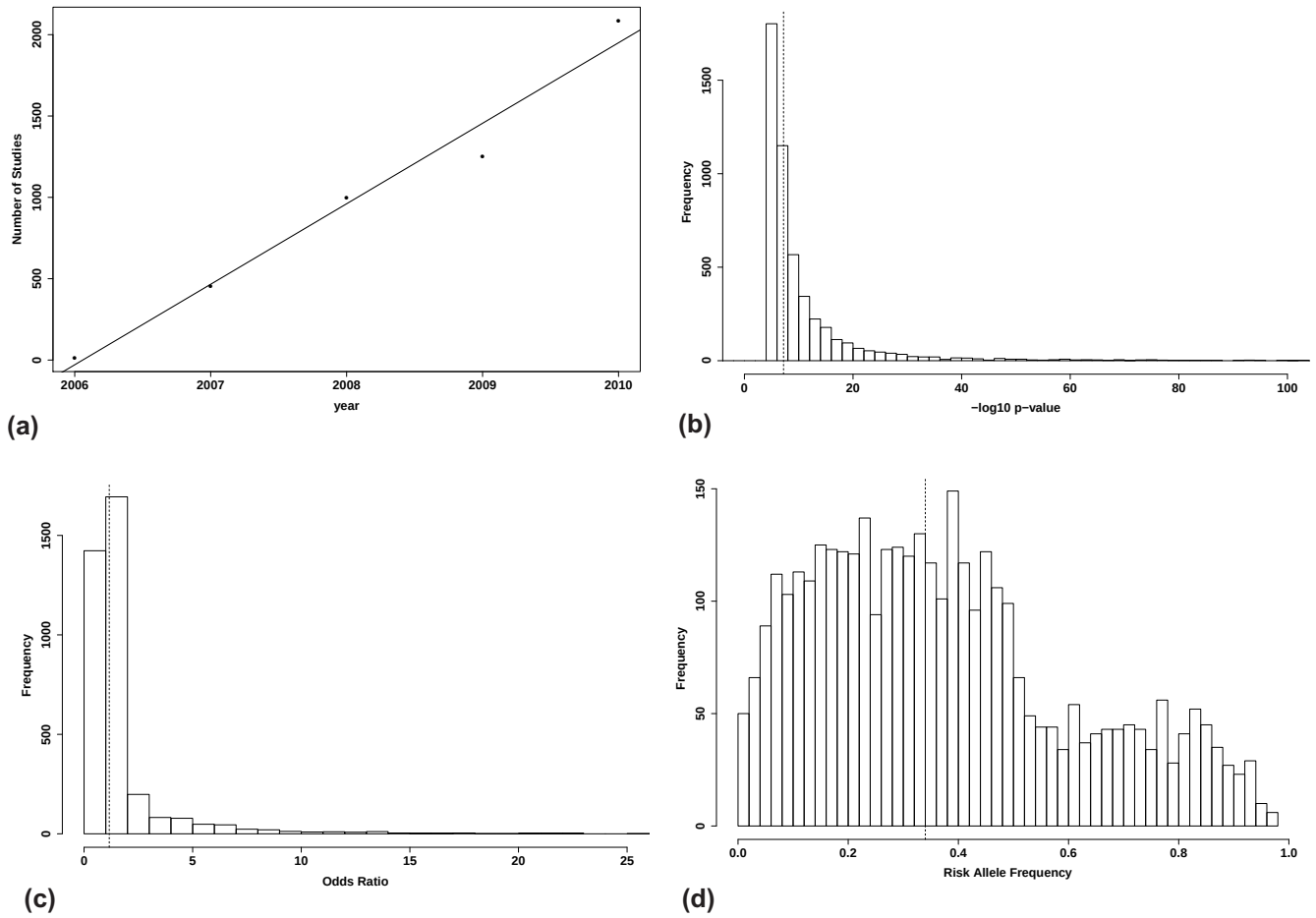

Fig. (2). Summary statistics for entries added to NGHRi database since 2005 http://www.genome.gov/gwastudies. **a**) shows the number of studies added per year (Regression r^2=0.97), **b**) shows the distribution of reported $-\log_{10}$ P-values – median is 7.2, **c**) shows the distribution of reported effects (e.g. odds ratio or beta) – median is 1.1, **d**) is histogram of risk allele frequencies – median is 0.34.

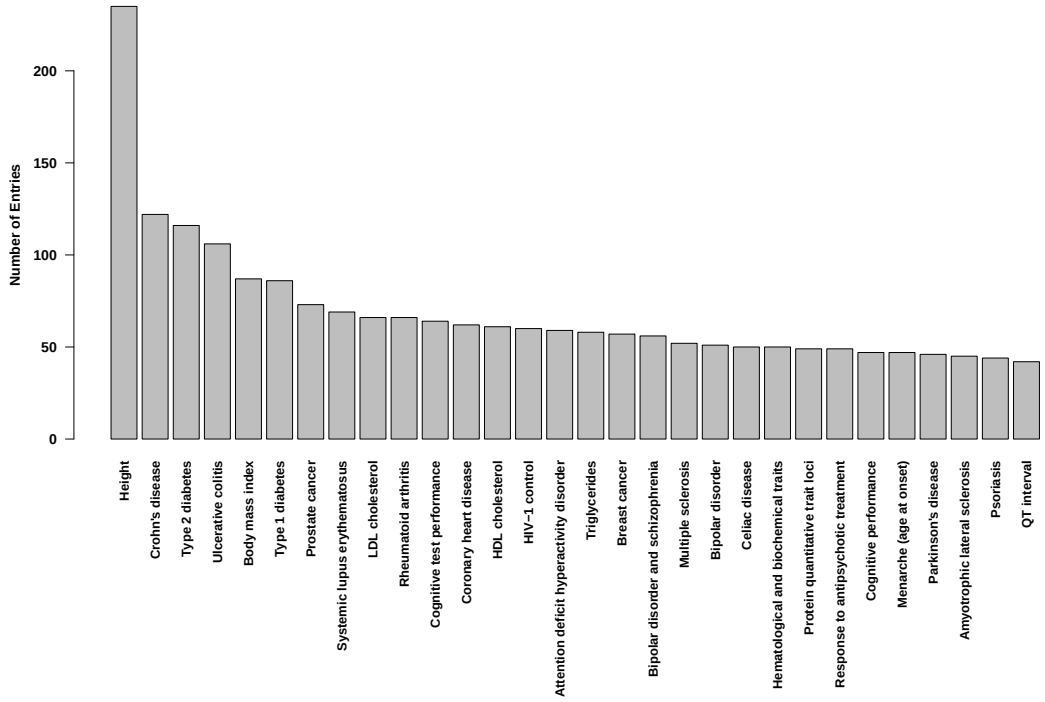

Fig. (3). Number of entries in NHGRi database for top 30 traits from 2005-2011.

Table 1. Examples of Number of Discovered Loci and Percentage of Heritability (h^2) Explained [22]

Disease/Trait	Number of discovered loci	% of h^2 explained	h^2
Type 1 diabetes	41	60	0.6
Fetal Heamaglobin levels	3	50	0.6
Macular degeneration	3	50	0.7
Type 2 diabetes	39	20-25	0.38
Crohn's disease	71	20-25	0.5
LDL and HDL levels	95	20-25	0.3
Height	180	12	0.8

*Narrow sense heritability of trait from published literature [22].

around 12% of the genetic variance [19]. Studies rarely seem to find variants explaining more than 1% of sibling recurrence risk. There are of course exceptions such as large variants found for Crohn's disease with an odds ratio of 3.99 [25] and activated partial thromboplastin time where three common variants explain 18% of the phenotypic variance [26]. It is increasingly likely, however, that most large common variants have now been discovered, therefore, the crucial question is how to capture the remaining variation - often dubbed the missing heritability or the dark matter of the genome. Furthermore, if we have already discovered the low hanging fruit, are there diminishing returns to be expected from further GWAS?

Although a proportion of the missing heritability may potentially be due to inflation of estimates of additive variance due to other non-linear sources of variation such as epistasis, epigenetics and gene by environment interaction [2,27], this is, in general, not supported by theoretical and empirical data [28]. Even if epistasis was widespread, its detection would be challenging due to the number of tests involved, lack of power and the difficulty of setting appropriate significance thresholds. These difficulties are reflected in the little evidence available from large genome-wide association studies to suggest interlocus interactions. It is highly probable that there are a large number of loci with effects too small to achieve significance by the stringent thresholds set. It is also possible that the remaining variants are rare and therefore poorly tagged by current arrays, which use common tag SNPs unlikely to be in sufficient LD with rarer variants. More powerful methodology can capture this hidden genetic variation by using all SNP information regardless of significance thus avoiding the problem of stringent significant thresholds set in GWAS. Yang *et al*. [22] used ~295K SNPs on ~4K individuals and explained 10-fold more variation of height than previously reported (i.e. the ~295K SNPs explained about 45% of variance).

Furthermore if the incomplete LD caused by the differences between the distributions of minor allele frequency for SNPs and causal variants is accounted for, 80% of the variation of height can be explained in line with literature estimates of narrow sense heritability. The approach of Yang *et al*. [29] can be readily extended to partition the genetic variance for designated regions (often per chromosomes) by using a linear mixed model to fit all SNPs in that region simultaneously. This provides a mechanism for locating regions of importance containing markers or groups of markers which may not meet significant thresholds on an individual basis.

Single marker based association models can also be extended further to incorporate combinations of markers or haplotypes. If the distribution of CTL differs from that of markers, the utilisation of haplotype information could capture associations with rarer variants eluded by single SNP analysis because allele frequencies at the CTL and the 'virtual' marker created by the haplotype will be better matched [30,31]. Combinations of multiple SNPs or haplotypes could potentially capture greater proportions of genetic variation than single SNPs [32].

A further explanation for the lack of genetic variation captured by GWAS is that the allelic architecture underlying complex traits is not described accurately by the CD/CV hypothesis. The common disease rare variant hypothesis (CD/RV) states that there are many low frequency variants of large effect segregating in the population and that each phenotype is due to combined effect of a few of these variants [33-35]. Under mutation selection balance theory the expectation would be an inverse correlation between deleterious SNPs and minor allele frequency with a probable upper limit of 1% applying to deleterious alleles [36]. Recent evidence for the (CD/RV) theory includes a rare variant in MHY6 associated with sick sinus syndrome or slow heart rate [37]. The risk allele has a frequency in the Icelandic population of 0.38% but has an associated odds ratio of 12.53. Lifetime risk in the population is 6%, however for carriers of the risk allele is 50%. Interestingly common variants of the gene modulate cardiac conduction.

The CD/RV theory also supports variation due to widespread allelic heterogeneity. There is increasing evidence to suggest that multiple independent signals within a locus exist for a number of complex traits. There are a large number of disease causing allelic variants in some known genes such as BRCA1 and MLH1. Haiman *et al*. [38] found multiple independent regions associated with prostate cancer within the 8q24 locus and Lango Allen *et al*. [19] found that out of 180 loci significant for height at least 19 loci had multiple independently associated variants. Overall, this suggests that previously discovered loci are strong candidates for harbouring further missing genetic variation.

The underlying distribution of effects in complex diseases may have a huge impact on the application of information discovered to date. One of the greatest hopes of GWAS was that it could be used for the detection of disease related CTL. Furthermore by identifying the genes involved there was much hope for the prediction and prevention of disease alongside new potential drug targets or therapeutics. If we were able to accurately estimate the effects of sufficient loci to explain half of the known genetic variance then genomic profiles for most common diseases would achieve sufficient discriminative ability to be of clinical validity. Even if accurately estimated loci explained only one quarter of genetic variance, for rare diseases (i.e. low prevalence) the genomic profile would be a more useful predictor of risk than self-reported family history [39]. These profiles can be used from birth enabling susceptible individuals to avoid environmental exposure to risk, thereby reducing absolute risk of disease.

The development of risk predictors for complex diseases has been slower than anticipated due to the small number of loci identified by current GWAS. For complex traits individual variants rarely explain enough variation to be utilised as risk predictors, however profiles based on many of these variants could potentially be used [40-42]. There is a mounting body of evidence showing that whole genome prediction methods developed over decades to estimate livestock breeding values offer the opportunity to increase the accuracy of prediction of disease risk [43-45]. Challenges include how to select and estimate the predictors of the model that minimises the mean square error of prediction of the phenotype. The genetic architecture of the trait determines the best strategy for model selection and the accuracy of the prediction model. Issues for model selection and expected accuracy include determining whether models that fit a subset of the available SNPs will perform better than models that fit all available SNPs simultaneously, how sensitive both approaches are to the misspecification of the genetic architecture of the trait [46,47], or the best strategy to shrink the estimates of the effects to prevent over-fitting of the models [48,49]. Furthermore in populations with high LD many loci will be correlated with each other affecting model choice and assumptions about prior distribution. Some of these issues are reviewed by Daetwyler *et al.* who give deterministic formulae for assessing the accuracy of genomic prediction [50].

The area under the receiver operator characteristic (ROC) curve (or its equivalent C-index) can be used to assess the discriminative ability of a prediction model [51] and has been used to assess the performance of genetic predictors and genomic profiling [52,53]. This is a technique for visualising, organising and selecting classifiers based on their performance often employed by the medical decision making community for diagnostic testing [54]. The performance of a diagnostic classifier over a range of thresholds can be examined to identify the threshold at which the classifier is most accurate.

For disease prediction, the ROC curve represents the trade off between true positive rate (sensitivity) and false positive rate (1-specificity) and the area under the receiver operator characteristic curve (AUC) the probability that for a randomly selected case and control, the case will be ranked higher by the prediction model than the control. This is equivalent to the Mann–Whitney–Wilcoxon test statistic [55]. ROC curves are a useful measure as they are not affected by the skewness of the data i.e. they are not affected by the proportion of cases and controls (other than sampling error) which might vary from one data set to the next [51]. Wray *et al*., [56] give parameter estimates for 17 common

complex diseases. They show that it is theoretically possible for a genomic profile for complex disease to exceed the threshold of discriminative ability of 0.75 that could, arguably, be considered clinically useful. They also show an AUC of 0.75 can imply anything from 0.1 to 0.74 of the genetic variance explained thus care should be taken in the interpretation of this statistic without some knowledge of the parameters used.

To explore this further we simulated data for prostate, breast and colorectal cancer to investigate the effect of various parameters on the prediction of risk. Ten thousand cases and 10,000 controls were simulated under a liability threshold model. The distribution of allele frequencies was taken from a beta distribution and the additive genetic effect sizes from a normal distribution. The data was randomly assigned to two equal sized groups, a training and a validation set. Three scenarios were examined a) the use of prediction or AUC when effects of all loci are known (i.e. we used the simulated effects), b) the use of prediction when the effects of all loci are estimated, and c) prediction using the 15 most significant loci. Results showed that, in particular, for the prediction of prostate cancer the discriminative ability across all ages (0.85) was higher than for the over 65's (0.79). This is counter intuitive for a disease primarily of late onset but can be explained by the fact that the prevalence across all ages is lower, therefore those that have genetic risk factors leading to clinical diagnosis are likely to be at the extremes of the distribution in the population making the probability of discriminating accurately higher. Accuracy of prediction is very dependent on the underlying genetic architecture of the trait. Fig. (**4**) shows the discriminative ability for breast cancer (BC), colorectal cancer (CRC) and prostate cancer (PC) cancers either across all ages or in the over 65s comparing models with 500 or 10,000 underlying additive loci. When 10,000 loci were simulated the discriminative ability when using the 15 most significant effects ranged from 0.53-0.57. Whilst these figures are low and do not appear to offer a hopeful prognosis for clinical utility, recent results for 32 loci associated with BMI [57] found that although AUC for prediction of obesity was only 0.57 this was still an increase from using family and environmental information alone which only yielded an accuracy of 0.51. When we simulated a prevalence of 0.004, a heritability of 0.42 and 500 loci underlying the trait, the discriminative ability for a model which estimated the effects of all loci was 0.93 which was the equivalent of using the known simulated values for the loci. Accuracy using the top 15 most significant estimated effects was 0.75. In general the discriminative ability of the prediction models using the estimates of genetic effects are as good as using the actual simulated values. The maximum AUC of 0.92, 0.87, and 0.91, are similar to estimates from Wray *et al*. [56] who estimated 0.90, 0.89 and 0.96 for Prostate. Breast, and Colon cancer respectively using a threshold liability model.

Park *et al*. [58] used summary statistics from existing GWAS to calculate the expected distribution and the number of loci that exist within the range of SNP effects observed on a trait by trait basis. They estimate discoveries for future GWAS for given sample sizes by integrating power over the number of unidentified loci that probably exist whilst accounting for the distribution of relative risk and allele frequency. Based on the assumption of a spectrum of low-penetrance common variants the predicted total number of loci within the range of effects currently detected by GWAS for height, Crohn's disease and BPC (breast, prostate, colorectal) cancers are 201, 142 and 67 explaining 16.4, 20 and 17.1% of the genotypic variance respectively. They use these predictions to estimate the AUC for Crohn's and the cancers. They predict that all 142 loci would give an AUC of 79.2% for Crohn's in comparison to 72.8% from the 30 loci discovered to date. An AUC for breast cancer given the 5-10

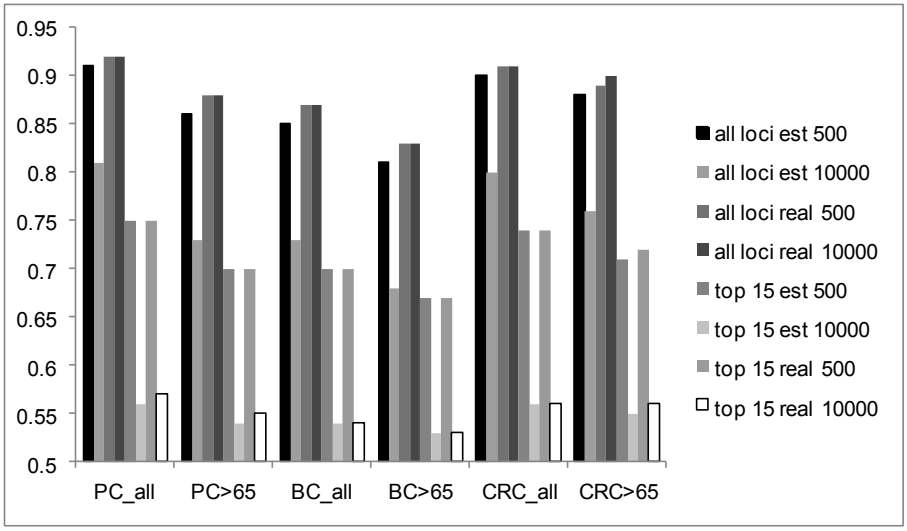

Fig. (4). Descriminative ability (AUC) for prostate (PC), breast (BC) and colorectal cancer (CRC) for all ages and for over 65 years of age. Either 500 or 10,000 loci were simulated with estimates or real/actual values for all loci or top 15 significant results used in prediction. Prevalence used for PC, BC and CRC for (all ages, >65) were (0.007,0.04),(0.007,0.025) and (0.004,0.01) respectively and heritabilities were 0.42(PC),0.27(BC),and 0.35(CRC). Prevalences were extracted from the Information Services Division from the National Health Service Scotland (http://www.isdscotland.org/index.asp).

loci that exist of 57% could be improved to 63.5% if all 67 loci were discovered. These results are based on modest estimates of heritability and MAF from published studies generally > 0.05 and do not reflect results from our simulation studies [59] or empirical data from Yang *et al*. [60].

A recent paper by Meuwissen *et al*., [44] takes genomic profiling a step further by exploring the prospect of prediction from whole genome sequencing data. They use simulation analyses to explore the accuracy of genomic prediction using sequence data which has the advantage of using all polymorphisms such as indels as well as SNPs. Furthermore the sequence data contains the causal variants. They conclude that if there are a finite number of causal variants a Bayesian approach is most successful, however should the distribution of effects follow an infinitesimal model with thousands of loci of small effects it is expected that BLUP methods will outperform the Bayesian analysis. Statistical methods for whole genome prediction are also reviewed by de los Campos *et al*. [43]. Ultimately, however, the success of prediction methods for complex diseases will be limited by the disease prevalence and the heritability. Even if the predictor explains 100% of the genetic variance, the maximum AUC is dependent on trait heritability ad prevalence. This is shown in Fig. (**2**) where even if the simulated values for all loci are used the AUC ranges from 0.87-0.93. Care must also be taken to note that genomic profiling infers genetic risk and not absolute risk. These profiles are a predictor of genotypic value rather than phenotype although they can be extended further to incorporate environmental risk factors such as smoking, diet or exercise. Furthermore, the discriminative ability of a model in a case-control study is different to discriminative ability at the population level and a model that is well calibrated for a case-control study may well yield a poor performance when screening the whole population. Strengths and weaknesses are further reviewed by Hand [55].

The advent of whole genome sequencing methods [61] provides the opportunity of increasing the information available for genetic mapping studies by the inclusion of all sources of genetic variation (SNPs, CNVs, indels, rearrangements, etc.) that may be causal and allows the unbiased screening of the genetic variation present in the sampled individuals. Of the 3200Mb of DNA in the human genome only 1.5% or ~50Mb is functional or coding DNA comprising approximately 20-25,000 genes. Large numbers of repetitive elements make up approximately 50% of the DNA. These include indels, copy number variants (CNVs), translocations, inversions, and chromosomal duplications. In a recent review of 75 cancer genes with germline mutations, 28 were reported to be mutated by genomic deletion or duplication [18].

The latest commercial arrays contain a combination of SNPs and structural variants [62]. CNVs have been associated with many traits [63] including starch digestion (AMY1) [64], HIV [62], Schizophrenia [65] and Crohn's disease [66], although in European populations, most common CNVs are likely to have been tagged with SNPs. High heritability, high mutation rates and greater variation in African populations still provide compelling arguments for sequencing thousands more genomes and using CNVs to complement SNPs.

Following the Hapmap project the thousand genomes project [10] http://www.1000genomes.org promises to deliver individual sequence variation by sequencing 1000 individuals. This will give insights into the variation in structural variation such as CNV's, indels and deletions and into regulatory elements. The genomes of 2500 individuals from 25 populations around the world will be sequenced using next-generation sequencing technologies [61]. This international collaboration is set to produce an unprecedented public catalog of human genetic variation, including SNPs and structural variants, and their haplotype contexts to support genome-wide association studies and other medical research studies. The use of this valuable resource as a reference panel for imputation of cohorts with genome-wide association data may help to screen for rarer genetic variants. However, it is yet unclear whether rarer genetic variants will be accurately imputed using general reference panels. It may be more useful that each cohort generates its own reference panel by either sequencing or genotyping a subset of it for a dense SNP array. Even in the latter case, it is likely that the imputation quality of rare variants in low LD with common tagging SNP will be low, making re-sequencing of large numbers of samples necessary [67].

Intermediate Phenotypes

It is becoming clear that SNP trait associations alone rarely lead to identifying the causal variant or the context in which the gene operates. This biological context is a necessary step for the generation of new biological hypotheses and the identification of drug targets for disease. There are many intermediate or endo-phenotypes which can be used to map complex variation. It is possible that these could be more heritable and represent a more comprehensive approach in the quest for the underlying causal variants of quantitative traits.

The most widely studied intermediate phenotype is the study of expression analysis or the abundance of mRNA transcripts using microarrays or RNASeq to look at which genes are expressed at a given timepoint in a given tissue. Expression QTL may be categorised into cis (local) or trans (distant) effects. There appears to be a current bias towards large cis effects inferring regulatory processes are involved [68,69]. Associating patterns of gene expression with genotypic and phenotypic values facilitates insights into biological function. Genes which are differentially expressed can be used to infer regulatory networks and underlying pathways associated with traits or diseases [70,71].

Regulatory networks and discovered pathways in turn provide another dimension to GWAS and can themselves be used as intermediate phenotypes. Pathway enrichment analysis involves assigning SNPs to genes and subsequent pathways in order to find associations at the pathway level providing greater insight into biological function. Enrichment scores for known pathways can be obtained to investigate whether there is over representation of genes in any one pathway associated with phenotype [72-75]. The 180 loci discovered for height [19] are enriched for genes

that are connected in biological pathways and underlie skeletal growth defects. Pathway analysis increasingly shows that pathways reputedly underlying common diseases are often common across diseases. It is difficult to ascertain whether this is due to annotation bias or whether there is genuine pleiotropy across the mechanisms underlying these diseases. It could be hypothesised some level of pleiotropy must exist given that there are approximately 21,000 genes and millions of traits. A recent study of coronary artery disease found that 5 out of 13 loci showed strong association with various other diseases or traits [76]. Interleukin receptor genes are linked with several common diseases such as Crohn's, lupus, and rheumatoid arthritis implicating common underlying mechanisms or pathways [75]. The prediction of biological mechanisms seems at best tenuous with many potential biases not only from limitations of annotation but from setting of significance thresholds, assignment of variant to gene or pathway, algorithm or method used to ascertain enrichment scores, and fundamentally the type of biological data selected [77,78]. It is possible that in using a statistical approach we are much closer to the accurate prediction of phenotype using molecular data, which is almost a black box whole genome approach, than we are to gaining real insights into the specific underlying biological mechanisms which remain rather more elusive for most traits.

Finally, an important source of heritable variation not seen in coding sequence is epigenetic modification. This includes promoter methylation, histone tail modifications and altered expression of non-coding RNAs that associate with chromatin modifying complexes. These modifications contribute to gene regulation in normal development, particularly in foetal growth, to gene expression in tumorigenesis and have been shown to mediate the influence of environment on gene expression. Technologies are now sufficiently advanced to carry out methylation profiling. Gibbs *et al*., [79] find abundant QTL for DNA CpG methylation across the genome in brain tissue. Methylation studies are likely to play a wider role in the post genomic era. There are arguments, yet to be proven, for epigenetic variation as a driving force for development, evolutionary adaptation, and disease [80].

CONCLUSIONS

Any GWAS is limited by depth of genomic coverage on commercial arrays. It remains to be seen whether results from current GWAS describe the full spectrum of existing genetic variation or merely what we have the ability to detect. It is highly probable that the latter is the case. The current status is that there are many variants detected but few explain more than 1% of trait variance and most genetic variance estimated by family studies is yet to be explained by allelic variation discovered from GWAS. Given extrapolation from current results it is likely that current studies are underpowered to detect all but the loci that explain the largest proportion of variance of complex traits. A common misconception is that rare variants of large effect will be easier to identify than common variants of small effect. Power depends on the proportion of variance explained by the SNP [31] and this is a function of the allele frequency and the effect. For instance, identified rare variants from GWAS (MAF 1-5%) have a mean OR ~3.74 [81] which, for disease prevalence ranging from 1-20% explain between 0.5-8% of the variance in the liability scale [82]. Identifying variants explaining this proportion of a quantitative trait variance at a genome-wide significance threshold of $5x10^{-8}$ with 80% power would require between 7900 and 460 samples, respectively. Finding a sufficient number of cases for some diseases in a single study or even at all may prove difficult and could be a limiting factor for future GWAS and whole-genome sequencing studies. As it is unlikely that the number of individuals needed to have enough power to detect these variants will come from a single study, the use of meta-analyses is and will be necessary to discover missing variants. Care must be taken that individuals are from the same population or that structure/admixture is properly identified and accounted for. It is important to note that whilst it is unlikely that increasing the SNP density of common variants will help to reveal new variants in European populations, there could still be much potential in populations of African descent where LD decays faster over distance and there is much more standing genetic variation.

New methods that exploit current GWA data are needed to tag rare variants which may be present in the population at lower minor allele frequencies than current SNP panels. Haplotype analysis may hold some promise if sufficiently large sample sizes are available. Poorly tagged structural and sequence variation may underlie some of the missing genetic variation and could explain allelic heterogeneity among common SNPs previously identified.

Whole genome methods are increasingly likely to be used for the estimation of heritability and the clinical prediction of susceptibility to complex disease given the mounting body of evidence that traits are affected by hundreds if not thousands of variants. The main advantages of these methods are that all loci can be used simultaneously removing bias from setting stringent arbitrary thresholds to control for false positive rates. These whole genome methods assume additivity and there is currently little evidence to suggest otherwise, however, there may be gene-gene and gene-environment interactions yet to be uncovered. It is yet to be shown how accurate predictors will be within and across populations.

Whole genome sequencing and the thousand genomes project offer unprecedented opportunities to tag many more variants. The likely next steps before moving to large-scale whole-genome sequencing will be exome sequencing and whole-genome sequencing of a small number of samples from the population under study that will be used as the reference panel for imputation. It remains to be seen whether and under what circumstances this will provide any advantage to the use of publicly available reference sets such as the thousand genomes project.

It is likely that the future of genomic mapping will involve incorporating methods such as re-sequencing and association analysis of all variants within a region of LD affecting a trait. To further elucidate the molecular and biological mechanisms involved it is likely that this will need to be followed by genomic analysis of gene expression and methylation in relevant human tissue and screens for somatic mutations on risk haplotypes. These steps are likely

to be necessary to develop reliable biomarkers for therapeutics and pharmacogenetics.

In the quest to solve cryptic human complex variation, researchers in human genomics have access to a greater toolbox than ever before. Increases in sample size and meta-analyses together with a new catalogue of markers will increase power to detect rare variants. Further, efficacious whole genome sequencing and the ability to explore beyond the DNA level mean the post GWAS era promises to be one of unprecedented discovery.

ACKNOWLEDGEMENTS

This work was funded by Cancer Research UK, grant C12229/A13154 and by the BBSRC *via* the Roslin Institute's Institute Strategic Programme Grant.

REFERENCES

[1] Falconer, D. S.; T.F.C.Mackay *Introduction to Quantitative Genetics;* Fourth Edition ed.; 4. Longmans Green, Harlow, Essex, UK: 1996.

[2] Visscher, P. M.; Hill, W. G.; Wray, N. R. Heritability in the genomics era--concepts and misconceptions. *Nat. Rev. Genet.,* **2008,** *9* (4), 255-266.

[3] Goldstein, D. B.; Cavalleri, G. L. Genomics: understanding human diversity. *Nature,* **2005,** *437* (7063), 1241-1242.

[4] Haseman, J. K.; Elston, R. C. The investigation of linkage between a quantitative trait and a marker locus. *Behav. Genet.,* **1972,** *2* (1), 3-19.

[5] Spielman, R. S.; Ewens, W. J. The TDT and other family-based tests for linkage disequilibrium and association. *Am. J. Hum. Genet.,* **1996,** *59* (5), 983-989.

[6] Lunetta, K. L.; Faraone, S. V.; Biederman, J.; Laird, N. M. Family-based tests of association and linkage that use unaffected sibs, covariates, and interactions. *Am J Hum. Genet.,* **2000,** *66* (2), 605-614.

[7] Allison, D. B. Transmission-disequilibrium tests for quantitative traits. *Am. J. Hum. Genet.,* **1997,** *60* (3), 676-690.

[8] Risch, N.; Merikangas, K. The future of genetic studies of complex human diseases. *Science,* **1996,** *273* (5281), 1516-1517.

[9] Reich, D. E.; Lander, E. S. On the allelic spectrum of human disease. *Trends Genet.,* **2001,** *17* (9), 502-510.

[10] Durbin, R. M.; Abecasis, G. R.; Altshuler, D. L.; Auton, A.; Brooks, L. D.; Durbin, R. M.; Gibbs, R. A.; Hurles, M. E.; McVean, G. A. A map of human genome variation from population-scale sequencing. *Nature,* **2010,** *467* (7319), 1061-1073.

[11] Price, A. L.; Patterson, N. J.; Plenge, R. M.; Weinblatt, M. E.; Shadick, N. A.; Reich, D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.,* **2006,** *38* (8), 904-909.

[12] Altshuler, D. M.; Gibbs, R. A.; Peltonen, L.; Altshuler, D. M.; Gibbs, R. A.; Peltonen, L.; Dermitzakis, E.; Schaffner, S. F.; Yu, F.; Peltonen, L.; Dermitzakis, E.; Bonnen, P. E.; Altshuler, D. M.; Gibbs, R. A.; de Bakker, P. I.; Deloukas, P.; Gabriel, S. B.; Gwilliam, R.; Hunt, S.; Inouye, M.; Jia, X.; Palotie, A.; Parkin, M.; Whittaker, P.; Yu, F.; Chang, K.; Hawes, A.; Lewis, L. R.; Ren, Y.; Wheeler, D.; Gibbs, R. A.; Muzny, D. M.; Barnes, C.; Darvishi, K.; Hurles, M.; Korn, J. M.; Kristiansson, K.; Lee, C.; McCarrol, S. A.; Nemesh, J.; Dermitzakis, E.; Keinan, A.; Montgomery, S. B.; Pollack, S.; Price, A. L.; Soranzo, N.; Bonnen, P. E.; Gibbs, R. A.; Gonzaga-Jauregui, C.; Keinan, A.; Price, A. L.; Yu, F.; Anttila, V.; Brodeur, W.; Daly, M. J.; Leslie, S.; McVean, G.; Moutsianas, L.; Nguyen, H.; Schaffner, S. F.; Zhang, Q.; Ghori, M. J.; McGinnis, R.; McLaren, W.; Pollack, S.; Price, A. L.; Schaffner, S. F.; Takeuchi, F.; Grossman, S. R.; Shlyakhter, I.; Hostetter, E. B.; Sabeti, P. C.; Adebamowo, C. A.; Foster, M. W.; Gordon, D. R.; Licinio, J.; Manca, M. C.; Marshall, P. A.; Matsuda, I.; Ngare, D.; Wang, V. O.; Reddy, D.; Rotimi, C. N.; Royal, C. D.; Sharp, R. R.; Zeng, C.; Brooks, L. D.; McEwen, J. E. Integrating common and rare genetic variation in diverse human populations. *Nature,* **2010,** *467* (7311), 52-58.

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

[14] Marchini, J.; Howie, B. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.,* **2010,** *11* (7), 499-511.

[15] Li, Y.; Willer, C.; Sanna, S.; Abecasis, G. Genotype imputation. *Annu. Rev. Genomics Hum. Genet.,* **2009,** *10*, 387-406.

[16] Li, Y.; Willer, C. J.; Ding, J.; Scheet, P.; Abecasis, G. R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.,* **2010,** *34* (8), 816-834.

[17] Hindorff, L. A.; Sethupathy, P.; Junkins, H. A.; Ramos, E. M.; Mehta, J. P.; Collins, F. S.; Manolio, T. A. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci U. S. A,* **2009,** *106* (23), 9362-9367.

[18] Fletcher, O.; Houlston, R. S. Architecture of inherited susceptibility to common cancer. *Nat. Rev Cancer,* **2010,** *10* (5), 353-361.

[19] Lango Allen, H.; Estrada, K.; Lettre, G.; Berndt, S. I.; Weedon, M. N.; Rivadeneira, F.; Willer, C. J.; Jackson, A. U.; Vedantam, S.; Raychaudhuri, S.; Ferreira, T.; Wood, A. R.; Weyant, R. J.; Segre, A. V.; Speliotes, E. K.; Wheeler, E.; Soranzo, N.; Park, J. H.; Yang, J.; Gudbjartsson, D.; Heard-Costa, N. L.; Randall, J. C.; Qi, L.; Vernon Smith, A.; Magi, R.; Pastinen, T.; Liang, L.; Heid, I. M.; Luan, J.; Thorleifsson, G.; Winkler, T. W.; Goddard, M. E.; Sin Lo, K.; Palmer, C.; Workalemahu, T.; Aulchenko, Y. S.; Johansson, A.; Carola Zillikens, M.; Feitosa, M. F.; Esko, T.; Johnson, T.; Ketkar, S.; Kraft, P.; Mangino, M.; Prokopenko, I.; Absher, D.; Albrecht, E.; Ernst, F.; Glazer, N. L.; Hayward, C.; Hottenga, J. J.; Jacobs, K. B.; Knowles, J. W.; Kutalik, Z.; Monda, K. L.; Polasek, O.; Preuss, M.; Rayner, N. W.; Robertson, N. R.; Steinthorsdottir, V.; Tyrer, J. P.; Voight, B. F.; Wiklund, F.; Xu, J.; Hua Zhao, J.; Nyholt, D. R.; Pellikka, N.; Perola, M.; Perry, J. R. B.; Surakka, I.; Tammesoo, M. L.; Altmaier, E. L.; Amin, N.; Aspelund, T.; Bhangale, T.; Boucher, G.; Chasman, D. I.; Chen, C.; Coin, L.; Cooper, M. N.; Dixon, A. L.; Gibson, Q.; Grundberg, E.; Hao, K.; Juhani Junttila, M.; Kaplan, L. M.; Kettunen, J.; Konig, I. R.; Kwan, T.; Lawrence, R. W.; Levinson, D. F.; Lorentzon, M.; McKnight, B.; Morris, A. P.; Muller, M.; Suh Ngwa, J.; Purcell, S.; Rafelt, S.; Salem, R. M.; Salvi, E.; Sanna, S.; Shi, J.; Sovio, U.; Thompson, J. R.; Turchin, M. C.; Vandenput, L.; Verlaan, D. J.; Vitart, V.; White, C. C.; Ziegler, A.; Almgren, P.; Balmforth, A. J.; Campbell, H.; Citterio, L.; De Grandi, A.; Dominiczak, A.; Duan, J.; Elliott, P.; Elosua, R.; Eriksson, J. G.; Freimer, N. B.; Geus, E. J. C.; Glorioso, N.; Haiqing, S.; Hartikainen, A. L.; Havulinna, A. S.; Hicks, A. A.; Hui, J.; Igl, W.; Illig, T.; Jula, A.; Kajantie, E.; Kilpelainen, T. O.; Koiranen, M.; Kolcic, I.; Koskinen, S.; Kovacs, P.; Laitinen, J.; Liu, J.; Lokki, M. L.; Marusic, A.; Maschio, A.; Meitinger, T.; Mulas, A.; Pare, G.; Parker, A. N.; Peden, J. F.; Petersmann, A.; Pichler, I.; Pietilainen, K. H.; Pouta, A.; Ridderstrale, M.; Rotter, J. I.; Sambrook, J. G.; Sanders, A. R.; Oliver Schmidt, C.; Sinisalo, J.; Smit, J. H.; Stringham, H. M.; Bragi Walters, G.; Widen, E.; Wild, S. H.; Willemsen, G.; Zagato, L.; Zgaga, L.; Zitting, P.; Alavere, H.; Farrall, M.; McArdle, W. L.; Nelis, M.; Peters, M. J.; Ripatti, S.; van Meurs, J. B. J.; Aben, K. K.; Ardlie, K. G.; Beckmann, J. S.; Beilby, J. P.; Bergman, R. N.; Bergmann, S.; Collins, F. S.; Cusi, D.; den Heijer, M.; Eiriksdottir, G.; Gejman, P. V.; Hall, A. S.; Hamsten, A.; Huikuri, H. V.; Iribarren, C.; Kahonen, M.; Kaprio, J.; Kathiresan, S.; Kiemeney, L.; Kocher, T.; Launer, L. J.; Lehtimaki, T.; Melander, O.; Mosley Jr, T. H.; Musk, A. W.; Nieminen, M. S.; 'Donnell, C. J.; Ohlsson, C.; Oostra, B.; Palmer, L. J.; Raitakari, O.; Ridker, P. M.; Rioux, J. D.; Rissanen, A.; Rivolta, C.; Schunkert, H.; Shuldiner, A. R.; Siscovick, D. S.; Stumvoll, M.; Tonjes, A.; Tuomilehto, J.; van Ommen, G. J.; Viikari, J.; Heath, A. C.; Martin, N. G.; Montgomery, G. W.; Province, M. A.; Kayser, M.; Arnold, A. M.; Atwood, L. D.; Boerwinkle, E.; Chanock, S. J.; Deloukas, P.; Gieger, C.; Gronberg, H.; Hall, P.; Hattersley, A. T.; Hengstenberg, C.; Hoffman, W.; Mark Lathrop, G.; Salomaa, V.; Schreiber, S.; Uda, M.; Waterworth, D.; Wright, A. F.; Assimes, T. L.; Barroso, I.; Hofman, A.; Mohlke, K. L.; Boomsma, D. I.; Caulfield, M. J.; drienne Cupples, L.; Erdmann, J.; Fox, C. S.; Gudnason, V.; Gyllensten, U.; Harris, T. B.; Hayes, R. B.; Jarvelin, M. R.; Mooser, V.; Munroe, P. B.; Ouwehand, W. H. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature,* **2010,** *467* (7317), 832-838.

[20] Voight, B. F.; Scott, L. J.; Steinthorsdottir, V.; Morris, A. P.; Dina, C.; Welch, R. P.; Zeggini, E.; Huth, C.; Aulchenko, Y. S.; Thorleifsson, G.; McCulloch, L. J.; Ferreira, T.; Grallert, H.; Amin, N.; Wu, G.; Willer, C. J.; Raychaudhuri, S.; McCarroll, S. A.; Langenberg, C.; Hofmann, O. M.; Dupuis, J.; Qi, L.; Segre, A. V.; van, H. M.; Navarro, P.; Ardlie, K.; Balkau, B.; Benediktsson, R.; Bennett, A. J.; Blagieva, R.; Boerwinkle, E.; Bonnycastle, L. L.; Bengtsson, B. K.; Bravenboer, B.; Bumpstead, S.; Burtt, N. P.; Charpentier, G.; Chines, P. S.; Cornelis, M.; Couper, D. J.; Crawford, G.; Doney, A. S.; Elliott, K. S.; Elliott, A. L.; Erdos, M. R.; Fox, C. S.; Franklin, C. S.; Ganser, M.; Gieger, C.; Grarup, N.; Green, T.; Griffin, S.; Groves, C. J.; Guiducci, C.; Hadjadj, S.; Hassanali, N.; Herder, C.; Isomaa, B.; Jackson, A. U.; Johnson, P. R.; Jorgensen, T.; Kao, W. H.; Klopp, N.; Kong, A.; Kraft, P.; Kuusisto, J.; Lauritzen, T.; Li, M.; Lieverse, A.; Lindgren, C. M.; Lyssenko, V.; Marre, M.; Meitinger, T.; Midthjell, K.; Morken, M. A.; Narisu, N.; Nilsson, P.; Owen, K. R.; Payne, F.; Perry, J. R.; Petersen, A. K.; Platou, C.; Proenca, C.; Prokopenko, I.; Rathmann, W.; Rayner, N. W.; Robertson, N. R.; Rocheleau, G.; Roden, M.; Sampson, M. J.; Saxena, R.; Shields, B. M.; Shrader, P.; Sigurdsson, G.; Sparso, T.; Strassburger, K.; Stringham, H. M.; Sun, Q.; Swift, A. J.; Thorand, B.; Tichet, J.; Tuomi, T.; van Dam, R. M.; van Haeften, T. W.; van, H. T.; van Vliet-Ostaptchouk, J. V.; Walters, G. B.; Weedon, M. N.; Wijmenga, C.; Witteman, J.; Bergman, R. N.; Cauchi, S.; Collins, F. S.; Gloyn, A. L.; Gyllensten, U.; Hansen, T.; Hide, W. A.; Hitman, G. A.; Hofman, A.; Hunter, D. J.; Hveem, K.; Laakso, M.; Mohlke, K. L.; Morris, A. D.; Palmer, C. N.; Pramstaller, P. P.; Rudan, I.; Sijbrands, E.; Stein, L. D.; Tuomilehto, J.; Uitterlinden, A.; Walker, M.; Wareham, N. J.; Watanabe, R. M.; Abecasis, G. R.; Boehm, B. O.; Campbell, H.; Daly, M. J.; Hattersley, A. T.; Hu, F. B.; Meigs, J. B.; Pankow, J. S.; Pedersen, O.; Wichmann, H. E.; Barroso, I.; Florez, J. C.; Frayling, T. M.; Groop, L.; Sladek, R.; Thorsteinsdottir, U.; Wilson, J. F.; Illig, T.; Froguel, P.; van Duijn, C. M.; Stefansson, K.; Altshuler, D.; Boehnke, M.; McCarthy, M. I. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat. Genet.,* **2010,** *42* (7), 579-589.

[21] Franke, A.; McGovern, D. P.; Barrett, J. C.; Wang, K.; Radford-Smith, G. L.; Ahmad, T.; Lees, C. W.; Balschun, T.; Lee, J.; Roberts, R.; Anderson, C. A.; Bis, J. C.; Bumpstead, S.; Ellinghaus, D.; Festen, E. M.; Georges, M.; Green, T.; Haritunians, T.; Jostins, L.; Latiano, A.; Mathew, C. G.; Montgomery, G. W.; Prescott, N. J.; Raychaudhuri, S.; Rotter, J. I.; Schumm, P.; Sharma, Y.; Simms, L. A.; Taylor, K. D.; Whiteman, D.; Wijmenga, C.; Baldassano, R. N.; Barclay, M.; Bayless, T. M.; Brand, S.; Buning, C.; Cohen, A.; Colombel, J. F.; Cottone, M.; Stronati, L.; Denson, T.; De, V. M.; D'Inca, R.; Dubinsky, M.; Edwards, C.; Florin, T.; Franchimont, D.; Gearry, R.; Glas, J.; Van, G. A.; Guthery, S. L.; Halfvarson, J.; Verspaget, H. W.; Hugot, J. P.; Karban, A.; Laukens, D.; Lawrance, I.; Lemann, M.; Levine, A.; Libioulle, C.; Louis, E.; Mowat, C.; Newman, W.; Panes, J.; Phillips, A.; Proctor, D. D.; Regueiro, M.; Russell, R.; Rutgeerts, P.; Sanderson, J.; Sans, M.; Seibold, F.; Steinhart, A. H.; Stokkers, P. C.; Torkvist, L.; Kullak-Ublick, G.; Wilson, D.; Walters, T.; Targan, S. R.; Brant, S. R.; Rioux, J. D.; D'Amato, M.; Weersma, R. K.; Kugathasan, S.; Griffiths, A. M.; Mansfield, J. C.; Vermeire, S.; Duerr, R. H.; Silverberg, M. S.; Satsangi, J.; Schreiber, S.; Cho, J. H.; Annese, V.; Hakonarson, H.; Daly, M. J.; Parkes, M. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat. Genet.,* **2010,** *42* (12), 1118-1125.

[22] Lander, E. S. Initial impact of the sequencing of the human genome. *Nature,* **2011,** *470* (7333), 187-197.

[23] Maher, B. Personal genomes: The case of the missing heritability. *Nature,* **2008,** *456* (7218), 18-21.

[24] Manolio, T. A.; Collins, F. S.; Cox, N. J.; Goldstein, D. B.; Hindorff, L. A.; Hunter, D. J.; McCarthy, M. I.; Ramos, E. M.; Cardon, L. R.; Chakravarti, A.; Cho, J. H.; Guttmacher, A. E.; Kong, A.; Kruglyak, L.; Mardis, E.; Rotimi, C. N.; Slatkin, M.; Valle, D.; Whittemore, A. S.; Boehnke, M.; Clark, A. G.; Eichler, E. E.; Gibson, G.; Haines, J. L.; Mackay, T. F.; McCarroll, S. A.; Visscher, P. M. Finding the missing heritability of complex diseases. *Nature,* **2009,** *461* (7265), 747-753.

[25] Barrett, J. C.; Hansoul, S.; Nicolae, D. L.; Cho, J. H.; Duerr, R. H.; Rioux, J. D.; Brant, S. R.; Silverberg, M. S.; Taylor, K. D.; Barmada, M. M.; Bitton, A.; Dassopoulos, T.; Datta, L. W.; Green, T.; Griffiths, A. M.; Kistner, E. O.; Murtha, M. T.; Regueiro, M. D.; Rotter, J. I.; Schumm, L. P.; Steinhart, A. H.; Targan, S. R.; Xavier, R. J.; Libioulle, C.; Sandor, C.; Lathrop, M.; Belaiche, J.; Dewit, O.; Gut, I.; Heath, S.; Laukens, D.; Mni, M.; Rutgeerts, P.; Van, G. A.; Zelenika, D.; Franchimont, D.; Hugot, J. P.; de, V. M.; Vermeire, S.; Louis, E.; Cardon, L. R.; Anderson, C. A.; Drummond, H.; Nimmo, E.; Ahmad, T.; Prescott, N. J.; Onnie, C. M.; Fisher, S. A.; Marchini, J.; Ghori, J.; Bumpstead, S.; Gwilliam, R.; Tremelling, M.; Deloukas, P.; Mansfield, J.; Jewell, D.; Satsangi, J.; Mathew, C. G.; Parkes, M.; Georges, M.; Daly, M. J. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat. Genet.,* **2008,** *40* (8), 955-962.

[26] Houlihan, L. M.; Davies, G.; Tenesa, A.; Harris, S. E.; Luciano, M.; Gow, A. J.; McGhee, K. A.; Liewald, D. C.; Porteous, D. J.; Starr, J. M.; Lowe, G. D.; Visscher, P. M.; Deary, I. J. Common variants of large effect in F12, KNG1, and HRG are associated with activated partial thromboplastin time. *Am. J. Hum. Genet.,* **2010,** *86* (4), 626-631.

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

[28] Hill, W. G.; Goddard, M. E.; Visscher, P. M. Data and theory point to mainly additive genetic variance for complex traits. *PLoS. Genet.,* **2008,** *4* (2), e1000008.

[29] Yang, J.; Manolio, T. A.; Pasquale, L. R.; Boerwinkle, E.; Caporaso, N.; Cunningham, J. M.; de, A. M.; Feenstra, B.; Feingold, E.; Hayes, M. G.; Hill, W. G.; Landi, M. T.; Alonso, A.; Lettre, G.; Lin, P.; Ling, H.; Lowe, W.; Mathias, R. A.; Melbye, M.; Pugh, E.; Cornelis, M. C.; Weir, B. S.; Goddard, M. E.; Visscher, P. M. Genome partitioning of genetic variation for complex traits using common SNPs. *Nat. Genet,* **2011,** *43* (6), 519-525.

[30] Tenesa, A.; Knott, S. A.; Carothers, A. D.; Visscher, P. M. Power of linkage disequilibrium mapping to detect a quantitative trait locus (QTL) in selected samples of unrelated individuals. *Ann. Hum. Genet.,* **2003,** *67* (Pt 6), 557-566.

[31] Tenesa, A.; Visscher, P. M.; Carothers, A. D.; Knott, S. A. Mapping quantitative trait loci using linkage disequilibrium: marker- versus trait-based methods. *Behav. Genet,* **2005,** *35* (2), 219-228.

[32] Wessel, J.; Schork, N. J. Generalized genomic distance-based regression methodology for multilocus association analysis. *Am J Hum. Genet.,* **2006,** *79* (5), 792-806.

[33] Stranger, B. E.; Stahl, E. A.; Raj, T. Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics,* **2011,** *187* (2), 367-383.

[34] Pritchard, J. K.; Cox, N. J. The allelic architecture of human disease genes: common disease-common variant...or not? *Hum. Mol. Genet.,* **2002,** *11* (20), 2417-2423.

[35] Pritchard, J. K. Are rare variants responsible for susceptibility to complex diseases? *Am. J. Hum. Genet.,* **2001,** *69* (1), 124-137.

[36] Bodmer, W.; Bonilla, C. Common and rare variants in multifactorial susceptibility to common diseases. *Nat. Genet.,* **2008,** *40* (6), 695-701.

[37] Holm, H.; Gudbjartsson, D. F.; Sulem, P.; Masson, G.; Helgadottir, H. T.; Zanon, C.; Magnusson, O. T.; Helgason, A.; Saemundsdottir, J.; Gylfason, A.; Stefansdottir, H.; Gretarsdottir, S.; Matthiasson, S. E.; Thorgeirsson, G. M.; Jonasdottir, A.; Sigurdsson, A.; Stefansson, H.; Werge, T.; Rafnar, T.; Kiemeney, L. A.; Parvez, B.; Muhammad, R.; Roden, D. M.; Darbar, D.; Thorleifsson, G.; Walters, G. B.; Kong, A.; Thorsteinsdottir, U.; Arnar, D. O.; Stefansson, K. A rare variant in MYH6 is associated with high risk of sick sinus syndrome. *Nat. Genet.,* **2011,** *43* (4), 316-320.

[38] Haiman, C. A.; Patterson, N.; Freedman, M. L.; Myers, S. R.; Pike, M. C.; Waliszewska, A.; Neubauer, J.; Tandon, A.; Schirmer, C.; McDonald, G. J.; Greenway, S. C.; Stram, D. O.; Le, M. L.; Kolonel, L. N.; Frasco, M.; Wong, D.; Pooler, L. C.; Ardlie, K.; Oakley-Girvan, I.; Whittemore, A. S.; Cooney, K. A.; John, E. M.; Ingles, S. A.; Altshuler, D.; Henderson, B. E.; Reich, D. Multiple regions within 8q24 independently affect risk for prostate cancer. *Nat. Genet.,* **2007,** *39* (5), 638-644.

[39] Wray, N. R.; Goddard, M. E.; Visscher, P. M. Prediction of individual genetic risk of complex disease. *Curr. Opin. Genet. Dev.,* **2008,** *18* (3), 257-263.

[40] Ashley, E. A.; Butte, A. J.; Wheeler, M. T.; Chen, R.; Klein, T. E.; Dewey, F. E.; Dudley, J. T.; Ormond, K. E.; Pavlovic, A.; Morgan, A. A.; Pushkarev, D.; Neff, N. F.; Hudgins, L.; Gong, L.; Hodges,

L. M.; Berlin, D. S.; Thorn, C. F.; Sangkuhl, K.; Hebert, J. M.; Woon, M.; Sagreiya, H.; Whaley, R.; Knowles, J. W.; Chou, M. F.; Thakuria, J. V.; Rosenbaum, A. M.; Zaranek, A. W.; Church, G. M.; Greely, H. T.; Quake, S. R.; Altman, R. B. Clinical assessment incorporating a personal genome. *Lancet,* **2010,** *375* (9725), 1525-1535.

[41] Spencer, K. L.; Olson, L. M.; Schnetz-Boutaud, N.; Gallins, P.; Agarwal, A.; Iannaccone, A.; Kritchevsky, S. B.; Garcia, M.; Nalls, M. A.; Newman, A. B.; Scott, W. K.; Pericak-Vance, M. A.; Haines, J. L. Using genetic variation and environmental risk factor data to identify individuals at high risk for age-related macular degeneration. *PLoS. One.,* **2011,** *6* (3), e17784.

[42] Wei, Z.; Wang, K.; Qu, H. Q.; Zhang, H.; Bradfield, J.; Kim, C.; Frackleton, E.; Hou, C.; Glessner, J. T.; Chiavacci, R.; Stanley, C.; Monos, D.; Grant, S. F.; Polychronakos, C.; Hakonarson, H. From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes. *PLoS. Genet.,* **2009,** *5* (10), e1000678.

[43] de los Campos, G.; Gianola, D.; Allison, D. B. Predicting genetic predisposition in humans: the promise of whole-genome markers. *Nat. Rev. Genet.,* **2010,** *11* (12), 880-886.

[44] Meuwissen, T.; Goddard, M. Accurate prediction of genetic values for complex traits by whole-genome resequencing. *Genetics,* **2010,** *185* (2), 623-631.

[45] Goddard, M.; Wray, N. R.; Verbyla, K.; Visscher, P. M. Estimating effects and making predictions from genome-wide marker data. *Statistical Science,* **2009,** *4*, 517-529.

[46] Meuwissen, T. H.; Hayes, B. J.; Goddard, M. E. Prediction of total genetic value using genome-wide dense marker maps. *Genetics,* **2001,** *157* (4), 1819-1829.

[47] Meuwissen, T. Genomic selection: marker assisted selection on a genome wide scale. *J Anim Breed. Genet.,* **2007,** *124* (6), 321-322.

[48] Tibshirani, R. Regression shrinkage and selection *via* the Lasso. *Journal of the Royal Statistical Society, Series B,* **1996,** *58*, 267-288.

[49] Gianola, D.; de los, C. G.; Hill, W. G.; Manfredi, E.; Fernando, R. Additive genetic variability and the Bayesian alphabet. *Genetics,* **2009,** *183* (1), 347-363.

[50] Daetwyler, H. D.; Villanueva, B.; Woolliams, J. A. Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PLoS. One.,* **2008,** *3* (10), e3395.

[51] Fawcett, T. An introduction to ROC analysis. *Pattern Recognition Letters,* **2006,** *27* (8), 861-874.

[52] Wacholder, S.; Hartge, P.; Prentice, R.; Garcia-Closas, M.; Feigelson, H. S.; Diver, W. R.; Thun, M. J.; Cox, D. G.; Hankinson, S. E.; Kraft, P.; Rosner, B.; Berg, C. D.; Brinton, L. A.; Lissowska, J.; Sherman, M. E.; Chlebowski, R.; Kooperberg, C.; Jackson, R. D.; Buckman, D. W.; Hui, P.; Pfeiffer, R.; Jacobs, K. B.; Thomas, G. D.; Hoover, R. N.; Gail, M. H.; Chanock, S. J.; Hunter, D. J. Performance of common genetic variants in breast-cancer risk models. *N. Engl. J. Med.,* **2010,** *362* (11), 986-993.

[53] Meigs, J. B.; Shrader, P.; Sullivan, L. M.; McAteer, J. B.; Fox, C. S.; Dupuis, J.; Manning, A. K.; Florez, J. C.; Wilson, P. W.; D'Agostino, R. B., Sr.; Cupples, L. A. Genotype score in addition to common risk factors for prediction of type 2 diabetes. *N. Engl. J. Med.,* **2008,** *359* (21), 2208-2219.

[54] Zweig, M. H.; Campbell, G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin. Chem.,* **1993,** *39* (4), 561-577.

[55] Hand, D. J. Evaluating diagnostic tests: The area under the ROC curve and the balance of errors. *Stat. Med.,* **2010,** *29* (14), 1502-1510.

[56] Wray, N. R.; Yang, J.; Goddard, M. E.; Visscher, P. M. The genetic interpretation of area under the ROC curve in genomic profiling. *PLoS. Genet.,* **2010,** *6* (2), e1000864.

[57] Speliotes, E. K.; Willer, C. J.; Berndt, S. I.; Monda, K. L.; Thorleifsson, G.; Jackson, A. U.; Allen, H. L.; Lindgren, C. M.; Luan, J.; Magi, R.; Randall, J. C.; Vedantam, S.; Winkler, T. W.; Qi, L.; Workalemahu, T.; Heid, I. M.; Steinthorsdottir, V.; Stringham, H. M.; Weedon, M. N.; Wheeler, E.; Wood, A. R.; Ferreira, T.; Weyant, R. J.; Segre, A. V.; Estrada, K.; Liang, L.; Nemesh, J.; Park, J. H.; Gustafsson, S.; Kilpelainen, T. O.; Yang, J.; Bouatia-Naji, N.; Esko, T.; Feitosa, M. F.; Kutalik, Z.; Mangino, M.; Raychaudhuri, S.; Scherag, A.; Smith, A. V.; Welch, R.; Zhao, J. H.; Aben, K. K.; Absher, D. M.; Amin, N.; Dixon, A. L.; Fisher, E.; Glazer, N. L.; Goddard, M. E.; Heard-Costa, N. L.; Hoesel, V.; Hottenga, J. J.; Johansson, A.; Johnson, T.; Ketkar, S.; Lamina, C.; Li, S.; Moffatt, M. F.; Myers, R. H.; Narisu, N.; Perry, J. R.; Peters, M. J.; Preuss, M.; Ripatti, S.; Rivadeneira, F.; Sandholt, C.; Scott, L. J.; Timpson, N. J.; Tyrer, J. P.; van, W. S.; Watanabe, R. M.; White, C. C.; Wiklund, F.; Barlassina, C.; Chasman, D. I.; Cooper, M. N.; Jansson, J. O.; Lawrence, R. W.; Pellikka, N.; Prokopenko, I.; Shi, J.; Thiering, E.; Alavere, H.; Alibrandi, M. T.; Almgren, P.; Arnold, A. M.; Aspelund, T.; Atwood, L. D.; Balkau, B.; Balmforth, A. J.; Bennett, A. J.; Ben-Shlomo, Y.; Bergman, R. N.; Bergmann, S.; Biebermann, H.; Blakemore, A. I.; Boes, T.; Bonnycastle, L. L.; Bornstein, S. R.; Brown, M. J.; Buchanan, T. A.; Busonero, F.; Campbell, H.; Cappuccio, F. P.; Cavalcanti-Proenca, C.; Chen, Y. D.; Chen, C. M.; Chines, P. S.; Clarke, R.; Coin, L.; Connell, J.; Day, I. N.; Heijer, M.; Duan, J.; Ebrahim, S.; Elliott, P.; Elosua, R.; Eiriksdottir, G.; Erdos, M. R.; Eriksson, J. G.; Facheris, M. F.; Felix, S. B.; Fischer-Posovszky, P.; Folsom, A. R.; Friedrich, N.; Freimer, N. B.; Fu, M.; Gaget, S.; Gejman, P. V.; Geus, E. J.; Gieger, C.; Gjesing, A. P.; Goel, A.; Goyette, P.; Grallert, H.; Grassler, J.; Greenawalt, D. M.; Groves, C. J.; Gudnason, V.; Guiducci, C.; Hartikainen, A. L.; Hassanali, N.; Hall, A. S.; Havulinna, A. S.; Hayward, C.; Heath, A. C.; Hengstenberg, C.; Hicks, A. A.; Hinney, A.; Hofman, A.; Homuth, G.; Hui, J.; Igl, W.; Iribarren, C.; Isomaa, B.; Jacobs, K. B.; Jarick, I.; Jewell, E.; John, U.; Jorgensen, T.; Jousilahti, P.; Jula, A.; Kaakinen, M.; Kajantie, E.; Kaplan, L. M.; Kathiresan, S.; Kettunen, J.; Kinnunen, L.; Knowles, J. W.; Kolcic, I.; Konig, I. R.; Koskinen, S.; Kovacs, P.; Kuusisto, J.; Kraft, P.; Kvaloy, K.; Laitinen, J.; Lantieri, O.; Lanzani, C.; Launer, L. J.; Lecoeur, C.; Lehtimaki, T.; Lettre, G.; Liu, J.; Lokki, M. L.; Lorentzon, M.; Luben, R. N.; Ludwig, B.; Manunta, P.; Marek, D.; Marre, M.; Martin, N. G.; McArdle, W. L.; McCarthy, A.; McKnight, B.; Meitinger, T.; Melander, O.; Meyre, D.; Midthjell, K.; Montgomery, G. W.; Morken, M. A.; Morris, A. P.; Mulic, R.; Ngwa, J. S.; Nelis, M.; Neville, M. J.; Nyholt, D. R.; O'Donnell, C. J.; O'Rahilly, S.; Ong, K. K.; Oostra, B.; Pare, G.; Parker, A. N.; Perola, M.; Pichler, I.; Pietilainen, K. H.; Platou, C. G.; Polasek, O.; Pouta, A.; Rafelt, S.; Raitakari, O.; Rayner, N. W.; Ridderstrale, M.; Rief, W.; Ruokonen, A.; Robertson, N. R.; Rzehak, P.; Salomaa, V.; Sanders, A. R.; Sandhu, M. S.; Sanna, S.; Saramies, J.; Savolainen, M. J.; Scherag, S.; Schipf, S.; Schreiber, S.; Schunkert, H.; Silander, K.; Sinisalo, J.; Siscovick, D. S.; Smit, J. H.; Soranzo, N.; Sovio, U.; Stephens, J.; Surakka, I.; Swift, A. J.; Tammesoo, M. L.; Tardif, J. C.; Teder-Laving, M.; Teslovich, T. M.; Thompson, J. R.; Thomson, B.; Tonjes, A.; Tuomi, T.; van Meurs, J. B.; van Ommen, G. J. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat. Genet.,* **2010,** *42* (11), 937-948.

[58] Park, J. H.; Wacholder, S.; Gail, M. H.; Peters, U.; Jacobs, K. B.; Chanock, S. J.; Chatterjee, N. Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nat. Genet.,* **2010,** *42* (7), 570-575.

[59] Tenesa, A.; Dunlop, M. G. New insights into the aetiology of colorectal cancer from genome-wide association studies. *Nat. Rev. Genet.,* **2009,** *10* (6), 353-358.

[60] Yang, J.; Weedon, M. N.; Purcell, S.; Lettre, G.; Estrada, K.; Willer, C. J.; Smith, A. V.; Ingelsson, E.; O'Connell, J. R.; Mangino, M.; Magi, R.; Madden, P. A.; Heath, A. C.; Nyholt, D. R.; Martin, N. G.; Montgomery, G. W.; Frayling, T. M.; Hirschhorn, J. N.; McCarthy, M. I.; Goddard, M. E.; Visscher, P. M. Genomic inflation factors under polygenic inheritance. *Eur J Hum. Genet.,* **2011**.

[61] Bateman, A.; Quackenbush, J. Bioinformatics for next generation sequencing. *Bioinformatics,* **2009,** *25* (4), 429.

[62] McCarroll, S. A. Extending genome-wide association studies to copy-number variation. *Hum. Mol. Genet.,* **2008,** *17* (R2), R135-R142.

[63] Craddock, N.; Hurles, M. E.; Cardin, N.; Pearson, R. D.; Plagnol, V.; Robson, S.; Vukcevic, D.; Barnes, C.; Conrad, D. F.; Giannoulatou, E.; Holmes, C.; Marchini, J. L.; Stirrups, K.; Tobin, M. D.; Wain, L. V.; Yau, C.; Aerts, J.; Ahmad, T.; Andrews, T. D.; Arbury, H.; Attwood, A.; Auton, A.; Ball, S. G.; Balmforth, A. J.; Barrett, J. C.; Barroso, I.; Barton, A.; Bennett, A. J.; Bhaskar, S.; Blaszczyk, K.; Bowes, J.; Brand, O. J.; Braund, P. S.; Bredin, F.; Breen, G.; Brown, M. J.; Bruce, I. N.; Bull, J.; Burren, O. S.; Burton, J.; Byrnes, J.; Caesar, S.; Clee, C. M.; Coffey, A. J.; Connell, J. M.; Cooper, J. D.; Dominiczak, A. F.; Downes, K.;

Drummond, H. E.; Dudakia, D.; Dunham, A.; Ebbs, B.; Eccles, D.; Edkins, S.; Edwards, C.; Elliot, A.; Emery, P.; Evans, D. M.; Evans, G.; Eyre, S.; Farmer, A.; Ferrier, I. N.; Feuk, L.; Fitzgerald, T.; Flynn, E.; Forbes, A.; Forty, L.; Franklyn, J. A.; Freathy, R. M.; Gibbs, P.; Gilbert, P.; Gokumen, O.; Gordon-Smith, K.; Gray, E.; Green, E.; Groves, C. J.; Grozeva, D.; Gwilliam, R.; Hall, A.; Hammond, N.; Hardy, M.; Harrison, P.; Hassanali, N.; Hebaishi, H.; Hines, S.; Hinks, A.; Hitman, G. A.; Hocking, L.; Howard, E.; Howard, P.; Howson, J. M.; Hughes, D.; Hunt, S.; Isaacs, J. D.; Jain, M.; Jewell, D. P.; Johnson, T.; Jolley, J. D.; Jones, I. R.; Jones, L. A.; Kirov, G.; Langford, C. F.; Lango-Allen, H.; Lathrop, G. M.; Lee, J.; Lee, K. L.; Lees, C.; Lewis, K.; Lindgren, C. M.; Maisuria-Armer, M.; Maller, J.; Mansfield, J.; Martin, P.; Massey, D. C.; McArdle, W. L.; McGuffin, P.; McLay, K. E.; Mentzer, A.; Mimmack, M. L.; Morgan, A. E.; Morris, A. P.; Mowat, C.; Myers, S.; Newman, W.; Nimmo, E. R.; O'Donovan, M. C.; Onipinla, A.; Onyiah, I.; Ovington, N. R.; Owen, M. J.; Palin, K.; Parnell, K.; Pernet, D.; Perry, J. R.; Phillips, A.; Pinto, D.; Prescott, N. J.; Prokopenko, I.; Quail, M. A.; Rafelt, S.; Rayner, N. W.; Redon, R.; Reid, D. M.; Renwick; Ring, S. M.; Robertson, N.; Russell, E.; St, C. D.; Sambrook, J. G.; Sanderson, J. D.; Schuilenburg, H.; Scott, C. E.; Scott, R.; Seal, S.; Shaw-Hawkins, S.; Shields, B. M.; Simmonds, M. J.; Smyth, D. J.; Somaskantharajah, E.; Spanova, K.; Steer, S.; Stephens, J.; Stevens, H. E.; Stone, M. A.; Su, Z.; Symmons, D. P.; Thompson, J. R.; Thomson, W.; Travers, M. E.; Turnbull, C.; Valsesia, A.; Walker, M.; Walker, N. M.; Wallace, C.; Warren-Perry, M.; Watkins, N. A.; Webster, J.; Weedon, M. N.; Wilson, A. G.; Woodburn, M.; Wordsworth, B. P.; Young, A. H.; Zeggini, E.; Carter, N. P.; Frayling, T. M.; Lee, C.; McVean, G.; Munroe, P. B.; Palotie, A.; Sawcer, S. J.; Scherer, S. W.; Strachan, D. P.; Tyler-Smith, C.; Brown, M. A.; Burton, P. R.; Caulfield, M. J.; Compston, A.; Farrall, M.; Gough, S. C.; Hall, A. S.; Hattersley, A. T.; Hill, A. V.; Mathew, C. G.; Pembrey, M.; Satsangi, J.; Stratton, M. R.; Worthington, J.; Deloukas, P.; Duncanson, A.; Kwiatkowski, D. P.; McCarthy, M. I.; Ouwehand, W.; Parkes, M.; Rahman, N.; Todd, J. A.; Samani, N. J.; Donnelly, P. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature,* **2010,** *464* (7289), 713-720.

[64] Mandel, A. L.; Peyrot des, G. C.; Plank, K. L.; Alarcon, S.; Breslin, P. A. Individual differences in AMY1 gene copy number, salivary alpha-amylase levels, and the perception of oral starch. *PLoS. One.,* **2010,** *5* (10), e13352.

[65] Magri, C.; Sacchetti, E.; Traversa, M.; Valsecchi, P.; Gardella, R.; Bonvicini, C.; Minelli, A.; Gennarelli, M.; Barlati, S. New Copy Number Variations in Schizophrenia. *PLoS ONE,* **2010,** *5* (10), e13422.

[66] Prescott, N. J.; Dominy, K. M.; Kubo, M.; Lewis, C. M.; Fisher, S. A.; Redon, R.; Huang, N.; Stranger, B. E.; Blaszczyk, K.; Hudspith, B.; Parkes, G.; Hosono, N.; Yamazaki, K.; Onnie, C. M.; Forbes, A.; Dermitzakis, E. T.; Nakamura, Y.; Mansfield, J. C.; Sanderson, J.; Hurles, M. E.; Roberts, R. G.; Mathew, C. G. Independent and population-specific association of risk variants at the IRGM locus with Crohn's disease. *Hum. Mol Genet.,* **2010,** *19* (9), 1828-1839.

[67] Servin, B.; Stephens, M. Imputation-Based Analysis of Association Studies: Candidate Regions and Quantitative Traits. *PLoS Genet,* **2007,** *3* (7), e114.

[68] Stranger, B. E.; Nica, A. C.; Forrest, M. S.; Dimas, A.; Bird, C. P.; Beazley, C.; Ingle, C. E.; Dunning, M.; Flicek, P.; Koller, D.; Montgomery, S.; Tavare, S.; Deloukas, P.; Dermitzakis, E. T. Population genomics of human gene expression. *Nat. Genet.,* **2007,** *39* (10), 1217-1224.

[69] Emilsson, V.; Thorleifsson, G.; Zhang, B.; Leonardson, A. S.; Zink, F.; Zhu, J.; Carlson, S.; Helgason, A.; Walters, G. B.; Gunnarsdottir, S.; Mouy, M.; Steinthorsdottir, V.; Eiriksdottir, G. H.; Bjornsdottir, G.; Reynisdottir, I.; Gudbjartsson, D.; Helgadottir, A.; Jonasdottir, A.; Jonasdottir, A.; Styrkarsdottir, U.; Gretarsdottir, S.; Magnusson, K. P.; Stefansson, H.; Fossdal, R.; Kristjansson, K.; Gislason, H. G.; Stefansson, T.; Leifsson, B. G.; Thorsteinsdottir, U.; Lamb, J. R.; Gulcher, J. R.; Reitman, M. L.; Kong, A.; Schadt, E. E.; Stefansson, K. Genetics of gene expression and its effect on disease. *Nature,* **2008,** *452* (7186), 423-428.

[70] Subramanian, A.; Tamayo, P.; Mootha, V. K.; Mukherjee, S.; Ebert, B. L.; Gillette, M. A.; Paulovich, A.; Pomeroy, S. L.; Golub, T. R.; Lander, E. S.; Mesirov, J. P. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A,* **2005,** *102* (43), 15545-15550.

[71] Zhong, H.; Yang, X.; Kaplan, L. M.; Molony, C.; Schadt, E. E. Integrating pathway analysis and genetics of gene expression for genome-wide association studies. *Am J Hum. Genet.,* **2010,** *86* (4), 581-591.

[72] Weng, L.; Macciardi, F.; Subramanian, A.; Guffanti, G.; Potkin, S. G.; Yu, Z.; Xie, X. SNP-based pathway enrichment analysis for genome-wide association studies. *BMC. Bioinformatics.,* **2011,** *12* (1), 99.

[73] Wang, K.; Li, M.; Bucan, M. Pathway-Based Approaches for Analysis of Genomewide Association Studies. *Am. J. Hum. Genet.,* **2007,** *81* (6).

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

[75] Torkamani, A.; Topol, E. J.; Schork, N. J. Pathway analysis of seven common diseases assessed by genome-wide association. *Genomics,* **2008,** *92* (5), 265-272.

[76] Schunkert, H.; Konig, I. R.; Erdmann, J. Molecular signatures of cardiovascular disease risk: potential for test development and clinical application. *Mol. Diagn. Ther.,* **2008,** *12* (5), 281-287.

[77] Jia, P.; Wang, L.; Meltzer, H. Y.; Zhao, Z. Pathway-based analysis of GWAS datasets: effective but caution required. *Int. J. Neuropsychopharmacol.,* **2011,** *14*, 567-572.

[78] Holmans, P. Statistical methods for pathway analysis of genome-wide data for association with complex genetic traits. *Adv. Genet.,* **2010,** *72*, 141-179.

[79] Gibbs, J. R.; van der Brug, M. P.; Hernandez, D. G.; Traynor, B. J.; Nalls, M. A.; Lai, S. L.; Arepalli, S.; Dillman, A.; Rafferty, I. P.; Troncoso, J.; Johnson, R.; Zielke, H. R.; Ferrucci, L.; Longo, D. L.; Cookson, M. R.; Singleton, A. B. Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS. Genet.,* **2010,** *6* (5), e1000952.

[80] Feinberg, A. P. Genome-scale approaches to the epigenetics of common human disease. *Virchows Arch.,* **2010,** *456* (1), 13-21.

[81] Asimit, J.; Zeggini, E. Rare variant association analysis methods for complex traits. *Annu. Rev. Genet.,* **2010,** *44*, 293-308.

[82] Wray, N. R.; Purcell, S. M.; Visscher, P. M. Synthetic associations created by rare variants do not explain most GWAS results. *PLoS Biol.,* **2011,** *9* (1), e1000579.