Fine scale mapping of genetic loci associated with human disease

Julian Maller

St Anne's College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

October 2012

For my son Douglas

ישמך אלהים
כאפרים וכמנשה׃
יברכך יי וישמרך׃
יאר יי פניו אליך ויחנך׃
ישא יי פניו אליך
וישם לך שלום׃

Acknowledgements

Peter Donnelly, I am very grateful for the opportunity to come here and do my D.Phil. Thanks for the guidance, encouragement, and support I have received over the years in Oxford. Beyond that you have been a really good friend, which I'm certain will continue in the years to come. On that note I would also like to thank everyone in the Donnelly group for providing friendship (and lots of beer) during my time here, in particular Jake, Damjan and Zhan who were excellent to work with on the fine mapping project and also the CCC2 team; Amy and Chris. It has been an honor to get to know and work with you all.

Behind every successful team you will find a wonderful administrator. I have been lucky enough to enjoy the support of Victoria and Kate during my D.Phil work. Without your help I would have had a much harder time (especially when trying to book time with Peter).

Thanks also to Mark McCarthy for a year of diabetes research and for your friendship and support here in Oxford.

I would like to thank the Wellcome Trust and the Statistics Department, for funding during my time here and for sponsoring the work described in this thesis. Thanks to St. Annes college in Oxford for providing such a nice community for expat students.

On a more personal note I would like to thank my parents Michael and Jennifer for always supporting and loving me, and Abby for being such a great sister and friend. You mean so very much to me.

Finally I would like to thank my wife Cecilia for going with me to Crystal Burger in New Orleans, for marrying me, for our son Douglas and for your unconditional love. It is wonderful to share my life with you and you mean the world to me.

Abstract

Genome-wide association studies (GWAS) use custom SNP arrays that provide effective genetic coverage, but do not detail the association of confirmed susceptibility regions with as dense a marker map as possible. In this thesis I will describe a fine mapping study across 14 associated regions, in 8,000 samples from three diseases (Type 2 diabetes mellitus (T2D), coronary artery disease (CAD) and Graves Disease (GD)) and a control group, including over 5,500 successfully genotyped SNPs. We defined using Bayes theorem sets of SNPs (credible sets) that were 95% likely (posterior probability) to contain the causal disease variants. In three of the 14 regions *TCF7L2* and *CDKN2A/B* in T2D and *CTLA4* in GD, we found that much of the posterior probability after the fine mapping rested on a single SNP. In seven of the 14 regions (*CDKN2A/B* in CAD, *CTLA4* in GD, and *CDKAL1*, *FTO*, *HHEX*, *TCF7L2*) and *CDKN2A/B* in T2D), including the three just mentioned, the credible sets are relatively small. For these regions, the fine mapping experiment has provided useful information, at least in excluding large numbers of SNPs from being causal. Almost none of the SNPs in our credible regions had obvious functions, illustrating our lack of knowledge of genome sequence in modulating gene expression and susceptibility to common disease. Based on this experiment, I outline and discuss the possibilities and challenges in identifying causal variants for complex traits through fine mapping of GWA signals. Further, I evaluate the possibility of using genotype imputation in the context of fine mapping, both as a method for fine mapping and as a way to increase efficiency in fine mapping studies.

Contents

Chapter 1

Literature Review

1.1 Introduction

In the past few years, genome-wide association studies (GWAS) have been performed on over 200 heritable diseases and traits with the aim to identify causal genetic variant(s). This has resulted in the identification of more than 1400 common genetic variants that are significantly associated with a variety of diseases and traits [http://www.genome.gov/gwastudies/].

Despite this massive success, a large proportion of heritability for a large majority of traits remains unexplained. For example, the current set of 56 established type 2 diabetes susceptibility-loci (Zeggini et al., 2008) (Kong et al., 2009) (Voight et al., 2010) (Dupuis et al., 2010) (Qi et al., 2010) (Tsai et al., 2010) (Shu et al., 2010) (Yamauchi et al., 2010) (Kooner et al., 2011) (Cho et al., 2012) captures around 10 percent of familial aggregation of the disease. Expanding current efforts to identify additional common, but also low-frequency/rare variants that contribute to these traits and disorders will hopefully aid in explaining more of the

heritability.

Also, even for the majority of the established associations there is little known about which variant(s) are actually causal for the downstream traits and diseases associated to them. To progress from association signal to identifying causal variants and after that to gain an understanding of the underlying biological mechanisms and pathways for the associated diseases and traits is one of the major hurdles to overcome before it is possible to move forward to translation. Given this, further investigation of the genetic variants over the entire allelic spectrum (i.e. fine-mapping) will be important to fully elucidate the complex underlying genetic architecture of genetic diseases and traits and holds promise for extending our knowledge of the underlying biological susceptibility mechanisms.

1.2 Genetic Disorders and Traits

Heritable human diseases and traits fall across a spectrum, at each end of which are two general categories:

i) Monogenic disorders, or so-called Mendelian disorders, which are rare disorders caused by mutations in a single gene and have a simple inheritance pattern;

ii) Complex disorders and traits, or so called multifactorial disorders and traits, where a number of genes and environmental factors contribute to susceptibility, but no clear inheritance pattern exists.

Over the last twenty years there has been much success in identifying the genetic loci predisposing to Mendelian disorders such as Huntington's disease (Gusella et al., 1983). In contrast, progress has been much slower for diseases and traits with a more complex underlying genetic architecture, such as Type 2

Diabetes Mellitus (T2D). With the aim to unravel the underlying genetic susceptibility of complex traits and disorders researchers have, in the past, primarily focused on two different strategies of identification: linkage analysis, later followed by association studies. The ultimate hope of these efforts is that the identification of associated genetic variants will lead to the illumination of new pathways connected to the trait, as well as the development of novel diagnostic and therapeutic tools (Risch and Merikangas, 1996, Altshuler et al., 2008).

1.2.1 Phenotypes

The phenotypes are the physical and biochemical characteristics (traits) of an individual, which result from interactions between its genetic makeup and the environment. Generally phenotypes can be divided into one of two categories:

Quantitative traits, such as height or blood pressure, vary continuously in the population; they are typically complex polygenic traits. Genetic association testing can be performed directly on raw quantitative traits, though it is important to test the extent to which the trait is really normally distributed in the population as this is often required for the statistical tests used to analyze them to be valid (Balding, 2006). Standard measures to determine whether a trait is normally distributed are to calculate deviations from normality in terms of skewness (the lack of symmetry of the distribution in either direction) and kurtosis (being more or less peaked compared to a normal distribution). Often, the data will be transformed to improve its fit to a normal distribution.

It is also common to search for and exclude outliers (observations that deviate notably from other values in a random sample from a population). Such unusual observations are often detected using box plots of the phenotypes in the popula-

tion. Routinely people exclude values that deviate from the mean of measures more than 2 standard deviation but importantly there is no universal mathematical rule for what an outlier definition is and one has to treat these carefully. It is good practice to investigate the reason for the extreme value (measurement error, unit errors and unit conversion errors are common sources) to determine the source.

Qualitative or *dichotomous* traits form the second category. For these traits, the phenotype is discrete and can take on one of only a few different values. By far the most common in this context is the classification of individuals as cases (suffering from the disease in question) or controls (strictly, not suffering from the disease, but in practice controls are often just a sample of individuals from the population.) Sometimes, the diagnosis of a disease (dichotomous trait) is based on a threshold of a quantitative trait. An example of this is T2D, where a fasting plasma glucose level of 7mmol/l or above is used for diagnosing an individual with T2D.

1.2.2 Genotypes and Genotyping Methods

Traditionally in genetics a genotype describes the genetic make-up of an individual (i.e. the specific allele make-up of the individual) but more recently the definition is modified to describe the specific variants (alleles) an individual has for a given genetic polymorphism.

In the early days of genetic experiments, an individuals genotype was not directly measured but inferred from the segregation of phenotypes of a pedigree of related individuals (Mendel, 1865). Later after DNA was discovered (Watson and Crick, 1953) and scientists started to construct maps of genetic variation the common method of analyzing genetic polymorphisms was restriction frag-

ment length polymorphism (RFLP) analysis of single base substitutions. These are labour-intensive, DNA consuming and difficult experiments to perform and as a result only be performed in small scale. They were soon replaced by PCR based investigations of tandem repeat polymorphisms, particularly microsatellite polymorphisms. These have the major advantage over base substitution RFLPs that they are far more informative (due to the higher number of alleles) and can be typed much faster (Weber, 1990).

Those early genetic efforts were limited by the limited number of available polymorphisms (Collins et al., 1997) but later it was realized that single base substitutions are much more common than initially thought (Nickerson et al., 1998).

The approaches to assessing DNA sequence differences between individuals offer considerable promise for reducing the cost and increasing the rate at which large numbers of single nucleotide polymorphisms (SNPs) can be genotyped either using DNA chips (Fodor et al., 1991) for genome wide detection and for smaller scale projects, versions of single base extension (or minisequencing primer extension) methods (Syvänen, 1999).

1.3 Recombination

Recombination is the natural process of crossing over between homologous chromosomes during meiosis. Measuring recombination is a technically challenging process but they have shown that recombination is non-random (Kauppi et al., 2004). In humans analysis of inheritance in pedigrees, sperm studies or population genetic data are used to estimate recombination rates (Auton and McVean, 2007).

These methods have been used to identify specific regions in which crossing-over events cluster in so called "recombination hotspots". These are abundant throughout the human genome, occurring non-randomly on average every 200 kb, preferentially outside genes (McVean et al., 2004). The discovery and subsequent map of the recombination hotspots throughout the genome have facilitated both the design and analysis of association studies through an improved understanding of the distribution of linkage disequilibrium (LD, see "Linkage Disequilibrium" in this chapter) in natural populations (Stumpf and McVean, 2003).

1.4 Linkage Analysis and Monogenic Disorders

Linkage analysis is one method traditionally used for searching for genetic loci harboring disease causing genetic variants. In linkage analysis, genetic loci are assayed sparsely across the genome in a set of related samples, some of which are affected by the disorder and some unaffected. Genetic polymorphisms are used to track co-segregation of chromosomal regions and a trait or a disorder within a set families. In families, two loci that are located near each other on a chromosome (e.g. within 10Mb in humans) will tend to be inherited together, i.e. they are "linked together". Recombination may occur during meiosis, resulting in the separation of previously linked loci (Cardon and Bell, 2001).

The probability of recombination is a function of the genetic distance between two loci. For each locus, the probability of recombination between the locus and a putative "disease locus" can be calculated. The key measure of evidence used in linkage is the logarithm of odds (*LOD*) score. This is the log of the likelihood comparing each locus being linked to the disease locus vs. the likelihood of them

being unlinked (Cardon and Bell, 2001).

In monogenic disorders, parametric (i.e. model based) linkage can sometimes be used, as the mode of inheritance may be known. In this scenario, the disease model must be specified, along with several other parameters including disease prevalence and allele frequencies. This is a tool with substantial power to detect effects, provided the model is right and accurate parameter estimates exist. However, when using linkage to study complex phenotypes, most or all of these parameters are unknown, and so model free (i.e. non-parametric) methods must be used. In multiply-affected families, the affected relatives will tend to display excess sharing of haplotypes that are identical by descent, in the region associated to disease. Non-parametric methods generally take advantage of this idea, most commonly using affected sib-pairs (Risch, 1990).

One desirable feature of using a linkage strategy is that it searches the entire genome without any a priori hypothesis about which loci or genes might be of importance, and so is "hypothesis free". While linkage analysis proved successful in many monogenic disorders, for all but a few common diseases such studies produced non-significant results which could not be consistently replicated (Risch, 2000).

Linkage efforts are most well powered for identification of low frequency or rare disease causing variants (minor allele frequency (MAF)$<<1\%$) with large effects on the disorder. It is becoming increasingly clear that we are not surveying this range of the allele frequency spectrum, as the tools hereto used to identify genetic susceptibility variants in common complex disorders have very limited detection power (McCarthy et al., 2008). For example, the type 2 diabetes (T2D) associated variant Pro12Ala on *PPARγ* would require a linkage study

involving around 3 million sib-pairs to obtain a significant LOD-score and be detected through linkage (Altshuler et al., 2000).

Recently, this progress in Mendelian disease loci identification is further sped up by the advent of exome sequencing which offers an alternative to linkage analysis as it allows the detection of underlying coding variants using a small number of unrelated, affected individuals, such as for Millers Syndrome (Ng et al., 2010).

1.5 Association Studies and Common Complex Traits

One hypothesis about allelic architecture for common, complex diseases and trait risk is that it is due to loci containing one or more common disease predisposing genetic variant(s). The common disease-common variant hypothesis (CD-CV) is based on the idea that common alleles ($MAF > 5\%$) will be, at least partly, underlying susceptibility to complex polygenic diseases. In contrast to monogenic disorders caused by rare variants ($MAF << 1\%$) with very large effect sizes, the CD-CV hypothesis suggests that each variant at each gene influencing a complex disease will have a small additive or multiplicative effect on the disease phenotype (Gabriel et al., 2002).

Based on these notions, an alternative primary approach for searching for genetic disease loci known as association studies was shown to be a powerful tool for detecting common loci of small effect (Risch and Merikangas, 1996). An association study looks for simple correlations between a phenotype (any observable characteristic or trait of an organism) and the genotype (combination of alleles) at a genetic variant (Cardon and Bell, 2001).

Association studies are typically performed using either a set of unrelated

cases and controls, or on a set of nuclear families with affected offspring. In a "case/control" setting, one collects a set of samples with a particular disorder (cases), and an unrelated set of samples without that disorder (controls). Genetic variants are genotyped in both sets of samples. Then, the variation data are analysed for statistical correlation between the phenotype and genotype; in other words, for each genetic variant, one ask the simple question "does this disorder occur more frequently in samples with allele *A* of this variant than allele *B*" (Risch and Merikangas, 1996). If the difference is observed to be statistically significant, typically at a stringent level of significance (see the section "Statistical significance"), then the variant is considered associated to the disorder. There are a number of potential difficulties to this approach however, most prominently that of "population stratification": the frequency of genetic loci varies across human population groups, and so any differences in ancestry between cases and controls can result in spurious associations (Cardon and Palmer, 2003). Until recently association studies were limited to studying only small regions of the genome (a limited number of markers) at a time due to cost and genotyping capacity.

1.6 Linkage Disequilibrium

The completion of the human genome reference sequence (Lander et al., 2001) enabled the systematic localization and categorization of genetic variants throughout the genome. The most common type of genetic variant in the human genome is the single nucleotide polymorphism (SNP) (Wang et al., 1998). A SNP is DNA sequence variation at a single nucleotide base, most often bi-allelic. Generally, a genomic location will only be described as a SNP when both the original al-

lele and a second allele are observed to be present in the relevant population. To date, current databases (e.g. dbSNP (Sherry et al., 2001a) and the HapMap (Frazer et al., 2007)) contain more than ten million common ($MAF > 5\%$) SNPs, which is a density of about one SNP per 290 base-pairs and these are thought to account for 90% of human genetic variation (Kruglyak and Nickerson, 2001) (Reich et al., 2003) (Carlson et al., 2003) (Goddard et al., 2000) (Stephens et al., 2001a).

The HapMap project was launched in 2002 with the aim to provide a catalog and description of common genetic variation (Frazer et al., 2007). This was followed in 2008 by the 1000 Genomes Project which is currently investigating the anonymized genomes of about 2500 people from about 25 populations around the world using next-generation sequencing technologies. This will expand the HapMap's initial catalogue of common variants to also include low-frequency variants (1000 Genomes Project Consortium, 2010).

The recent and rapid progress in genetic association studies has partly been fueled by the characterization of the allelic correlation between nearby genetic variants, known as "linkage disequilibrium" (LD). This correlation is a non-random association of alleles at two or more loci, resulting from the loci being located close to each other on the same chromosome, and thus tending to be inherited together (Lewontin, 1964).

While the concept of LD had long been known, theoretical analyses based on population genetic models suggested that strong correlation would only be present for very short distances and would drop off quickly (Kruglyak, 1999). One of the first demonstrations that this might not be the case occurred during a search for genes predisposing to Inflammatory Bowel Disease (Daly et al., 2001). Daly and colleagues reported that the correlation between SNPs extended much farther than

expected; in fact, across long stretches of the region, a small number of combinations of alleles seemed sufficient to represent variation across regions as large as 92 kb. Further, when looking at regions that spanned more than one of these segments virtually no correlation was observed, suggesting that recombination in the region may be limited (Daly et al., 2001, Wall and Pritchard, 2003, Jeffreys et al., 2001). It was concluded that the genome consists of stretches of high LD regions (also called "LD-blocks" or "haplotype blocks") separated by short distinct segments of very low LD, which are now known to be caused by "recombination hotspots" (Altshuler et al., 2005). These common LD-blocks display only limited haplotype diversity and a small number of distinct common haplotypes account for the majority of chromosomes in the population (Daly et al., 2001) (Gabriel et al., 2002) (Patil et al., 2001).

This block structure within high LD regions might aid the association studies of common alleles, because recombination sites could make up boundaries for causative candidate genes and variants and also indicate how far to extend the search for functional variants. With this realization about the common haplotype blocks in high LD regions, allelic correlation between SNPs allow a subset of them to act as proxies for (or "tag") the majority of common SNPs. As a result, it is possible to capture a large proportion (80 %) of allelic diversity across the genome by genotyping only a fraction of known SNPs (300,000-500,000 SNPs) (Barrett and Cardon, 2006) (Pe'er et al., 2006).

Another important consequence of the correlation structure is that semi-dense genotyping captures the information at both known and unknown common variants. Thanks to projects such as 1000genomes we now have a nearly complete catalogue of common variants for many populations, but that is a quite recent

development.

1.6.1 LD Measures

More specifically, LD refers to association between the alleles at two loci. If we have two biallelic loci located near each other on a chromosome, a and b, with allele frequencies p_{a_1} and p_{a_2}, and p_{b_1} p_{b_2}, then the probabilities of observing pairs of alleles at these loci are $p_{a_1b_1}, p_{a_1b_2}$, and $p_{a_2b_1}$. In a state of linkage equilibrium, the alleles are independent between loci, and so the multi-allele probabilities can be calculated as $p_{a_ib_j} = p_{a_i}p_{b_j}$. In the presence of LD, alleles at the two loci are not independent, and this relationship does not exist (Wray NR, 2008). These multi-allele probabilities may also be referred to as haplotype probabilities (or frequencies). The term *haplotype* is used to describe pairs of alleles at loci nearby on the same chromosome and which tend to be transmitted together.

One of the most common and simplest measures of LD is the covariance D (Falconer DS, 1996) (Lynch M, 1998). For biallelic loci this is defined as:

$$\begin{aligned} D &= p_{a_1b_1} - p_{a_1}p_{b_1} \\ &= p_{a_2b_2} - p_{a_2}p_{b_2} \\ &= -(p_{a_1b_2} - p_{a_1}p_{b_2}) \\ &= -(p_{a_2b_1} - p_{a_2}p_{b_1}) \end{aligned}$$

The sign of D is determined by which allele is assigned to $a1$ or $a2$, and so

in comparisons between loci the absolute value can be used. The maximum and minimum values of D depend on the allele frequencies of the loci. The absolute value of the maximum value, $|D_{max}|$, is the smaller of $p_{a_1}p_{b_2}$ and $p_{a_2}p_{b_1}$ when $D > 0$ and the smaller of $p_{a_1}p_{b_1}$ and $p_{a_1}p_{b_2}$ when $D < 0$ (Wray NR, 2008). Another measure of LD is a scaled version of D, known as D': $D' = D/D_{max}$. The absolute value of this measure is generally used. $|D'|$ is much less dependent on allele frequency than D, which more readily allows comparisons of value between different pairs of loci (Wray NR, 2008, Wang et al., 2005).

Another key metric for measuring LD between SNPs is r^2, which is the squared correlation between alleles at two SNPs (Ardlie et al., 2002, Hill and Robertson, 1968).

r^2 is defined as:

$$r^2 = \frac{D^2}{p_{a_1}p_{a_2}p_{b_1}p_{b_2}}$$

While r^2 and D (and D') are closely related, r^2 is more commonly used in the context of association studies. One reason for this is that there is a direct relationship between the power to detect a causal variant and the r^2 between the genotyped and causal variant (Wray NR, 2008, Wang et al., 2005).

1.7 Genome Wide Association Studies

As of 2004, little progress had been made towards dissecting the genetic mechanisms underlying common complex disease. In the time since, the possibility to perform association studies on a far larger scale has transformed this picture.

Rapid technological improvements culminated in the availability of commer-

cial genotyping assays allowing the simultaneous, accurate and affordable genotyping of hundreds of thousands of SNPs across the entire genome (Fan et al., 2006). By leveraging the available tagSNP information from the HapMap, these assays were able to characterize the majority of common variation across the genome, while only directly genotyping about 10% of common SNPs (Barrett and Cardon, 2006) (Pe'er et al., 2006) (Gabriel et al., 2002). Two early examples of commonly used genotyping arrays are the Affymetrix GeneChip Human Mapping 500k Array Set and the Illumina 317K chip. The technology behind these arrays continues to improve, and genotyping arrays are now available containing $>$ 5 million SNPs.

One of the first published reports of a GWAS described a case control study of age-related macular degeneration (AMD). The results of this study were striking. Using only 96 cases and 50 controls, Klein et al. discovered a SNP in the gene complement factor H with an odds ratio of 2.72 (Klein et al., 2005) (Haines et al., 2005). This finding was later repeatedly replicated (Maller et al., 2006) (Despriet et al., 2006). Also in 2005, Rivera et al. discovered a SNP in an uncharacterised gene with an odds ratio of 2.69 for AMD (Rivera et al., 2005). These two findings combined with two further discoveries resulted in a report suggesting that the combined effects of the four loci explained greater than 50% of the genetic component of AMD (measured using sibling recurrence risk) (Maller et al., 2006) (Maller et al., 2007). This was an early and striking success story for complex disease genetics. After almost two decades of searching unsuccessfully for genes predisposing to AMD, half of the genetic predisposition had been explained in just two years. The success in AMD provided hope for many planned or ongoing GWAS. It quickly became clear that the common SNPs with high odds ratios

observed in AMD were atypical, with odds ratios in the highest 1% of SNPs found via GWAS (as of November 2011) (NHGRI, 2011). Much more common for diseases have been odds ratios on the order of 1.1-1.3, requiring substantially larger sets of samples. With this realization that the effect sizes of associations would be much smaller than anticipated it became clear that to find true associations, studies would need to become far larger (thousands or tens of thousands of samples). No studies of that size had been previously performed or even considered, and this necessity has catalysed large-scale international collaboration around association analysis.

As the odds ratio drops the contribution to the overall heritability of the disorder rapidly drops. A typical example is T2D. While for AMD 5 SNPs explained approximately 50% of the heritability, the first 18 replicated SNPs for T2D explained about 6% (Manolio et al., 2009). Another example is Crohn's disease (CD), where the first 32 replicated SNPs explain approximately 20% of the genetic risk (Manolio et al., 2009).

One challenge that arose quickly for GWAS was that of technical artifacts. Even subtle genotyping error or bias is a non-trivial problem in the context of a GWAS. This can result from a number of causes including problems in the lab handling of samples, different handling of cases vs. controls, among many others (Hirschhorn and Daly, 2005). While these are always important issues in an experiment, the very large increase in the scale of data being collected in GWAS has made this much more prominent and presents a continuing challenge in the analysis and interpretation of data (WTCCC, 2007). This can be minimized through rigorous quality control of genotype data and careful analysis as well as strict replication of any promising findings.

Due to their design (i.e. to genotype common variants that act as tags or proxies of other common variants) GWAS are well-powered to characterize common variants down to relatively small effect sizes, as SNP variation is very effectively captured down to a MAF of 5% (Pe'er et al., 2006). Sample size plays an important role in the design of these studies, and power will depend very much on the sample size (both in cases and controls). For variants below MAF of 5% "coverage" (the proportion of variation at untyped SNPs which can be recovered as a result of LD) drops off rapidly. Current large-scale resequencing efforts in both public (1000Genomes , (1000 Genomes Project Consortium, 2010)) and disease specific (eg. GoT2D) settings hold promise to aid identification of lower frequency variants relevant for diseases and traits.

1.7.1 GWAS Significance Levels

Controlling the false positive rate is of central importance when testing hundreds of thousands of SNPs for association to a phenotype. The most common approach to this is via a threshold for statistical significance. From a Bayesian of Frequentist perspective there are (different) norms for having stringent significance thresholds for GWAS. There is no "right" answer as to what this threshold should be, but the field has settled on a p-value threshold of 5×10^{-8}, which seems to work quite well for avoiding false positives. For a general discussion of GWAS significance see (WTCCC, 2007). However, no matter how significant an initial finding may appear, replication of a putative association in an independent sample is considered essential.

1.7.2 Success of GWAS

In the seven years since the first GWAS results were published it has proven a major success. As of June 2011 there were over 1,600 independent associations at genome wide significance ($p \leq 5 \times 10^{-8}$, see previous section on statistical significance) for 237 diseases and traits reported including AMD, T2D, T1D, cancer, inflammatory bowel disease (IBD) as well as a variety of anthropometric and laboratory traits (Manolio et al., 2008).

Figure 1.1 displays associations which have been detected in GWA studies, labelled according to their position in the genome.

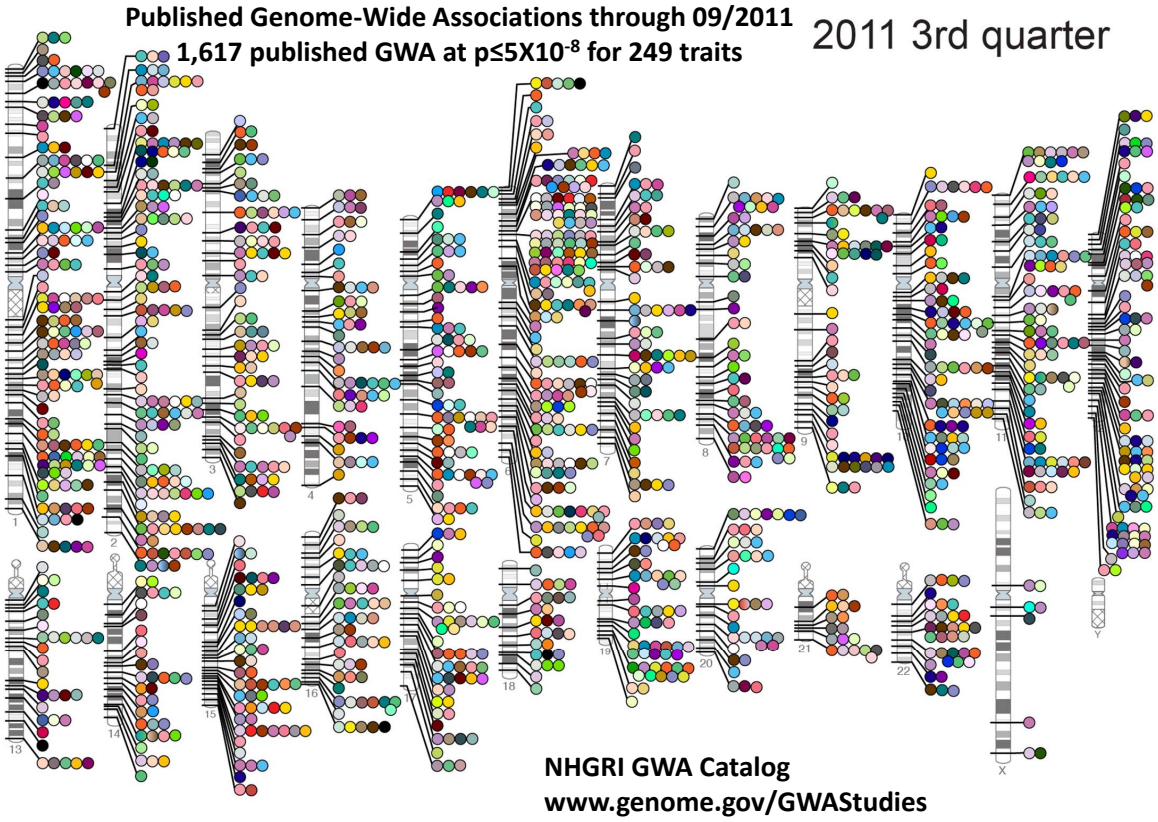


Figure 1.1: SNP-trait associations detected in GWA studies. Associations significant at $P < 9.9 \times 10^7$ are shown according to chromosomal location and involved or nearby gene, if any. Colored boxes indicate similar diseases or traits. From: Hindorff LA, MacArthur J (European Bioinformatics Institute), Morales J (European Bioinformatics Institute), Junkins HA, Hall PN, Klemm AK, and Manolio TA. A Catalog of Published Genome-Wide Association Studies. Available at: www.genome.gov/gwastudies. Accessed 11/8/2012.

As mentioned above, the identified common susceptibility variants taken together only account for a small portion ($< 10\%$) of the heritability of the studied diseases and traits . There are exceptions to this, such as IBD, where meta-analysis of GWAS have identified 71 loci associated with one of the sub-forms of the disease (CD) and these associated loci taken together account for about 23% of the heritable portion of disease risk (Franke et al., 2010) .

One of the strengths of the GWAS strategy is that as a hypothesis free screening strategy it will lead to the discovery of associations without any prior knowledge or assumptions about disease and trait aetiology. In the GWAS results yielded so far we see the implication of genes of both known and unknown function in both known and new etiological pathways. Example of novel pathways that are highlighted are the autophagy pathway in IBD (Rioux et al., 2007) and the complement pathway in AMD (Klein et al., 2005). The hope is that the identifications of genes and pathways that are associated to diseases may accelerate the development of new therapies.

It is important to highlight the fact that while GWAS has been very successful at identifying true associations, we have thus far rarely been able to translate this success into identification of the actual causal variant(s) and through which gene(s) the associations are exerting their effect.

1.8 Beyond GWAS

While the field is continually expanding the GWAS efforts into larger collaborations to increase samples sizes and identify additional common variation associated with traits and diseases, loci already associated often remain largely unex-

plored.

While there is no clear path for how to best move forward, a number of research avenues are open to investigate these including: fine-mapping and bioinformatic efforts to identify the causal variant(s), use expression data and phenotype refinements to identify the gene(s) through which the association exerts an effect on the disease or trait in question.

1.8.1 Fine-scale Mapping of Associated Regions

Once a significant and replicated association has been established, a natural next step would be to exhaustively fine map the region to fully characterize the genotype (and haplotype) patterns.

The first step in fine-mapping would be to generate as complete a list of genetic variants as possible in the associated locus. New sequencing technologies allow full re-sequencing of individuals across the entire region of association to identify all variants present. Understanding and characterizing the association and the role of genetic variants within low LD regions will be challenging and will require deeper re-sequencing efforts in larger samples to get a comprehensive catalogue of variants.

The second step would be confirmatory genotyping efforts to expand the number of individuals available for association testing of the newly discovered markers which will hopefully refine the association signal(s) and localize the causative variant(s).

One option to supplement the information from the re-sequencing effort is to capitalize on the 1000Genomes (1000 Genomes Project Consortium, 2010) effort. The high density of SNPs in the 1000Genomes dataset allows the imputation of a

very high proportion of common ($MAF > .05$) SNPs. These imputed datasets may help to narrow the boundaries of an associated region before costly re-sequencing and genotyping efforts take place. This is standard practice today but these resources were not available when this project and thesis were initiated.

Another possibility is to utilize differences in LD patterns between samples of different ethnicities to fine-map associated regions. See Discussion chapter for a description of this and other approaches to fine mapping.

1.8.2 Association Tests

Table 1.1: Genotypes

	0	1	2	Total
Cases	r_0	r_1	r_2	R
Controls	s_0	s_1	s_2	S
Total	n_0	n_1	n_2	N

By far the most common association test, under the assumption that genetic effect is multiplicative on the log odds scale, is known by various names: the additive test, Cochran-Armitage trend test or simply the trend test. For a SNP a, with genotypes AA, AB and BB represented, respectively, as $\{0,1,2\}$ and genotype counts represented as in Table 1.1 the trend test statistic X_G^2 is defined, as:

$$X_G^2 = \frac{N\{N(r_1+2r_2)-R(n_1+2n_2)\}^2}{R(N-R)\{N(n_1+4n_2)-(n_1+2n_2)^2\}}$$

Most of the association testing in this thesis is based on the trend test except where stated otherwise such as testing for secondary signal. In such cases the test being used will be specified in the relevant sections. For more background and further details of the trend test see (Sasieni, 1997) (Balding, 2006) (Zeggini and Morris, 2010).

Logistic Regression

Another approach to association testing when working with a binary phenotype is using logistic regression. One key benefit of logistic regression is that it readily accommodates a number of different inheritance models (Balding, 2006). Let Y_i represent the phenotype of individual i, with $Y_i = 0$ for controls (for both population or cohort controls) and $Y_i = 1$ for cases. For an additive model, let $X_1[i]$ represent the genotype of an individual, with coding of 0,1,2 corresponding to genotypes AA,AB and BB (Balding, 2006) . Let the disease risk of individual i conditional on their genotype be $p = E(Y_i|X_1[i])$. Then, using the logit transformation we have

$$logit(p_i) = \log \frac{p_i}{1-p_i} \tag{1.1}$$

And

$$logit(p) \sim \beta_0 + \beta_1 X_1 \tag{1.2}$$

. This model is then tested using the likelihood ratio test, against the general model in which $\beta_0 = \beta_1$. The corresponding β_1 represents the expected log odds

ratio for the SNP represented by X_1 for cases and controls, or the log relative risk when using a population based sample (WTCCC, 2007) (Balding, 2006).

1.8.3 Imputation of Genetic Variants

Imputation is a method for predicting genotypes at unassayed sites in a sample of individuals. Generally, imputation is used in a situation where there is a reference panel of densely genotyped samples, and a larger set of study samples typed at a subset of the SNPs in the reference panel. Imputation uses SNP correlation patterns from the reference panel and a recombination map to statistically predict genotypes at untyped sites in the set of study samples. (Marchini and Howie, 2010)

There are a number of uses for imputation. First, it can be used to predict genotypes at SNPs that are completely untyped in a study. This is one of the most frequent scenarios in which imputation is used, due to the public availability of large, very dense public reference panels from the HapMap and 1000 genome projects. Other common uses for imputation include filling in missing genotype data or identifying genotyping errors. Imputation also has the potential to be particularly useful in the context of fine mapping (Ioannidis et al., 2009). After the identification of a region associated with a particular phenotype, the identification of the SNP or SNPs underlying the association is of key interest. Imputed genotype data can help narrow the set of possible candidates and limit the set of SNPs to be typed. The drop in the number of SNPs required to be genotyped can be substantial, as I will show in the chapter discussing imputation. It may also be possible to directly identify potentially causal variant, though to do this accurately requires, at minimum, a large and quite complete reference panel. Even with such

a reference, confidently differentiating between highly correlated SNPs is quite challenging due to the inherent uncertainty of imputed genotypes. See Chapter 4 for further discussion of these and other issues.

Figure 1.2 illustrates how imputation works in several settings.

1.9 Aims of thesis

In Chapter 2 and Chapter 3 of this thesis I describe a fine mapping experiment, undertaken with a number of collaborators. Chapter 2 describes the design, implementation and primary analysis of this study. Chapter 3 describes a number of secondary analyses of the data, including annotation of potentially causal variants, secondary/conditional tests and evaluation of GWAS and imputation contributions to the results.

In Chapter 4 I describe an evaluation of how genotype imputation can be used in the context of fine mapping. I first evaluate the potential for imputation as a direct fine mapping approach. I then investigate whether imputation can be used in the design stage of fine mapping studies as a mechanism to increase efficiency in such studies.

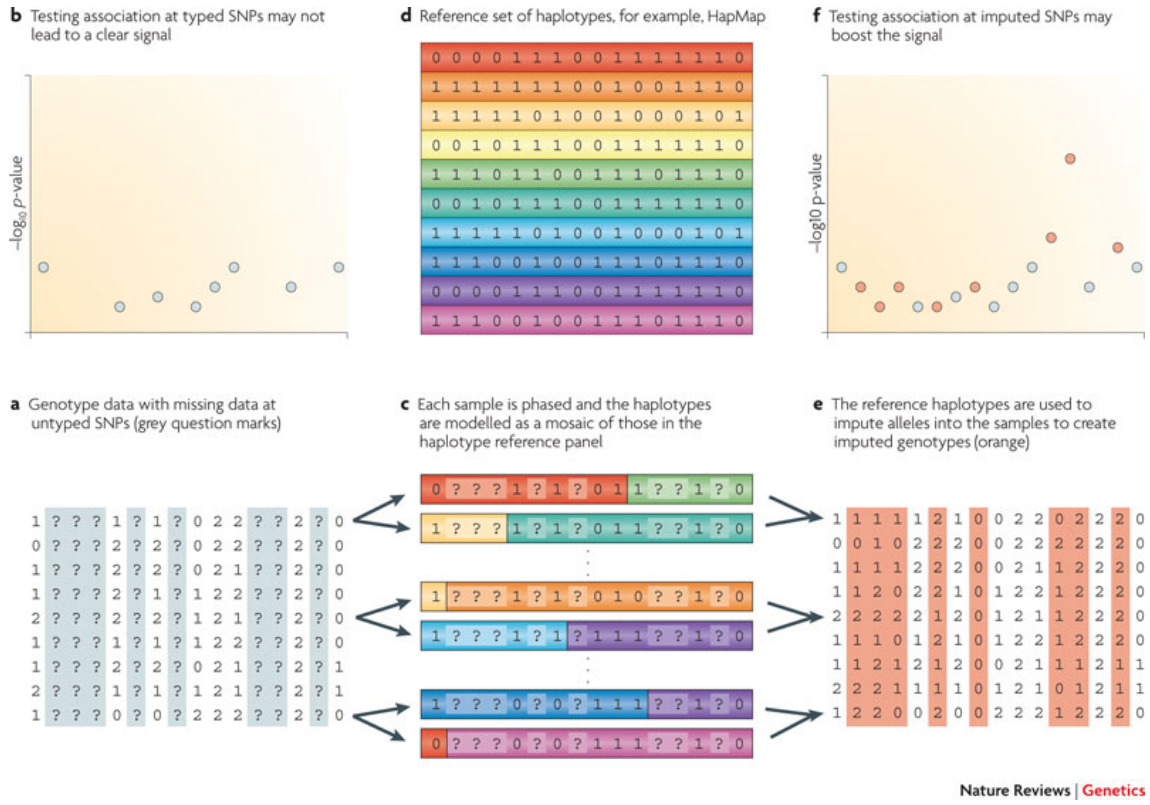


Figure 1.2: The figure above illustrates imputation for a sample of unrelated individuals. The raw data consist of a set of genotyped SNPs that has a large number of SNPs without any genotype data (part a). Testing for association at just these SNPs may not lead to a significant association (part b). Imputation attempts to predict these missing genotypes. Algorithms differ in their details but all essentially involve phasing each individual in the study at the typed SNPs. The figure highlights three phased individuals (part c). These haplotypes are compared to the dense haplotypes in the reference panel (part d). The phased study haplotypes have been coloured according to which reference haplotypes they match. This highlights the idea implicit in most phasing and imputation models that the haplotypes of a given individual are modeled as a mosaic of haplotypes of other individuals. Missing genotypes in the study sample are then imputed using those matching haplotypes in the reference set (part e). Testing these imputed SNPs can lead to more significant associations (part f) and a more detailed view of associated regions. Reprinted from: Marchini J., Howie B. Genotype imputation for Genome-wide association studies. Nature Reviews Genetics 11, 499-511 doi:10.1038/nrg2796

Chapter 2

Design and Primary Analysis of a Fine-Mapping Experiment

2.1 Aims of this chapter

2.1.1 Fine Mapping of Genetic Loci Associated with Human Disorders

In the quest to illuminate the aetiology of common complex disorders in humans, GWA studies are a powerful tool that can provide substantial insight into associated genomic variants. However, these studies are not geared towards identifying specific variants associated with a disorder; their power lies in identifying small regions of the genome associated with a disorder. In nearly all cases, further study will be required to identify the variant(s) functionally influencing the disorder. In this way, a GWAS can be thought of as Hypothesis generating rather than a way of addressing a particular hypothesis.

Once an association has been established between a region of the genome and a phenotype, a natural next step is to try to refine and localize the signal, as well as determine the underlying mechanism. This process is commonly referred to as "fine mapping" an associated locus. Put simply, the goal of fine mapping is to fully characterize the genotype (and haplotype) patterns for a particular associated region. The hope is that this would make it possible to determine the variant functionally influencing the phenotype, known as the causal variant (or variants). While full characterization is the ideal outcome, it is not currently realistic for a number of reasons. One limiting factor is genotyping technology; the best medium to large scale genotyping assay currently available can only successfully genotype around 80% of SNPs. Another limitation is the lack of a complete catalogue of variant sites in the genome, although this has recently improved substantially thanks to the 1000 Genomes Project and the continued efforts of the HapMap project.

These and other issues affecting fine mapping studies will be covered in greater depth in this chapter. An overview diagram of the design and analysis of the project described in this chapter is shown in Figure 2.1.

2.2 Wellcome Trust Case Control Consortium

The Wellcome Trust Case Control Consortium (WTCCC) GWAS was an early, large scale study involving seven common disorders: bipolar disorder, coronary artery disease, Crohn's disease, hypertension, rheumatoid arthritis, type 1 diabetes and type 2 diabetes (WTCCC, 2007). Earlier family studies had shown all seven disorders to be consistently heritable, but at the time the WTCCC study com-

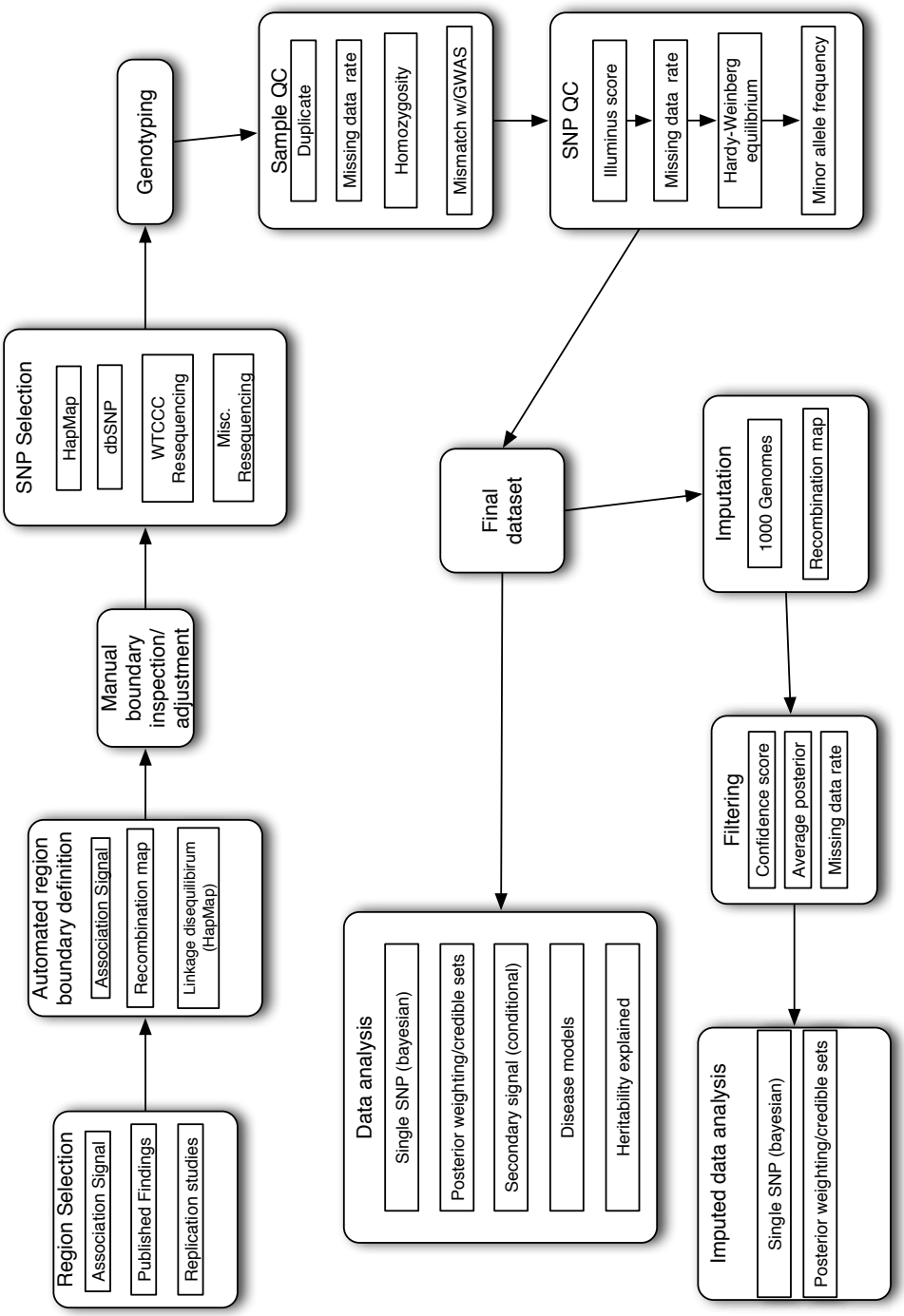


Figure 2.1: Flowchart illustrating the data generation, QC, and analytical pipelines used in the fine mapping project. These will be discussed in this chapter and the chapter that follows.

menced little was known about the genetic architecture of these disorders. For each disorder 2000 unrelated case samples were collected. In addition, two sets of control samples were collected: 1500 individuals from each of the 1958 Birth Cohort and the UK Blood Services. All 17000 samples were genotyped using the Affymetrix GeneChip 500k Human Mapping Array Set, which contains 500,568 SNPs from across the genome.

As one of the first GWAS', and particularly the first very large one, the WTCCC study demonstrated both the utility of the GWAS approach, as well as the potential pitfalls involved. Once the genotype data was cleaned and analysed, 24 independent association signals ($p < 5 \times 10^{-7}$) were found across six of the disorders. Nearly all of these associations were later replicated in independent studies. As well as demonstrating the feasibility and utility of this powerful new approach to disease genetics, the results shed light on a central property of the genetics of heritable common complex disease. The results shed light on the range of effect sizes at common SNPs for common disorders, finding not a single SNP of large effect ($OR > 2$) aside from a small number of previously known loci.

2.3 WTCCC Fine-mapping Study

In order to evaluate possible strategies for fine-mapping GWAS hit regions, we undertook a large scale fine mapping study. The design of the study included associated regions for 3 diseases, which were chosen with a mind towards representing a diversity of allele frequencies, effect sizes, region sizes and LD structure. SNPs for genotyping were drawn from a number of sources including dbSNP and a number of regional resequencing efforts; all known SNPs which successfully

designed were genotyped.

This chapter describes the design, data gathering, QC and primary analysis of this fine-mapping study. The subsequent chapter describes secondary analyses of this study. The study was a part of a large collaboration, and my contribution is only a portion of the overall effort. I will explain in detail the portions I have worked on. Descriptions of work performed by others in the project are noted, such as the annotation work performed together with Jake Byrnes.

2.3.1 Phenotypes and Samples

The experiment included 13 regions known to be associated to one of type-2 diabetes (T2D), coronary artery disease (CAD), or Graves' disease (GD). Six regions were associated with T2D, five regions were associated with CAD and three regions were associated with GD One of the regions is associated to both T2D and CAD, though the associations were independent and separated by a recombination hotspot). For GD, 1,963 samples were included, 1,934 for CAD, and 2,037 for T2D, as well as 1,930 controls taken from the 1958 Birth Cohort and the NBS, for a total of 7,894 samples. Samples were genotyped at 8000 SNPs using a custom designed Illumina iSelect chip.

2.4 Defining Regions

While it's generally quite clear where the most strongly associated variant lies, deciding how far the signal extends in an associated region is a non-trivial task, particularly given the lack of a complete catalogue of genomic variation. Given the relatively small number of comprehensive fine mapping efforts applied to com-

mon variants connected with common complex phenotypes, not much is known about any general properties of causal SNPs or mechanisms.

It follows that even less is known about the relationship between an initial finding and the underlying causal variant(s). Imputing genotypes at untyped SNPs can assist in estimating how far the signal extends, but the reference datasets used in imputation are also incomplete.

There are many methods one could use to define region boundaries, but here we use the 3 common approaches of using pairwise r^2, using the association signal, and using information on recombination.

When using pairwise r^2 to refine the signal, the core idea is to define the region based on correlation to the focal SNP. One way of doing this would be to choose an r^2 threshold and define the region boundaries such that all SNPs with an r^2 above the threshold are within the region. Correlations would generally be calculated using a reference dataset such as the HapMap in order to include as many SNPs as possible (assuming there is a suitable population in the reference).

The approach using the association signal across the region is similar to the correlation based method, but defines boundaries to include all SNPs where the measure of association exceeds some threshold. The threshold can be either an absolute threshold or relative to the focal SNP. Note that as this uses association information, data from the GWAS is used rather than a reference data set.

The third approach uses information on recombination levels in order to define the boundaries. There are different ways to do this, but a natural approach is to define the region to extend far enough that any common SNPs beyond the boundaries would be minimally correlated to the focal SNP . This requires estimates of recombination rates across the region for a sufficiently similar population. Given

those estimates, one can specify the region by starting from the focal SNP, extend the boundaries until the genetic distance, which is the sum of recombination rate across the interval, exceeds some threshold (performed both to the left and right of the focal SNP).

For this fine mapping study we employed a combined strategy across the three approaches. Our approach was deliberately generous in terms of boundary placement, as we wanted to minimise the chance of missing any potential signal. While the recombination estimates and correlations were fairly robust as both were calculated using HapMap data, the available association signal data varied across phenotypes. There was no genome-wide data available for GD, and so we could not confidently assess how far the signal stretched for the three GD regions. For T2D we generally had strong association signals, anddense genotype data in a substantial number of samples available. However, for CAD there were several regions in which the signal was very weak or effectively non-existent in the initial WTCCC GWAS (this is likely to be due at least in part to the number of samples and the modest effect size in some CAD regions). Given these limitations, our strategy was to first choose a set of boundaries using recombination information, and then adjust those boundaries as necessary based on correlation and association signal. The boundaries were chosen to be a distance of at least 0.1 centimMorgans both upstream and downstream from the focal SNP. We then checked if any SNPs outside these boundaries had an $r^2 > 0.2$ to the focal SNP and expanded the boundaries to include any such SNPs. The boundaries were then further adjusted outwards if necessary to include any SNPs with a p-value within 2 orders of magnitude of the p-value of the focal SNP, where this information was available. Generally, the initial boundaries chosen solely based on recombination ended up

as the final boundaries. Table 2.1 includes information on each of the regions.

2.5 SNP Selection

After defining a region, the next step was to select which variants to assay. We limit the discussion of SNP selection methods here to the simple and conservative approach of comprehensively genotyping any sites reported to exhibit variation in an appropriate population. There are a number of more sophisticated approaches, such as using public reference genotype data (i.e. HapMap) together with GWAS data to increase efficiency by reducing the set of variants to genotype, perhaps by avoiding typing of pairs of SNPs which are apparently perfectly correlated. Given the relative lack of knowledge about the structure and properties of causal alleles in regions found via association with common variants, the more conservative approach of comprehensively typing all known variants in each region seemed far safer than a strategy aimed at genotyping efficiency. Indeed, one of the goals of the study was to investigate exactly this sort of question through improved understanding of the structure of the underlying association.

2.5.1 SNP Sources

While public databases of genetic variation have grown exponentially in size in the past several years, they do not as yet contain a complete catalogue of common variation across the genome. The most prominent of these databases is dbSNP, which is comprised of reports of variation submitted by researchers worldwide. The level of detail in these reports ranges from a simple report of observed variation by a single researcher, to complete genotype data on hundreds of samples

Table 2.1: SNP counts across regions, with sources and stages reflected by columns.For each SNP source, counts are shown for the total number of SNPs, the number successfully assay-designed and genotyped, the number that passed QC filters, and the number of SNPs that were common in our samples (Here we define common as $]MAF > 0.01$).

Phenotype	Chromosome	Start (Mb)	End (Mb)	Gene/Symbol	Affymetrix 500k				Hapmap				Reseq				dbSNP			
					Total	Genotyped	Pass QC	Common	Total	Genotyped	Pass QC	Common	Total	Genotyped	Pass QC	Common	Total	Genotyped	Pass QC	Common
CAD	1	109.54	109.65	*SORT1*	49	19	17	13	76	75	50	50	141	77	42	22	539	239	173	127
	9	21.92	22.13	*CDKN2A/B*	107	40	37	30	176	174	128	128	691	443	184	64	984	464	314	206
	10	44.01	44.15	*CXCL12*	115	42	40	33	161	159	111	111	74	35	23	16	957	416	303	238
	1	220.78	221.04	*1q41*	68	23	23	22	11	11	6	6	109	56	28	25	970	399	305	266
	2	226.73	226.91	*2q36*	67	25	24	18	100	96	72	72	378	216	121	41	691	330	223	138
T2D	16	52.35	52.41	*FTO*	42	15	14	13	75	74	53	52	356	199	98	59	313	128	97	88
	10	94.19	94.49	*HHEX*	48	18	16	14	152	145	115	114	974	564	248	162	1019	457	309	253
	6	20.63	20.84	*CDKAL1*	84	30	28	26	241	229	167	167	573	364	135	74	1115	503	347	265
	10	114.71	114.82	*TCF7L2*	42	14	14	14	53	53	34	34	75	32	24	19	369	159	118	92
	7	28.01	28.23	*JAZF1*	101	37	34	30	157	152	104	104	63	34	15	14	873	369	290	214
GD	2	204.38	204.53	*CTLA-4*	91	35	33	23	135	133	104	104	59	33	13	13	754	322	246	186
	10	6.06	6.17	*CD25/IL2RA*	97	38	32	27	111	111	71	71	17	13	2	2	1222	566	391	265
	1	155.81	156.09	*FCRL3*	125	46	44	35	206	200	131	130	43	28	10	5	1637	743	552	342
				Totals:	1036	382	356	298	1654	1612	1146	1143	3553	2094	943	516	11443	5095	3668	2680

from multiple populations. As such, many sites listed as exhibiting variation may turn out to be monomorphic when genotyped (though this can also be due to population differences or low frequency). Additionally, an appreciable proportion of variable sites are not present in dbSNP, particularly in the lower end of the frequency spectrum. One way to address this is to resequence the region of interest in order to find variants not present in the public datasets.

The variability in the scope of reports in dbSNP presented a challenge. If the stated plan for the study was to genotype "all known variants" in each region, how does one define "all known variants"? There is no doubt that many reports of variants in dbSNP are inaccurate; one need only look at the substantial number of cases in which multiple reported "SNPs" are folded into a single SNP (Sherry et al., 2001a).

This is not surprising given, among other things, the many difficulties involved in resequencing, as well as the tendency of small structural variants ("indels") to look like one or more SNPs. It became clear that while dbSNP would be hugely useful in identifying SNPs to type, we could not simply type all sites reported to vary in dbSNP, as a large portion of sites would turn out to be monomorphic. There are many criteria upon which dbSNP can be filtered, including number of distinct reports of variation, number of samples in which variation is observed, whether actual genotypes or frequency data have been uploaded and whether any of the aforementioned information was generated by the HapMap project. In the end, we took forward any SNP in dbSNP which had genotype or frequency data showing variation and any which had been reported by more than one group. We felt this provided a good balance between limiting the number of non-varying sites that we typed, while not excessively restricting ourselves to sites which had

already been characterized and thus less likely to provide fresh insight into the structure of these regions.

WTCCC Resequencing Project

Prior to the fine mapping study, the WTCCC embarked on a regional resequencing project. The goal of this project was to assess how effectively GWAS associated regions could be resequenced, and how useful this data would be towards characterizing the underlying signal in those regions. In all, sixteen regions were chosen from nine diseases in the original WTCCC study, with resequencing performed on samples used in the HapMap project. Of these regions, five overlap with the fine mapping project. This provided a potentially useful source of previously unknown SNPs in those regions that could be typed in the fine mapping study. Indeed, in choosing sites to genotype for the fine mapping project, we included any novel sites identified in the resequencing project (WTCCC, 2012).

The regions sequenced and their properties are shown in Table 2.2.

Combining across sources

In addition to dbSNP and the WTCCC resequencing study, we drew upon several other sources for sites to type. These include SNPs from the publicly available Craig Venter and James Watson genomes, which were found using comparisons with the human genome reference sequence. We also included SNPs found in regional resequencing by several collaborators. While this provided a rich set of variants to potentially type, we were also presented with an additional informatics challenge. The issue of reference genome build cropped up a number of times, which often occurs in projects assaying genomic data and results involving dis-

Region	Associated Diseases	Chrom. (Mb)	Start (b36)	End (b36)	Focal SNP	MAF (CEU)	Length (kb)	Recombination rate (cM/Mb)	STS Coverage	Mapped bases	10x coverage
IL23R	AS	1	67.37	67.54	rs11209026	0.067	170	1.12	87.6	89.4	71.3
SORT1	CAD	1	109.589	109.64	rs4970834	0.28	51	3.14	96.7	95.6	62.3
IFIH1	T1D	2	162.745	163.03	rs3788964	0.142	285	0.56	98.1	97.8	83.3
2q36	CAD	2	226.731	226.893	rs2943634	0.358	161	0.31	100	100	87.1
MAP3K1	BC	5	56.017	56.315	rs889312	0.308	298	0.77	99.6	99.3	79.8
ARTS1	AS	5	96.108	96.224	rs30187	0.3	116	0.52	97.8	97.9	88.5
MST150	CD	5	150.15	150.31	rs1000113	0.033	160	0.31	85.5	85.4	65.3
6*q*23.3	RA	6	137.99	138.09	rs6920220	0.175	100	1	100	100	90
CDKN2A/B	CAD/T2D	9	21.923	22.131	rs1333049	0.492	208	3.27	89.2	85.5	67.2
HHEX	T2D	10	94.191	94.491	rs5015480	0.432	300	0.57	90.9	91.7	54.3
NKX23	CD	10	101.260	101.32	rs10883365	0.5	60	3.67	89.2	87.3	71.6
LSP1	BC	11	18.23	20.0	rs3817198	0.342	177	2.66	61.3	59.4	22.3
KIAA0350	T1D	16	109.0	11.25	rs2542151	0.192	350	1.29	99.2	98.7	84.2
FTO	T2D/OB	16	52.354	52.406	rs9939609	0.45	52	3.65	100	100	82.8
PTPN2	T1D/CD	18	12.722	12.932	rs12708716	0.288	210	1.52	90.3	88.4	59.7
JSRP1	ATD	19	2.2	2.258	rs7250822	0.025	58	4.31	55.3	55.2	5
Average							172.3	1.79	90	89.5	67.2

Table 2.2: Regions sequenced as part of the WTCCC resequencing project. The diseases listed in the column "associated diseases" are: AS: Ankylosing spondylitis, CAD: coronary artery disease, T1D: type I diabetes, BC: breast cancer, CD: Crohns disease, RA: rheumatoid arthritis, T2D: type II diabetes, OB: obesity, ATD: autoimmune thyroid disease. For each region the start and end coordinates are shown in Mb (build 36 coordinates). The focal SNP from the original finding, as well as that SNPs Minor allele frequency in HapMap data. Length corresponds to the size of the region in kilobases, and Recombination rate across the region is then shown. STS coverage refers to the percent of bases with at least one successful PCR amplicon at design stage. Mapped bases refers to the percent of bases with at least one read mapping in one individual. 10x coverage refers to the average percent of bases with at least 10x coverage.
From: WTCCC (2012). Bayesian refinement of association signals for 14 loci in three common diseases. *Nature Genetics* - in press.

parate sources of data and multiple research groups. Physical positions frequently shift subtly between versions of the Human genome reference, and the difficultly in updating data annotation results in differences in build version between sources. This difficulty combined with differences in SNP-id schemas meant much care was needed to ensure that the final list of regions and variants were consistently and correctly annotated. Table 2.1 displays information on SNP sources. Individual SNPs may be counted more than once if they appear in multiple categories.

2.6 Quality Control

Quality control is a key step in any genetic study, and particularly so in fine-mapping. If we are to have any success in teasing apart and characterizing highly correlated SNPs, it is critical to have a clear sense of the data quality and ideally to remove any SNPs or samples that performed poorly. Genotype calling was performed in two stages at the Sanger Institute. First, genotypes were called using Illuminus (Teo et al., 2007). Only SNPs which Illuminus called with high confidence were taken forward directly; the remainder underwent manual cluster inspection and re-calling where appropriate. The first QC filter applied was to remove individual genotypes with an Illuminus "confidence score" < 0.2, which is the recommended threshold (Teo et al., 2007). Samples with call rates $< 90\%$ were then filtered out. Sample heterozygosity was calculated, as it is a useful marker for sample failure or contamination, but the structure of the data (a small number of relatively small regions) made it unlikely that we would be able to clearly identify outliers. After applying these filters, we then performed SNP QC. We filtered SNPs based on the following criteria:

- SNP call rate < 0.95
- Hardy-Weinberg p-value < 0.001
- minor allele frequency < 0.001.

Table 2.1 shows SNP counts by region, both pre and post QC.

2.7 Expanded Reference Group

As stated above, all samples were typed at all SNPs on the assay, which presented the opportunity to substantially expand the reference set (this was also the case in the original WTCCC study) (WTCCC, 2007). This is feasible thanks to the independence between T2D and GD, and between CAD and GD. This allows us to include the GD samples as controls when analysing T2D regions as well as CAD regions, and to include both the CAD and T2D samples as controls when analysing the GD regions. We compared the results using only the common controls (58BC + NBS cohorts) to the results using the expanded set just described. The signal patterns did not differ between the two sets, and using just the common controls showed somewhat weaker signal as expected (data not shown). Unless stated otherwise, all analyses that follow were performed using the expanded sample set.

2.8 Single Marker Analysis

Single SNP association tests under the additive model were performed on the QC-cleaned data using SNPTEST (Marchini et al., 2007).

In the later section titled "Results by region", the first figure for each region shows the strength of evidence for association for each region. Strength of evidence is illustrated here using the Bayes Factor (BF) for each SNP. The BF is defined as the ratio of two probabilities: the denominator is the probability of the genotype configuration at that SNP in cases and controls under the null hypothesis that disease status is independent of genotype at that SNP, while the numerator is the probability of the genotype configuration at that SNP in cases and controls under the alternative hypothesis, namely the assumption that the SNP is associated with disease status. Large values of the BF correspond to strong evidence for association.

2.8.1 Identifying Causal Variants

Identifying causal variants is one of the primary goals of fine-mapping. While there are a number of methods for functionally determining which of a set of partially-correlated variants is causal, they are labor intensive, expensive, and only feasible for small numbers of SNPs. As such, any robust information we extract from in-silico analyses is of large utility.

One feature of the Bayesian perspective is that BFs for different SNPs can be compared quantitatively. For example, for a particular region, under specific assumptions about how many causal SNPs are in the set of genotyped SNPs, it is straightforward to calculate for each SNP, the posterior probability that that SNP is causal, in the light of the data from the fine mapping experiment. For definiteness, we will calculate these posterior probabilities under the simple assumption that exactly one of the genotyped SNPs in each region is causal, that we have typed all variation in the region, and that it is equally likely, a priori, to be any of the

genotyped SNPs in the region. We cannot know that these prior assumptions are true for any particular region, and they may well not be. Even so, the resulting probabilities can be thought of as the relative strength of evidence in favour of each of the SNPs studied. Writing BF_k for the BF for SNP k, it can be shown that: ((WTCCC, 2012) (Vukcevik, 2009)):

Posterior probability for SNP $k = \frac{BF_k}{\sum_j BF_j}$.

Let M_0 be the null model, in which the phenotype is independent of all SNPs, and M_i be the model under which SNP i is the sole causal SNP in the region. "Let BF_i be the BF that compares M_i and M_0. Under the assumption that there is exactly one causal SNP, we show that the posterior that a given SNP is causal is proportional to its BF. By Bayes' Theorem," (WTCCC, 2012)

$$\Pr(M_i \mid data, M) = \frac{\Pr(data \mid M_i, M)\Pr(M_i \mid M)}{\Pr(data \mid M)} \tag{2.1}$$

$$= \frac{1}{k}\frac{\Pr(data \mid M_i)}{\Pr(data \mid M)} \tag{2.2}$$

$$= \frac{1}{k}\frac{\Pr(data \mid M_i)/\Pr(data \mid M_0)}{\Pr(data \mid M)/\Pr(data \mid M_0)} \tag{2.3}$$

$$= \frac{BF_i}{kBF_{\text{reg}}} \propto BF_i. \tag{2.4}$$

(WTCCC, 2012)

In each region one can choose the smallest set of SNPs accounting for 95% (respectively 99%) of the cumulative posterior probability. These sets, which we call credible sets, are somewhat analogous to confidence intervals: in a particular region, if the true causal SNP has been genotyped, we can be reasonably confident that it will be in the relevant credible set. The size of these sets give a rough sense

of the success of fine-mapping in localising a region's causal variant. The smaller the set, the closer we have come to identifying the causal variant. A large set suggests that we may not have added much resolution, or that there are multiple variants very highly correlated with the causal variant.

In the later section titled "Region by region results", each regions second figure plots the posterior probability associated with each SNP. SNPs in the 95% credible set are coloured yellow, and those in the 99% credible set which are not in the 95% set are coloured purple. The number of SNPs in each set is also indicated next to the title of the plot.

2.8.2 Imputation

While we attempted to genotype all known SNPs, a non-trivial proportion were not present in our final dataset due to assay design failure, genotyping failure, and QC filtering. All told, we obtained QC passing genotypes for 73% of the SNPs for which we attempted assay design. Given our dense genotyping, we were able to infer data for many of the missing SNPs using imputation. By using a densely typed set of reference haplotypes along with data on recombination, imputation can infer data at many untyped SNPs in a study sample using SNPs which overlap between the study and the reference. We used the tool IMPUTE (Marchini et al., 2007), which uses a population genetics model-based approach to imputation. For the reference panel we used 286 haplotypes from the CEU, GBR, IBS and TSI 1000 Genomes panels (June 2011 release), all of european ancestry. Given the substantial overlap of SNPs between the reference panel and our data, we were able to impute genotypes at a substantial number of failed SNPs with a high degree of accuracy. We filtered the output of impute using the recommended cut-offs of

< 0.9 for average maximum posterior,< 0.4 for imputation confidence and > 0.1 for missing data proportion (Marchini et al., 2007). Across all regions, a total of 1,943 polymorphic SNPs were imputed confidently. In the section titled "Region by region results", the third (bottom) panel adds the set of imputed SNPs to the regional results shown in the top panel. Imputed SNPs are shown as green. For further discussion of this imputation and various analyses of the imputed data see the next chapter.

2.9 Results by Region

2.9.1 Coronary Artery Disease (CAD)

Coronary artery disease (CAD) occurs when the arteries and vessels that provide oxygen to the heart are narrowed or blocked by an accumulation of fatty materials on the inner linings of arteries (atherosclerosis). This restricts the blood flow to the heart and when the blood flow is completely cut off, the result is a heart attack (or myocardial infarction), which is the clinical manifestation of CAD (Maouche and Schunkert, 2012). We fine mapped five regions for CAD, all detected through GWAS efforts.

SORT1

This locus was not highlighted for CAD in the initial WTCCC study (rs599839 chr1: 109534208, risk allele A with frequency 0.77 ; $OR(95\%CI) : 1.24(1.12, 1.38)$ and $p = 2.19 \times 10^{-5}$) but it did reach significance in the follow up study of CAD: $OR(95\%CI) : 1.29(1.18, 1.4)$ and $p = 4.05 \times 10^{-9}$ (Samani et al., 2007).

The SNPs in the 1p13 locus are located in a noncoding DNA region between two genes of unknown function: *CELSR2* and *PSRC1*.

It was first associated with low-density lipoprotein cholesterol (*LDL-C*), a causal risk factor for MI and CAD (Willer et al., 2008) (Kathiresan et al., 2008) and this locus has the strongest association with LDL-C of any locus in the genome (beta:-5.65, $p = 1 \times 10^{-170}$) (Teslovich et al., 2010). It was associated with early onset MI (MIGC, 2009). The lead SNP (rs646776) is 3636 bp away from the initial WTCCC marker for which it is also a very good proxy ($r^2 = 0.895$, $D' = 1.000$) at position 109,530,572. The T risk allele had frequency of 0.81 and $OR(95\%CI) : 1.18(1.12 - 1.25)$ $p = 9.36 \times 10^{-11}$.

The locus provides a useful positive control for our analysis, because the functional variant has subsequently been identified (Musunuru et al., 2010) (independently of our genotyping experiment).

The minor allele of rs646776 associated with mRNA levels in liver of three genes in and around the associated region: *CELSR2*, *PSRC1* and *SORT1*. The functional variant, rs12740374 (940 bp away from the MI/lipid lead SNP for which it is a perfect proxy, and 4576 bp from the WTCCC lead SNP, has been shown to create a *C/EBP* (*CCAAT*/enhancer binding protein) transcription factor binding site which affects the expression of *SORT1* which in turn affects intracellular apolipoprotein B (apoB) processing (Musunuru et al., 2010). We designed genotyping assays for the functional SNP in our assay, but unfortunately the SNP failed genotyping QC. Figure 2.2 shows the result of imputation from the 1000 Genome data set (June 2011 release) into our fine mapping data in this region. With CAD as the phenotype of interest, the imputed version of the functional SNP is the sixth strongest association signal, with non-trivial, but not large, poste-

rior probability. Our data do not allow an assessment of the relative contributions of imputation error, and our use of an indirect phenotype, to the failure to see a large posterior probability.

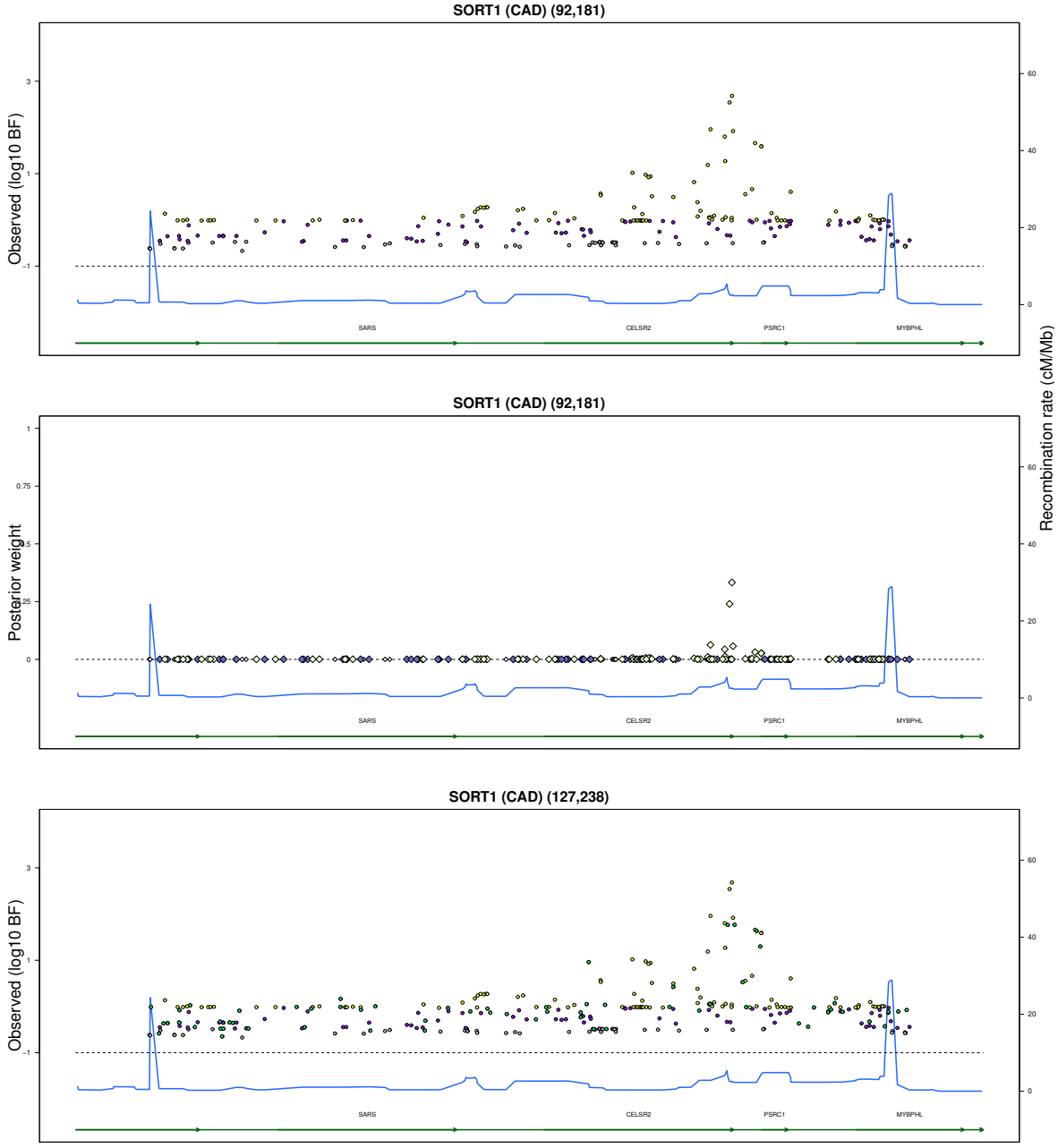


Figure 2.2: Regional results. All panels have genomic position on the x-axis. Recombination rate (cM/Mb) is shown in blue, with the scale on the right y-axis. SNP counts for 95% and 99% sets are in the title of each panel. a) Additive log_{10} Bayes Factors for each SNP in the *SORT1* region for CAD. SNPs in the 95% credible set are coloured yellow and SNPs exclusively in the 99% set are coloured purple. The left y-axis is log_{10} BF scale. b) The posterior probability associated with each SNP in the *SORT1* region for CAD. SNPs in the 95% credible set are coloured yellow and SNPs exclusively in the 99% set are coloured purple. c) Single marker additive log_{10} Bayes Factors across the *SORT1* region for CAD including imputed SNPs. Imputed SNPs are coloured green. The left y-axis is log_{10} BF scale.

CDKN2A/B (CAD)

This region on chromosome 9p21.3 contains the most notable new finding for CAD in the original WTCCC study. The strongest signal is observed at rs1333049 (chr9: 22115503. Risk allele C $MAF = 0.53$, OR (95% CI) 1.37 $(1.27 - 1.49)$ $p = 1.8 \times 10^{-14}$), but the association spans approximately 100 kb around the SNP.

This region has not been associated with lipids hereto but has been shown to be associated with risk of abdominal aortic (rs10757278 $OR = 1.31$, $p = 1.2 \times 10^{-12}$) and intra-cranial aneurysms (rs10757278 $OR = 1.29$, $p = 2.5 \times 10^{-6}$). rs10757278 is 1026 bp away from and in perfect LD with the WTCCC SNP ($r^2 = 1.00$, $D = 1.00$) (Helgadottir et al., 2008, Low et al., 2012). This region is also associated with MI (rs4977574, chr9: 22,088,574, risk allele G MAF = 0.56, OR (95% CI): 1.28 $(1.24 - 1.33)$, $p = 1.08 \times 10^{-41}$) and the SNP is 27kb away but in high LD with the WTCCC lead marker ($r^2 = 0.967$, $D = 1.000$). It is also associated with T2D (see below) and suggested to be associated to several cancers, frailty in the elderly (Murabito et al., 2012), and glaucoma (Wiggs et al., 2012).

This locus contains two cyclin dependent kinase inhibitors genes, *CDKN2A* (encoding *p16INK4a*) and *CDKN2B* (encoding *p15INK4b*) which both play a role of cell cycle regulation. The only other known gene in the region is *MTAP*, which encodes methylthioadenosine phosphorylase. This is an enzyme that contributes to polyamine metabolism (Schmid et al., 2000).

The best SNP after the fine mapping is rs1537370 ($RR = 1.40$; 95% $CI = 1.29 - 1.51$), accounting for 26% of the posterior weight. There are 12 other SNPs in the 95% credible set.

All SNPs in the 99% credible set for this region are within the 3 half of *ANRIL*

and within a region showing weak evidence of selective sweep in the Bantu samples of HGDP. Our top SNP (rs1537370) appears to be in a nuclease accessible domain and has evidence for polymerase II binding. It also has evidence of histone modifications associated with active transcription including H3K4me3 and H3K9ac in human mammary epithelial cells (HMEC) and H3K27ac in HMEC, normal human epithelial keratinocytes (NHEK) and normal human lung fibroblasts (NHLF). Perhaps of interest is the fact that four SNPs (rs1333045, rs10757278, rs10757279, rs10757277) in the 95% credible set appear to be in regions that bind *TCF7L2* and one shows some evidence of conservation (rs10757278).

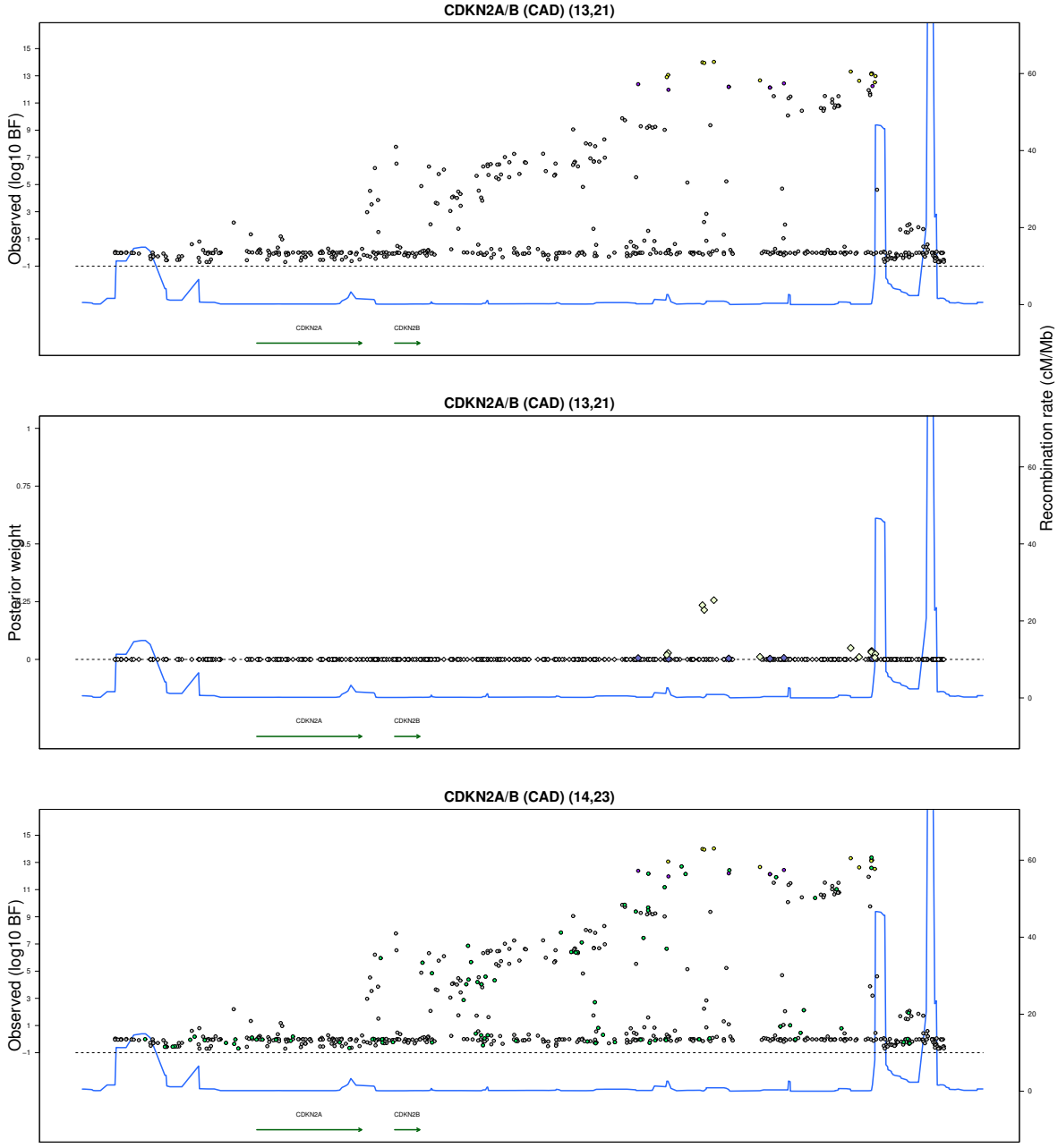


Figure 2.3: Regional results. All panels have genomic position on the x-axis. Recombination rate (cM/Mb) is shown in blue, with the scale on the right y-axis. SNP counts for 95% and 99% sets are in the title of each panel. a) Additive log_{10} Bayes Factors for each SNP in the *CDKN2A/B* region for CAD. SNPs in the 95% credible set are coloured yellow and SNPs exclusively in the 99% set are coloured purple. The left y-axis is log_{10} BF scale. b) The posterior probability associated with each SNP in the *CDKN2A/B* region for CAD. SNPs in the 95% credible set are coloured yellow and SNPs exclusively in the 99% set are coloured purple. c) Single marker additive log_{10} Bayes Factors across the *CDKN2A/B* region for CAD including imputed SNPs. Imputed SNPs are coloured green. The left y-axis is log_{10} BF scale.

CXCL12

The association of a locus at $10q11.21$ with CAD was not significant in the WTCCC (rs501120 position 44073873 bp, T allele frequency 0.87, $OR(95\%CI) = 1.24(1.09, 1.41)$, $p = 1.31 \times 10^{-3}$). However, a close to significant association was observed in the combined analysis of the WTCCC and German MI I studies ($OR(95\%CI) = 1.33(1.2 - 1.48)$, $p = 9.46 \times 10^{-8}$) (Samani et al., 2007).

This locus has been supported by at least one further MI study and is now significantly associated (rs1746048; 44,095,830 bp, C risk allele with frequency 0.86, combined $OR(95\%CI) = 1.19(1.13 - 1.25)$, $p = 8.14 \times 10^{-11}$). rs1746048 is a perfect proxy of the initially described WTCCC marker rs501120 ($r^2 = 1.000$, $D' = 1.000$, 21957 Bp away) (MIGC, 2009) (CADC, 2009). There is some evidence that the effect of the locus on CAD risk may be stronger in women than in men (CADC, 2009), although this requires further confirmation.

The association signal cluster is about 100 kb downstream of the *CXCL12* gene, which is the only coding gene in a 500 kb window of the association. It encodes stromal cell-derived factor-1 (*SDF-1*), which is a chemokine that plays a key role in stem-cell homing and tissue regeneration in ischemic cardiomyopathy (Askari et al., 2003) and in promoting angiogenesis through recruitment of endothelial progenitor cells (Zheng et al., 2007).

Our fine mapping data failed to refine the signal in this region. The top SNP (rs34161818) accounted for 7% of the posterior weight, and 266 SNPs were included in the 95% credible set.

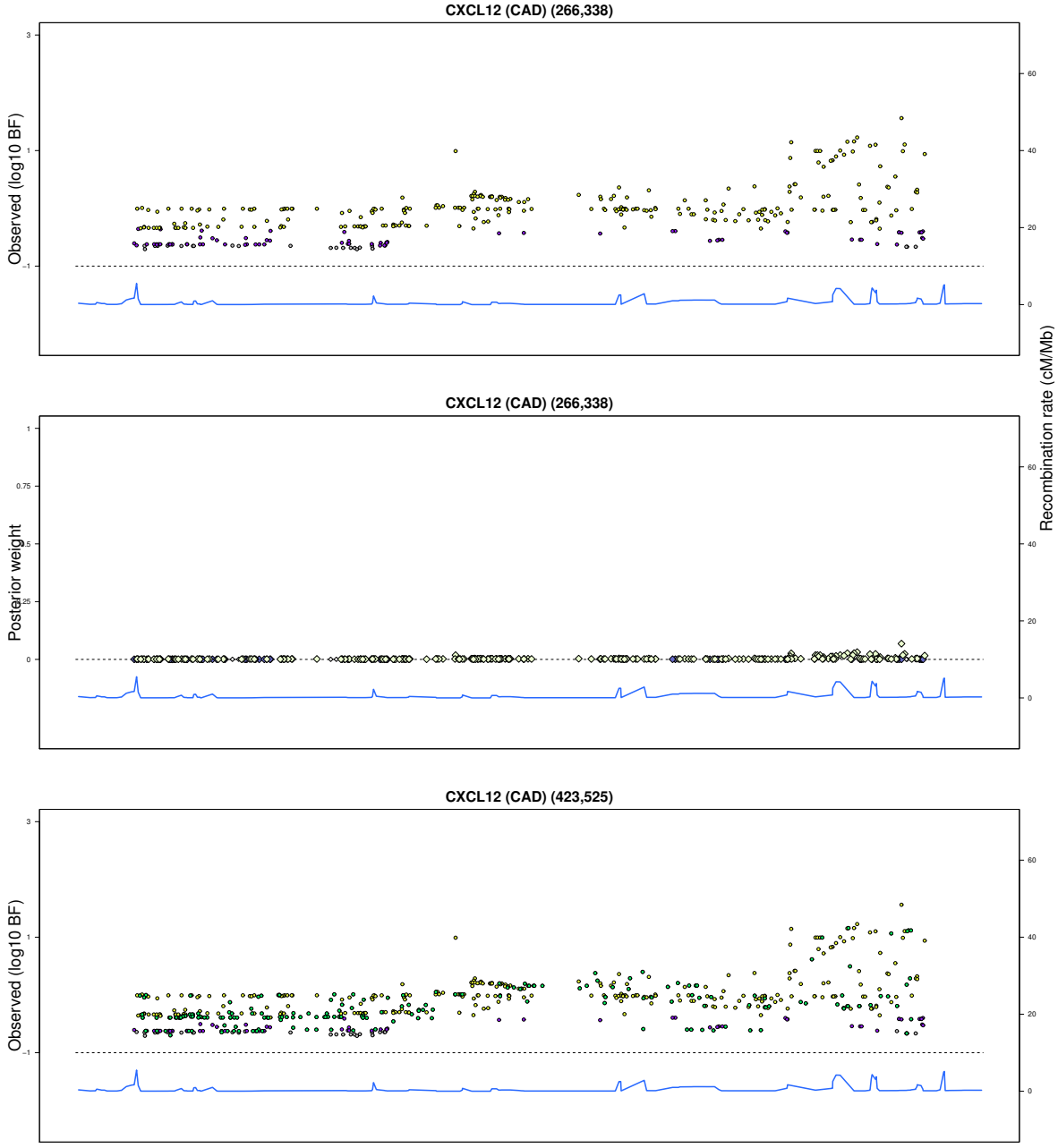


Figure 2.4: Regional results. All panels have genomic position on the x-axis. Recombination rate (cM/Mb) is shown in blue, with the scale on the right y-axis. SNP counts for 95% and 99% sets are in the title of each panel. a) Additive log_{10} Bayes Factors for each SNP in the *CXCL12* region for CAD. SNPs in the 95% credible set are coloured yellow and SNPs exclusively in the 99% set are coloured purple. The left y-axis is log_{10} BF scale. b) The posterior probability associated with each SNP in the *CXCL12* region for CAD. SNPs in the 95% credible set are coloured yellow and SNPs exclusively in the 99% set are coloured purple. c) Single marker additive log_{10} Bayes Factors across the *CXCL12* region for CAD including imputed SNPs. Imputed SNPs are coloured green. The left y-axis is log_{10} BF scale.

1q41

The nominal association signal for CAD on the locus on chromosome 1q41 (rs17465637 at position 220,890,152, risk allele C, $MAF = 0.71$, $OR(95\%CI) = 1.23(1.12 - 1.34)$, $p = 1.00 \times 10^{-5}$) has been confirmed when incorporating additional studies (rs17465637; chr1 220,890,152bp, risk allele C with frequency 0.72. $OR(95\%CI) = 1.13(1.09 - 1.19)$, $p = 1.33 \times 10^{-8}$) (MIGC, 2009) (CADC, 2009).

The most strongly associated SNPs lie within the melanoma inhibitory activity family, member 3 (*MIA3*) gene. This gene has not been very well characterized functionally but may play a role in cell growth or inhibition (Bosserhoff and Buettner, 2002). The association signal include at least four other coding genes: *TAF1A*, *AIDA*, *C1orf58* and *FAM177B*, non of which are very well characterized.

Our fine mapping data failed to refine the signal in this region. The top SNP (rs2936023) accounted for 4% of the posterior weight, and 240 SNPs were included in the 95% credible set.

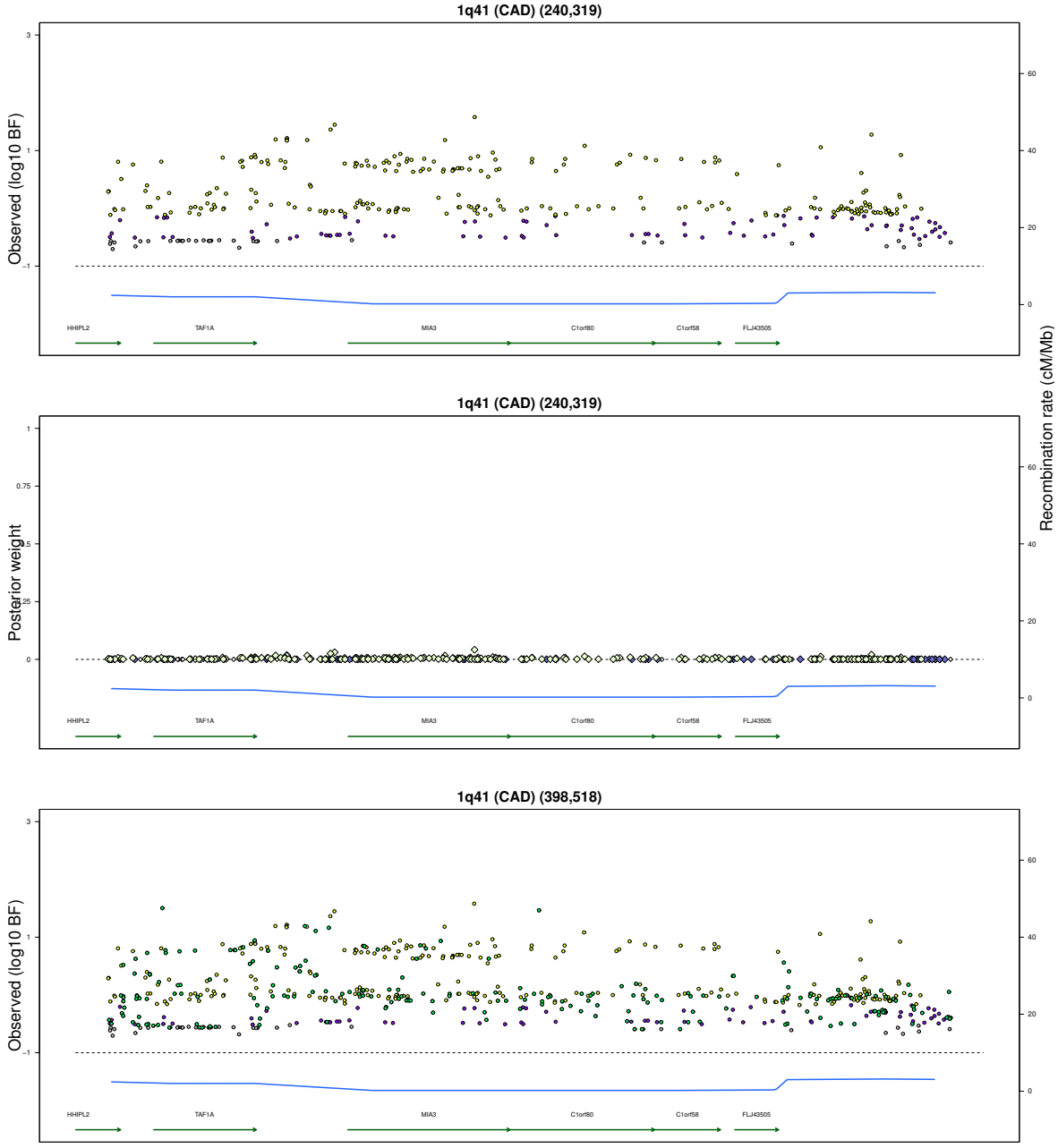


Figure 2.5: Regional results. All panels have genomic position on the x-axis. Recombination rate (cM/Mb) is shown in blue, with the scale on the right y-axis. SNP counts for 95% and 99% sets are in the title of each panel. a) Additive log_{10} Bayes Factors for each SNP in the *1q41* region for CAD. SNPs in the 95% credible set are coloured yellow and SNPs exclusively in the 99% set are coloured purple. The left y-axis is log_{10} BF scale. b) The posterior probability associated with each SNP in the *1q41* region for CAD. SNPs in the 95% credible set are coloured yellow and SNPs exclusively in the 99% set are coloured purple. c) Single marker additive log_{10} Bayes Factors across the *1q41* region for CAD including imputed SNPs. Imputed SNPs are coloured green. The left y-axis is log_{10} BF scale.

2q36

The chromosome 2q36 association signal for CAD was nominally significant in the WTCCC study (rs2943634- 226893585 bp, risk allele C with frequency 0.66, $OR(95\%CI) = 1.22(1.11-1.33)$, $p = 1.19 \times 10^{-5}$). In the combined analysis of the WTCCC and German MI Study it was also nominally significant ($OR(95\%CI) = 1.20(1.12-1.3)$, $p = 1.27 \times 10^{-6}$) (WTCCC, 2007) and after adding in further studies (MIGC, 2009) (CADC, 2009) there is still some evidence of association of the locus with CAD (rs2943634 on 2q36 (OR $(95\%CI) = 1.05(1.01-1.10)$, $p = 0.01$) albeit no definite confirmation.

There are no coding genes within a 500 kb window around the lead SNP in this region. Our fine mapping data failed to refine the signal in this region. The top SNP (rs2673145) accounted for 6% of the posterior weight, and 86 SNPs were included in the 95% credible set.

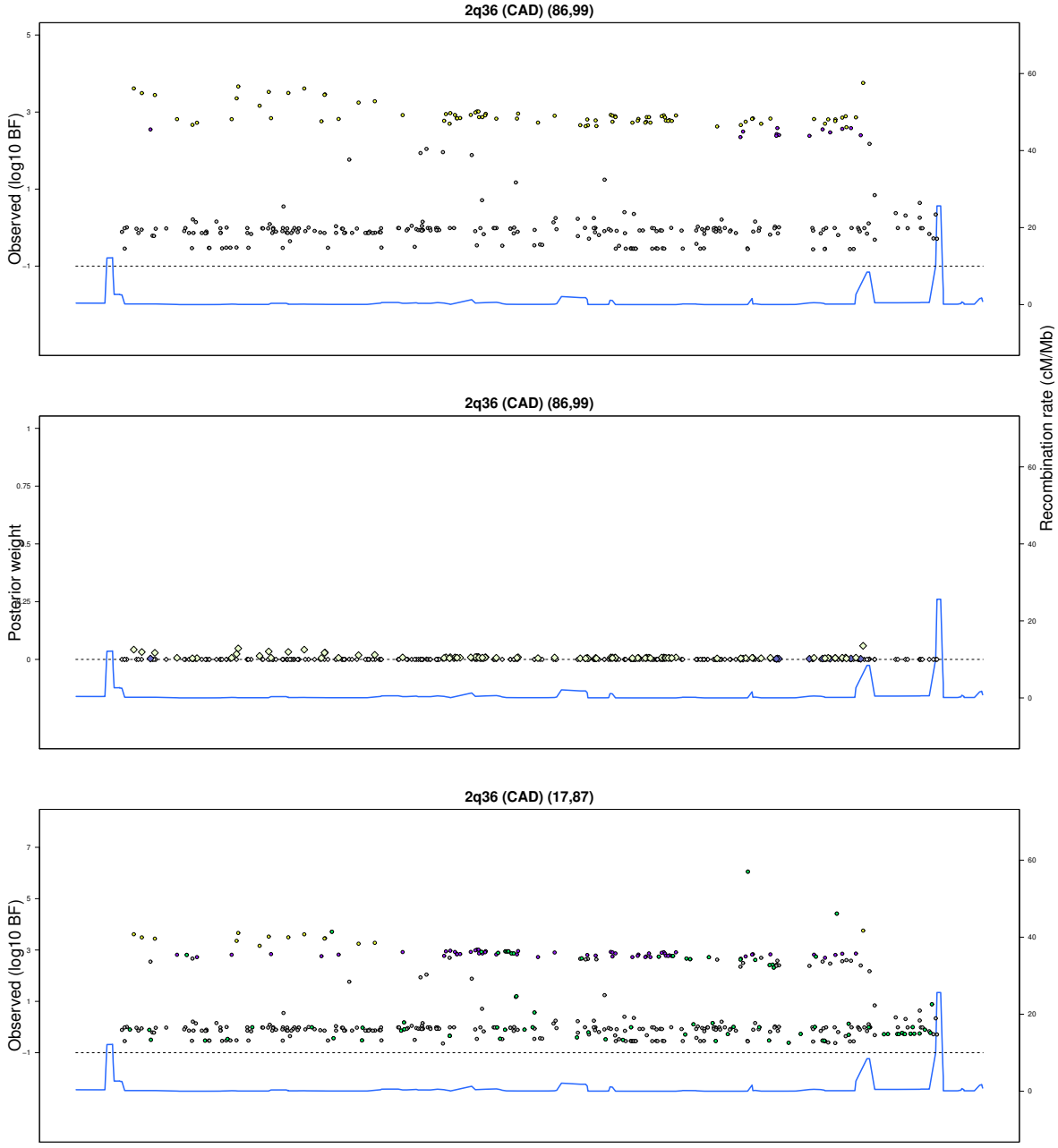


Figure 2.6: Regional results. All panels have genomic position on the x-axis. Recombination rate (cM/Mb) is shown in blue, with the scale on the right y-axis. SNP counts for 95% and 99% sets are in the title of each panel. a) Additive log_{10} Bayes Factors for each SNP in the *2q36* region for CAD. SNPs in the 95% credible set are coloured yellow and SNPs exclusively in the 99% set are coloured purple. The left y-axis is log_{10} BF scale. b) The posterior probability associated with each SNP in the *2q36* region for CAD. SNPs in the 95% credible set are coloured yellow and SNPs exclusively in the 99% set are coloured purple. c) Single marker additive log_{10} Bayes Factors across the *2q36* region for CAD including imputed SNPs. Imputed SNPs are coloured green. The left y-axis is log_{10} BF scale.

2.9.2 Graves' Disease (GD)

Autoimmune thyroid disease (AITD) primarily consists of Graves' Disease (GD) which is characterized by hyperthyroidism, and Hashimoto's Thyroiditis (HT) which is characterized hypothyroidism. AITD is characterized by T and B cell infiltration of the thyroid and by the production of thyroid autoantibodies.

We fine mapped three regions for GD, of which two were candidate regions that were known previously and one detected through GWAS efforts.

CTLA-4

This locus was known before the GWAS era and early studies of *CTLA-4* in Graves' disease (GD) reported association of a limited number of variants, including, a dinucleotide (AT)n repeat within the $3'$ untranslated region of exon 3 (Yanagawa et al., 1995), an A/G SNP within exon 1 (Ala/Thr position 17) (Heward et al., 1999) a C/T SNP within the *CTLA-4* promoter region -318bp from the ATG start codon (Braun et al., 1998) and a C/T SNP within the non-coding intron 1 of *CTLA-4* (Vaidya et al., 2003).

A more detailed fine mapping study genotyped 108 SNPs across 330 kb containing *CD28* , *CTLA-4* and *ICOS* (Ueda et al., 2003). The association (maximum $OR = 1.51$) was refined to within a 6.1 kb region, 3' of *CTLA-4*, containing at least four SNPs *CT60* (rs3087243), *JO31* (rs11571302), *JO30* (rs7565213) and *JO27_1* (rs11571297) all highly associated with GD (Ueda et al., 2003). In 672 cases and 844 controls the rs3087243 A>G SNP was nominally stronger associated with GD than the other markers and the susceptibility allele (A) of rs3087243 was associated with reduced mRNA levels of a soluble *CTLA-4* isoform (Ueda et al., 2003).

Also, in a recent meta-analysis of 10 studies (4,906 GD subjects and controls) and two of the SNPs studied commonly supported findings that rs3087243 is the most associated SNP and calculated a combined $OR = 1.49$ for rs3087243 (Kavvoura et al., 2007). It is noted, however, that accurate genotyping of the functional candidate, the polymorphic (AT)n microsatellite variants near the main polyA site of the *CTLA-4* $3'$ UTR, in large numbers of samples has not been reported, and the entire region of linkage disequilibrium (LD) has not been thoroughly re-sequenced to reveal all the variants that might be strongly associated with GD in the region.

The best SNP after our fine mapping is rs11571297 ($RR = 1.39$; 95% $CI = 1.29 - 1.50$), accounting for 77% of the posterior weight. This SNP is 6 kb away from the initially reported variant (rs3087243) and is in close LD with it ($r^2 = 0.875$ and $D' = 1.000$). It was not typed in many of the studies included the earlier meta-analysis so we cannot compare our results with previous ones. There are five other SNPs in the 95% credible set, with rs3087243, the top SNP from the earlier meta-analysis ranking in a group of three similarly ranked SNPs behind the top SNP. Its posterior weight was 5.2%. rs3087243, located near the 3 end of *CTLA-4* also has some evidence of regulatory potential and conservation.

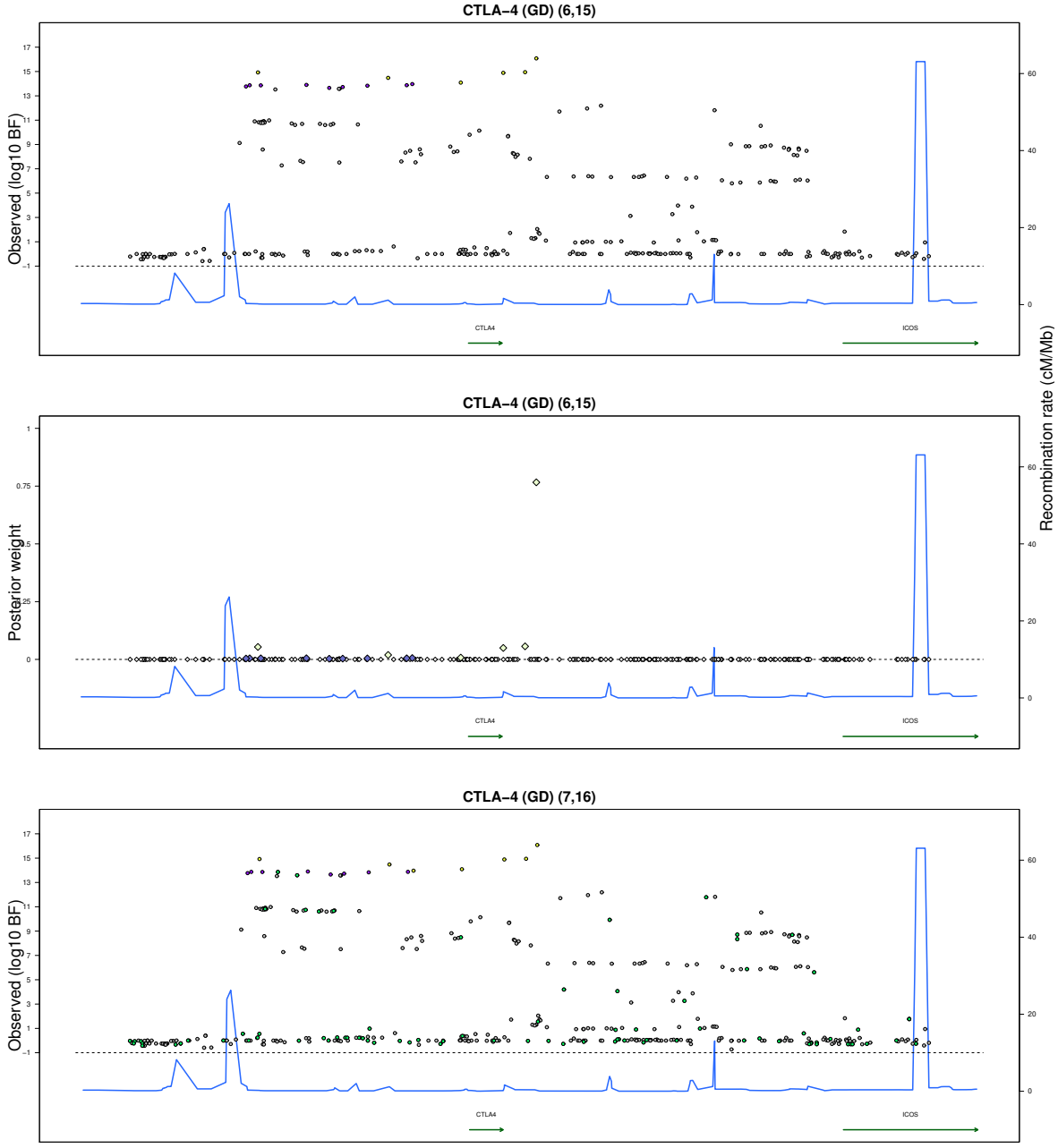


Figure 2.7: Regional results. All panels have genomic position on the x-axis. Recombination rate (cM/Mb) is shown in blue, with the scale on the right y-axis. SNP counts for 95% and 99% sets are in the title of each panel. a) Additive log_{10} Bayes Factors for each SNP in the *CTLA-4* region for GD. SNPs in the 95% credible set are coloured yellow and SNPs exclusively in the 99% set are coloured purple. The left y-axis is log_{10} BF scale. b) The posterior probability associated with each SNP in the *CTLA-4* region for GD. SNPs in the 95% credible set are coloured yellow and SNPs exclusively in the 99% set are coloured purple. c) Single marker additive log_{10} Bayes Factors across the *CTLA-4* region for GD including imputed SNPs. Imputed SNPs are coloured green. The left y-axis is log_{10} BF scale.

CD25/IL2RA

A genetic association between type 1 diabetes (T1D) and the interleukin-2 receptor alpha (*IL-2R*)/*CD25* locus was first identified in a study using tag SNPs and an LD mapping approach (Vella et al., 2005). Further fine mapping in T1D suggested a localization of association to two independent groups of SNPs, a combined odds ratio of 2 and which span two overlapping regions of 14 and 40 kb, including *IL2R* intron 1 and the 5 regions of *IL2R* and *RBM17* (Lowe et al., 2007).

Using the same 20 tag SNPs used in the original T1D study and applying a multilocus test upon a case-control study of 1896 GD cases and 1892 matched controls, evidence for association between GD and the *CD25* region was found ($p = 4.5 \times 10^{-4}$), with the pattern of association similar to that seen in T1D (Brand et al., 2007) with SNPs rs7093069 ($OR = 1.15$; 95% $CI = 1.04 - 1.31$, $P = 2 \times 10^{-3}$) and rs12722592 ($OR = 1.24$; 95% $CI = 1.09 - 1.45$, $P = 6 \times 10^{-3}$). Subsequently, studies in T1D and in multiple sclerosis (MS) have reported evidence of three association signals in the *IL2RA* region, with SNPs, rs41295061, rs11594656 and rs2104286 being required to explain the risk of T1D region, and two of these, rs11594656 and rs2104286, being associated with MS risk (Maier et al., 2009) (Dendrou et al., 2009).

The top SNP after the fine mapping, rs10905669 ($RR = 1.20$; 95% $CI = 1.10 - 1.30$), accounted for 21% of the posterior weight, however there were 76 SNPs in the 95% credible set, so that the experiment failed to exclude many possible SNPs.

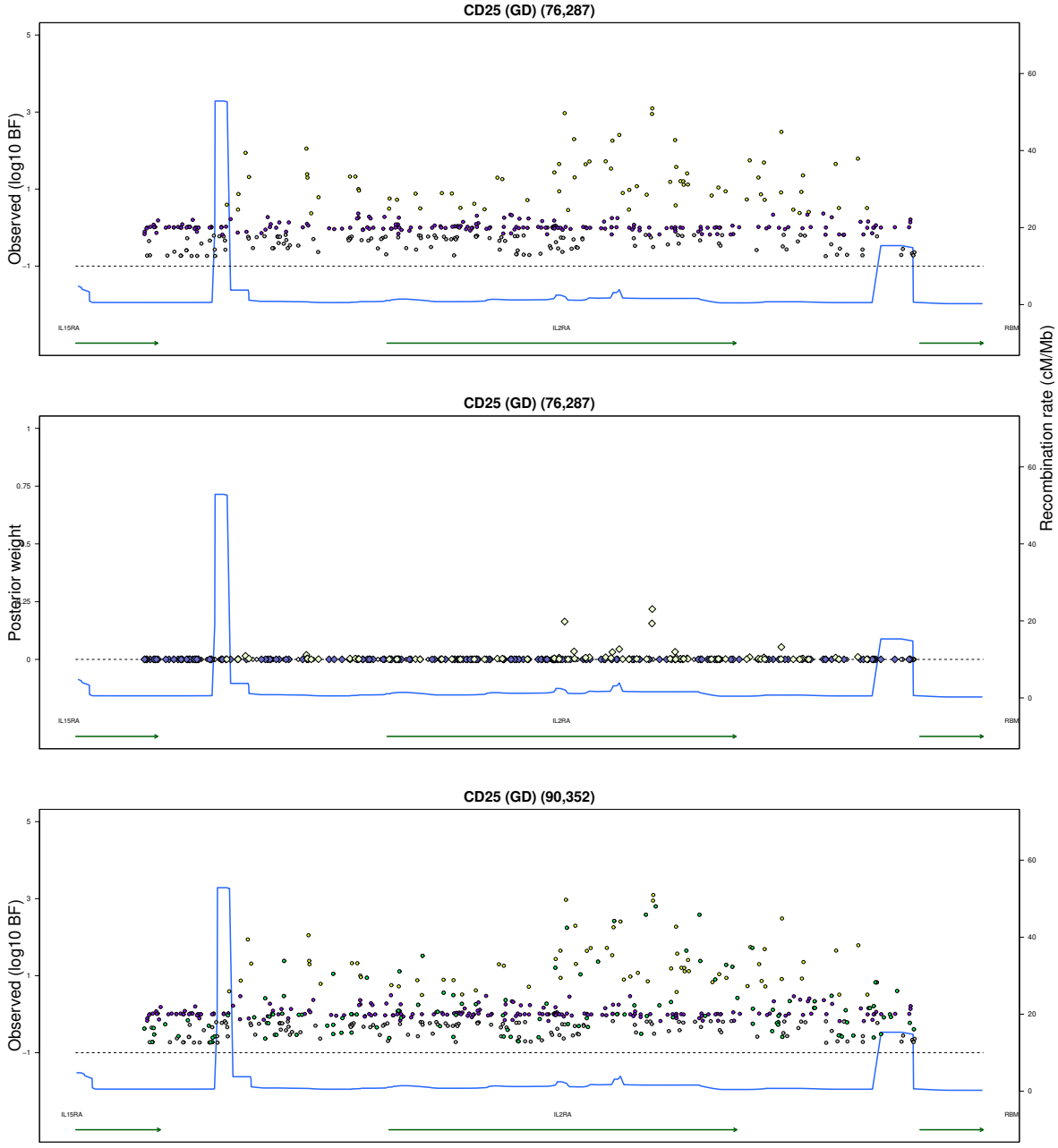


Figure 2.8: Regional results. All panels have genomic position on the x-axis. Recombination rate (cM/Mb) is shown in blue, with the scale on the right y-axis. SNP counts for 95% and 99% sets are in the title of each panel. a) Additive log_{10} Bayes Factors for each SNP in the *CD25/IL2RA* region for GD. SNPs in the 95% credible set are coloured yellow and SNPs exclusively in the 99% set are coloured purple. The left y-axis is log_{10} BF scale. b) The posterior probability associated with each SNP in the *CD25/IL2RA* region for GD. SNPs in the 95% credible set are coloured yellow and SNPs exclusively in the 99% set are coloured purple. c) Single marker additive log_{10} Bayes Factors across the *CD25/IL2RA* region for GD including imputed SNPs. Imputed SNPs are coloured green. The left y-axis is log_{10} BF scale.

FCRL3

Association of the FC receptor like (*FCRL*) region with autoimmune disease was first detected within a Japanese rheumatoid arthritis cohort and with maximum effect seen with SNP rs7528684 within *FCRL3* ($OR = 2.15$). In the same study association of rs7528684 was also reported in a GD cohort ($OR = 1.79$) (Kochi et al., 2005). Subsequently, in a UK Caucasian cohort of 983 GD patients and 733 controls association with rs7528684 was also detected ($OR = 1.17$) but less significant than that seen in the Japanese cohort (Simmonds et al., 2006).

Later, the results from the WTCCC 14,500 nsSNP screen detected association of rs7522061 SNP (which tagged rs7528684, 2.4 kb away and $r^2 = 0.872$, $D = 1.000$).

Seven tag SNPs capturing 11 common variants (MAF > 0.05) within *FCRL3* (including one tag SNP that captured rs7528684) were typed in the National AITD cohort consisting of 5000 Graves cases and controls (Burton et al., 2007). Out of these 7 tags, four SNPs (rs3761959, rs11264794, rs11264798 and rs11264793) were found to be associated with GD ($p = 0.01$ – 1.6×10^{-5}) with rs11264798 now producing the strongest signal ($p = 1.6 \times 10^{-5}$, $OR = 1.22$). Our fine mapping data failed to refine the signal in this region. The top SNP, rs11264798, accounted for 7% of the posterior weight, and 114 SNPs were included in the 95% credible set.

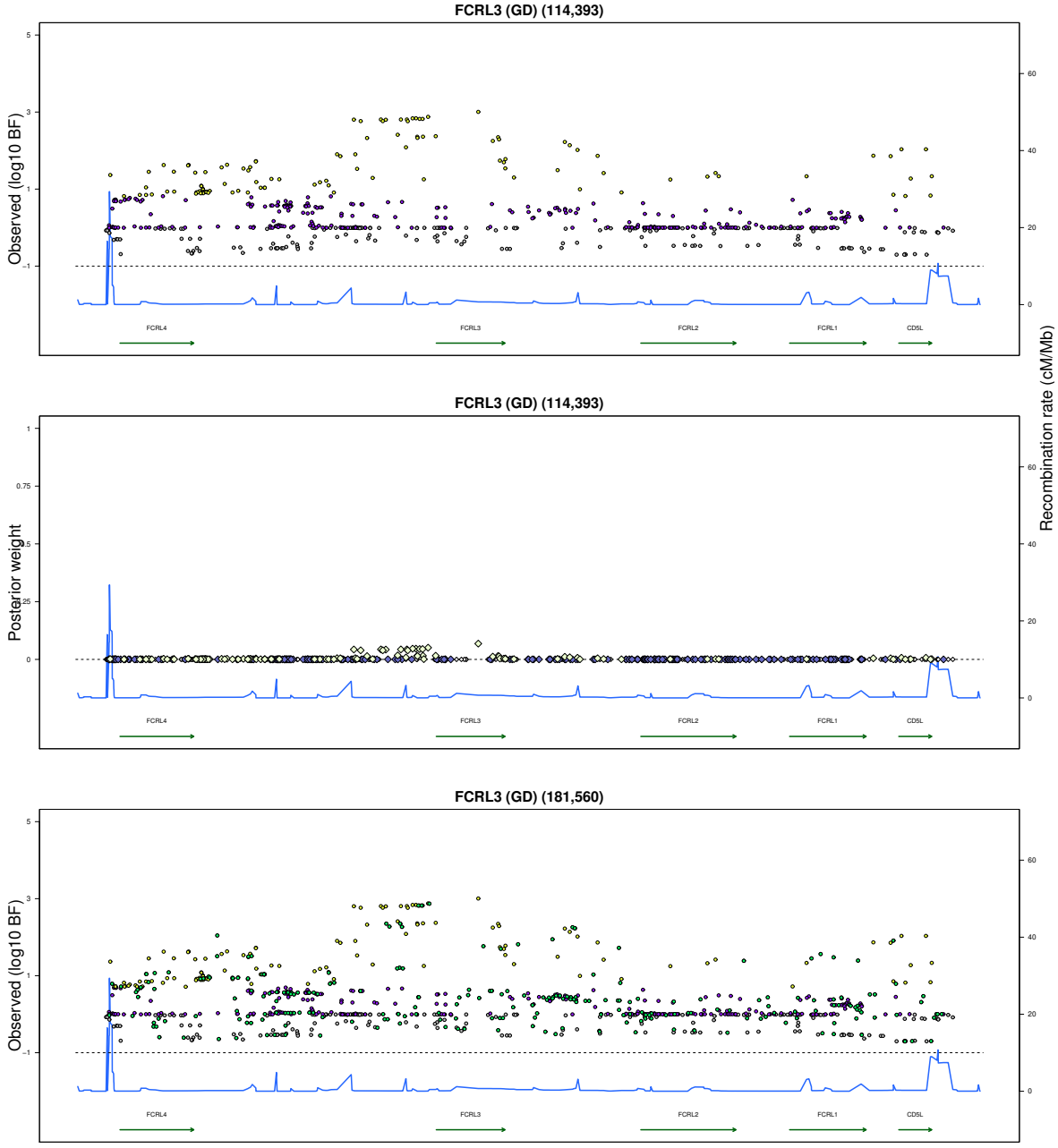


Figure 2.9: Regional results. All panels have genomic position on the x-axis. Recombination rate (cM/Mb) is shown in blue, with the scale on the right y-axis. SNP counts for 95% and 99% sets are in the title of each panel. a) Additive log_{10} Bayes Factors for each SNP in the *FCRL3* region for GD. SNPs in the 95% credible set are coloured yellow and SNPs exclusively in the 99% set are coloured purple. The left y-axis is log_{10} BF scale. b) The posterior probability associated with each SNP in the *FCRL3* region for GD. SNPs in the 95% credible set are coloured yellow and SNPs exclusively in the 99% set are coloured purple. c) Single marker additive log_{10} Bayes Factors across the *FCRL3* region for GD including imputed SNPs. Imputed SNPs are coloured green. The left y-axis is log_{10} BF scale.

2.9.3 Type 2 diabetes (T2D)

CDKN2A/B

The T2D association with variants in the 9p21 region close to *CDKN2A/B* was first identified through the attempts to replicate sub-significant signals of the GWA analysis of WTCCC, DGI and FUSION samples where the region generated a promising signal in all three scans. Replication was observed in all three efforts (rs10811661, risk allele: T. OR (95%CI)=1.19 (1.11-1.28), p=4.9×10^{-7}; DGI samples, OR (95%CI)=1.20 (1.12-1.28), p=5.4×10^{-8}; FUSION: OR (95%CI)=1.20 (1.07-1.36), $p = 2.2 \times 10^{-3}$. All samples combined: OR (95%CI)=1.20 (1.14-1.25), $p = 7.8 \times 10^{-15}$) (WTCCC, 2007) (Zeggini et al., 2007) (Saxena et al., 2007) (Scott et al., 2007).

The variants influencing susceptibility to CAD (rs1333049) (Samani et al., 2007) (Helgadottir et al., 2007) and aneurysm formation (rs10757278 and rs10811661) (Helgadottir et al., 2008) at this locus (see above) are distinct from those involved in T2D ($r^2 = 0.009$, $D' = 0.179$ between).

Based on functional data from mouse (Krishnamurthy et al., 2006), the T2D-association signal is likely to involve gain of function of CDKN2A, potentially mediated via the non-coding RNA *ANRIL* (or *CDKN2BAS*) (Pasmant et al., 2007).

In our analysis, the best T2D-associated SNP after the fine mapping is rs12555274 ($RR = 1.26$; 95% $CI = 1.15 - 1.37$), accounting for 68% of the posterior weight. There are 4 other SNPs in the 95% credible set.

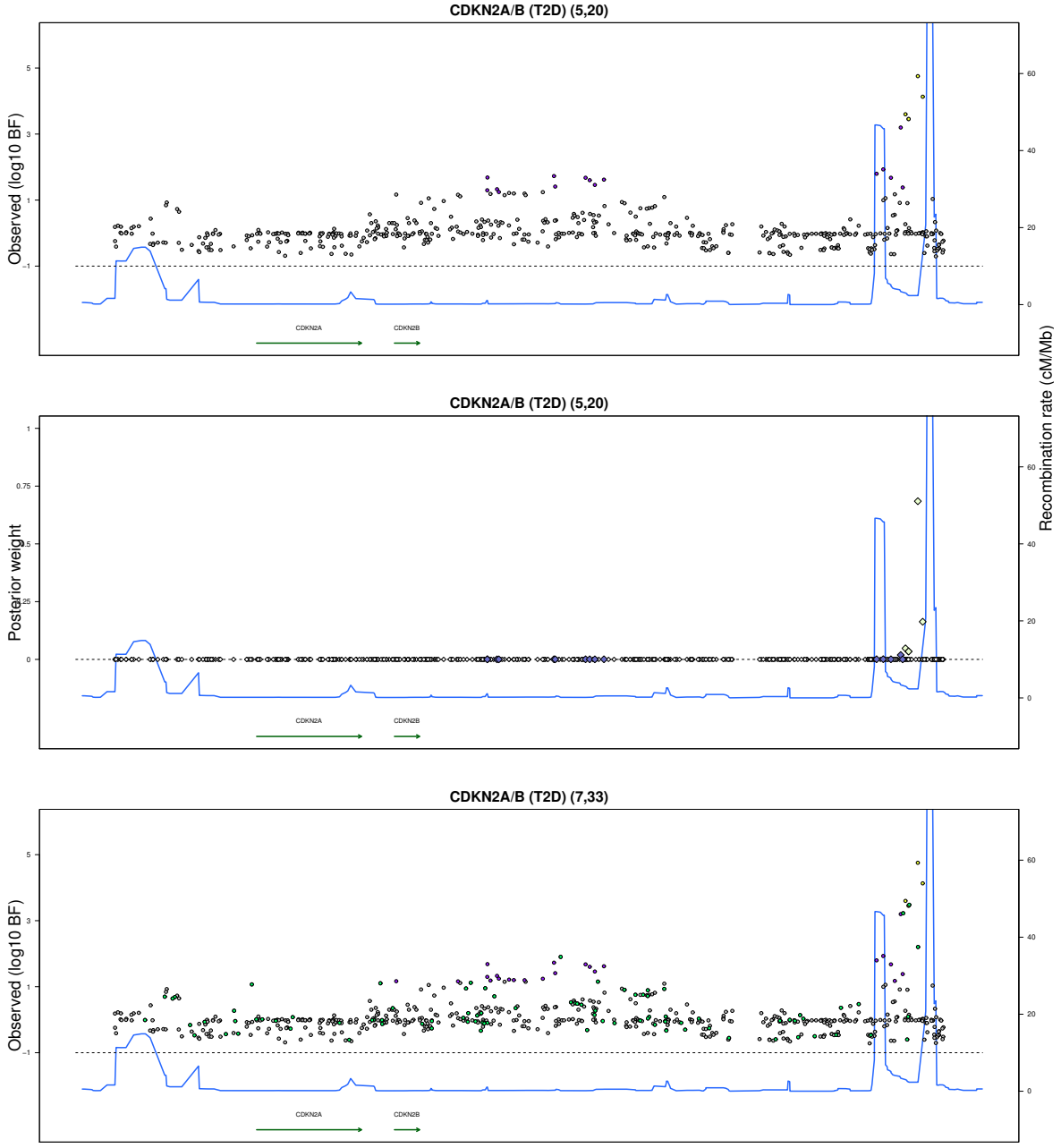


Figure 2.10: Regional results. All panels have genomic position on the x-axis. Recombination rate (cM/Mb) is shown in blue, with the scale on the right y-axis. SNP counts for 95% and 99% sets are in the title of each panel. a) Additive log_{10} Bayes Factors for each SNP in the *CDKN2A/B* region for T2D. SNPs in the 95% credible set are coloured yellow and SNPs exclusively in the 99% set are coloured purple. The left y-axis is log_{10} BF scale. b) The posterior probability associated with each SNP in the *CDKN2A/B* region for T2D. SNPs in the 95% credible set are coloured yellow and SNPs exclusively in the 99% set are coloured purple. c) Single marker additive log_{10} Bayes Factors across the *CDKN2A/B* region for T2D including imputed SNPs. Imputed SNPs are coloured green. The left y-axis is log_{10} BF scale.

CDKAL1

The association between variants mapping close to *CDKAL1* and T2D was identified first in the original WTCCC paper (rs9465871, risk allele C with frequency 0.178, $OR(95\%CI) = 1.18(1.04-1.34)$, $p = 1.02 \times 10^{-6}$) (WTCCC, 2007). This locus was subsequently confirmed and strengthened in several GWA studies (Zeggini et al., 2007) (Saxena et al., 2007) (Scott et al., 2007) (Steinthorsdottir et al., 2007) and the associated variants near *CDKAL1* that are implicated in psoriasis (lead SNP: rs6908425) (Quaranta et al., 2009) and Crohns disease (lead SNP: rs6908425) (Barrett et al., 2008) appear to be distinct from those associated with T2D, though they might reside on the same haplotype ($r^2 = 0.029$, $D' = 1.000$).

The mechanism of action of this association on T2D is unclear though *CDKAL1* shows homology with *CDK5RAP1*, a known inhibitor of *CDK5* activation; in turn *CDK5* has been implicated in the regulation of pancreatic beta cell function (Ubeda et al., 2006) (Wei et al., 2005).

The best SNP after the fine mapping is rs7756992 ($RR = 1.29$; 95% $CI = 1.19-1.40$), accounting for 35% of the posterior weight. The risk-variant previously highlighted at this locus, rs10946398 ($r^2 = 0.73$ with rs7756992) accounts for only 1.7% of the posterior in our analysis. There are 32 other SNPs in the 95% credible set.

All SNPs in this region are located in the third and fourth introns of *CDKAL1*. In the 95% credible set, there are 17 SNPs which are within a block with evidence of extended haplotype homozygosity in the HGDP samples from the Mideast, Europe, and South Asia. Four SNPs (rs9460544, rs9460545, rs4712526, rs35456723) in the 95% set show conservation and regulatory potential, with one

(rs35456723) among the most conserved regions in a 28-way alignment of mammalian genomes. The entire 99% credible set is within a broad domain containing histone modifications associated with active transcription.

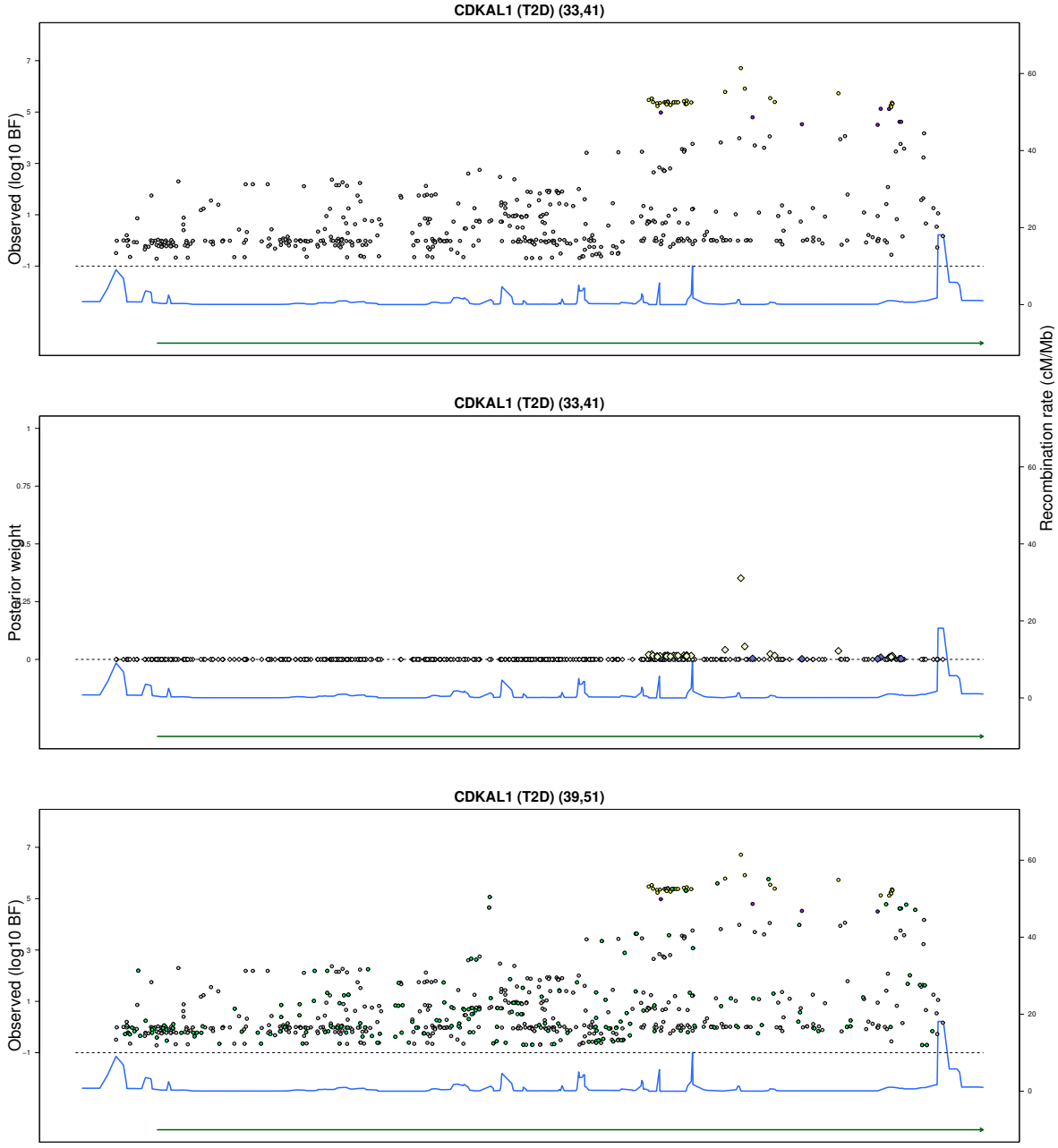


Figure 2.11: Regional results. All panels have genomic position on the x-axis. Recombination rate (cM/Mb) is shown in blue, with the scale on the right y-axis. SNP counts for 95% and 99% sets are in the title of each panel. a) Additive log_{10} Bayes Factors for each SNP in the *CDKAL1* region for T2D. SNPs in the 95% credible set are coloured yellow and SNPs exclusively in the 99% set are coloured purple. The left y-axis is log_{10} BF scale. b) The posterior probability associated with each SNP in the *CDKAL1* region for T2D. SNPs in the 95% credible set are coloured yellow and SNPs exclusively in the 99% set are coloured purple. c) Single marker additive log_{10} Bayes Factors across the *CDKAL1* region for T2D including imputed SNPs. Imputed SNPs are coloured green. The left y-axis is log_{10} BF scale.

FTO

The association between *FTO* variants and T2D was detected as one of the strongest signals in the WTCCC effort (rs9939609 with risk allele A with frequency 39.8%, $OR(95\%CI) = 1.34(1.17 - 1.52)$ and $p = 5.04 \times 10^{-8}$) (WTCCC, 2007). Interestingly this association was not observed in subsequent, independent efforts in leaner T2D patients (Saxena et al., 2007) (Scott et al., 2007) and it was discovered that this association is mediated through a strong, primary effect on BMI and risk of obesity (Frayling et al., 2008) (Scuteri et al., 2007) (Dina et al., 2007).

The T2D/BMI associated variants map to a 50kb interval in intron 1 of the *FTO* gene. Recent evidence from both mouse and human studies, implicates *FTO* as the gene through which the effect is mediated (Fischer et al., 2009) (Boissel et al., 2009) though it is not proven beyond doubt yet. *FTO*, localizes in the brain, shares sequence motifs with Fe(II)- and 2-oxoglutarate-dependent oxygenases and catalyzes the Fe(II)- and 2OG-dependent demethylation of 3-methylthymine in single-stranded DNA, with concomitant production of succinate, formaldehyde, and carbon dioxide. Taken together this suggests a potential role for *FTO* in nucleic acid demethylation in the brain (Gerken et al., 2007).

In the analyses here, we investigate SNPs in the region in terms of their effect on the endpoint of T2D, rather than directly on obesity (as BMI information was not available on many of our samples). We would expect use of T2D rather than BMI as an endpoint to attenuate the primary signal somewhat, but would not expect it to affect broad conclusions.

There is a set of 33 SNPs, for which all pairwise r^2 values are at least 0.95: together, these account for 95% of the posterior weight after fine mapping. The

RR associated with the top SNPs is 1.26 (95% $CI = 1.16 - 1.36$). A single common haplotype spans the entire region, with multiple SNPs segregating with the haplotype. In settings such as this, fine mapping focuses attention on the haplotype background, but will not be able to distinguish between the correlated SNPs unless very large numbers of samples are typed or transethnic analyses provide access to distinct haplotypic structures. Functional annotations for the associated SNPs offer few clues to which of these might be causal.

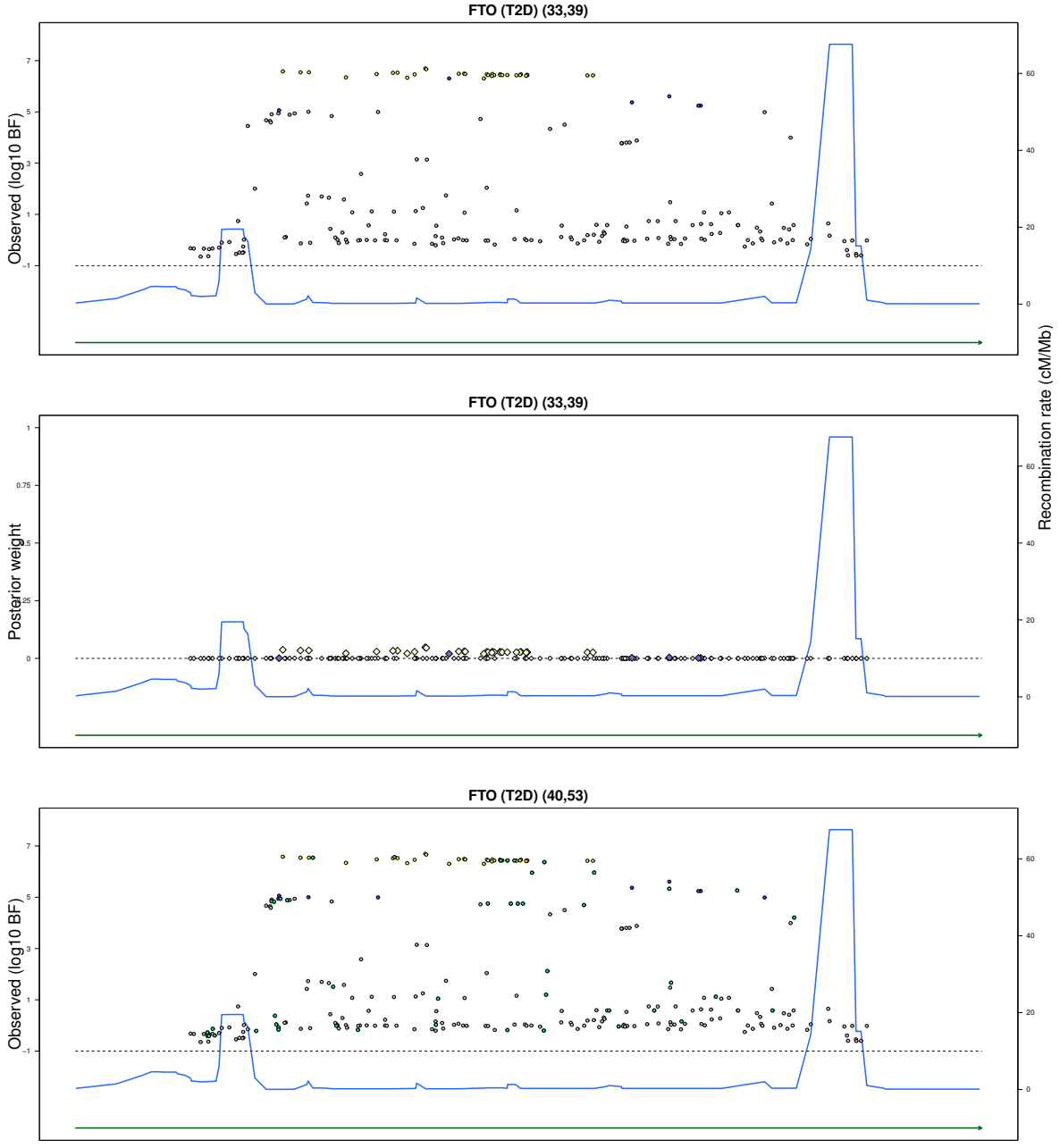


Figure 2.12: Regional results. All panels have genomic position on the x-axis. Recombination rate (cM/Mb) is shown in blue, with the scale on the right y-axis. SNP counts for 95% and 99% sets are in the title of each panel. a) Additive log_{10} Bayes Factors for each SNP in the *FTO* region for T2D. SNPs in the 95% credible set are coloured yellow and SNPs exclusively in the 99% set are coloured purple. The left y-axis is log_{10} BF scale. b) The posterior probability associated with each SNP in the *FTO* region for T2D. SNPs in the 95% credible set are coloured yellow and SNPs exclusively in the 99% set are coloured purple. c) Single marker additive log_{10} Bayes Factors across the *FTO* region for T2D including imputed SNPs. Imputed SNPs are coloured green. The left y-axis is log_{10} BF scale.

HHEX

The association between variants mapping within an interval containing the genes *HHEX*, *KIF11* and *IDE* was first reported in a GWAS of French subjects (rs1111875 with risk allele G and frequency 0.6, OR (with 95% CI)$= 1.19 \pm 0.19$, $p = 3.0 \times 10^{-6}$ (Sladek et al., 2007) and subsequently confirmed in other studies (Saxena et al., 2007) (Scott et al., 2007) (Zeggini et al., 2007).

The recombination interval contains two genes with strong biological candidacy for a role in T2D pathogenesis: *HHEX* which encodes a homeobox protein implicated in pancreatic development (Bort et al., 2006), and *IDE* (insulin degrading enzyme) has a role in insulin metabolism (Fakhrai-Rad et al., 2000) (Farris et al., 2003).

The best SNP after the fine mapping is rs10882098 ($RR = 1.21$; 95% $CI = 1.12 - 1.31$), accounting for 20% of the posterior weight. This SNP is well correlated ($r^2 = 0.73$) with the previously reported lead SNP for this region. There are 13 other SNPs in the 95% credible set, of which two (rs10882099, rs10882106) appear to lie in polymerase II binding sites: rs10882106 is also in a region of high conservation and suggestive DNase hypersensitivity.

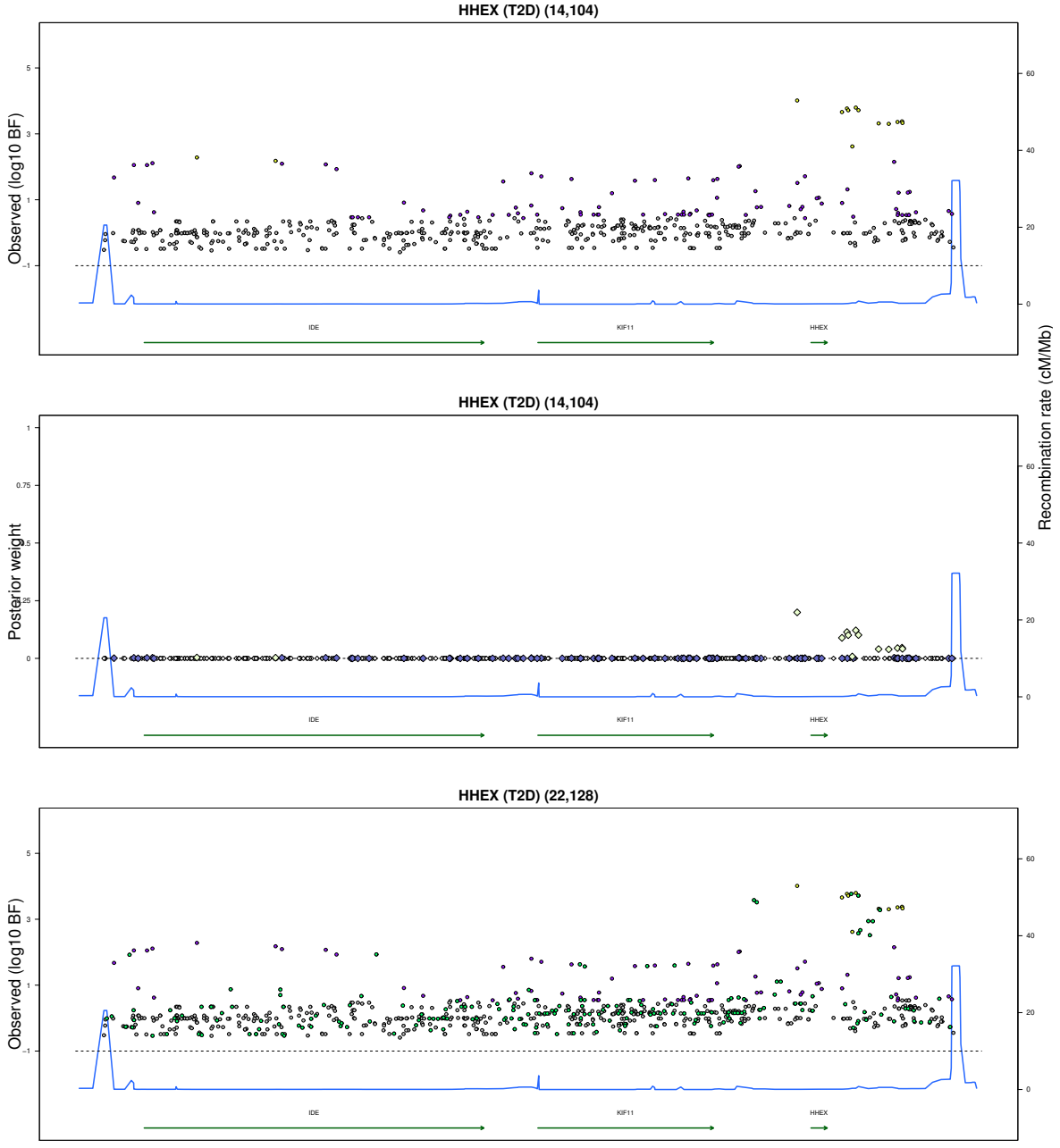


Figure 2.13: Regional results. All panels have genomic position on the x-axis. Recombination rate (cM/Mb) is shown in blue, with the scale on the right y-axis. SNP counts for 95% and 99% sets are in the title of each panel. a) Additive log_{10} Bayes Factors for each SNP in the *HHEX* region for T2D. SNPs in the 95% credible set are coloured yellow and SNPs exclusively in the 99% set are coloured purple. The left y-axis is log_{10} BF scale. b) The posterior probability associated with each SNP in the *HHEX* region for T2D. SNPs in the 95% credible set are coloured yellow and SNPs exclusively in the 99% set are coloured purple. c) Single marker additive log_{10} Bayes Factors across the *HHEX* region for T2D including imputed SNPs. Imputed SNPs are coloured green. The left y-axis is log_{10} BF scale.

TCF7L2

Transcription factor 7-like 2 (*TCF7L2*) was initially detected by linkage and fine-mapping efforts localized the likely causal variant(s) to chromosome 10 and an intron within *TCF7L2* (Helgason et al., 2007).

In the WTCCC effort, a cluster of SNPs within *TCF7L2*, generates the strongest association signal for T2D (rs4506565 T allele with frequency 0.32, $OR(95\%CI) = 1.36(1.2-1.54)$ and $p = 5.68 \times 10^{-13}$) and rs7903146 is in strong LD with rs4506565 ($r^2 = 0.892$ and $D' = 1.000$) and it remains the strongest common variant effect for T2D-predisposition in GWAS and meta-analysis efforts (Saxena et al., 2007) (Scott et al., 2007) (Sladek et al., 2007) (WTCCC, 2007) (Zeggini et al., 2007). The T2D association maps to a 64kb interval including exon 4 and flanking introns, with rs7903146 identified as the strongest causal candidate (Helgason et al., 2007). *TC*F7L2 acts within the WNT-signaling pathway, and effects on diabetes risk seem to be mediated predominantly through beta-cell dysfunction. More specifically, rs7903146 maps to an enhancer region active in pancreatic islets identified using FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements), and has allele-specific differences in both islet enhancer activity and chromatin accessibility (Gaulton et al., 2010).

In our analysis, rs7903146 remains the best SNP after fine mapping ($RR = 1.40$; 95% $CI = 1.29-1.52$), accounting for 75% of the posterior weight. There are four other SNPs in the 95% credible set (6 in the 99% set), all mapping within the largest intron of *TCF7L2*. rs7903146 maps to an enhancer region active in pancreatic islets identified using FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements), and has allele-specific differences in both islet enhancer activity

and chromatin accessibility (Gaulton et al., 2010).

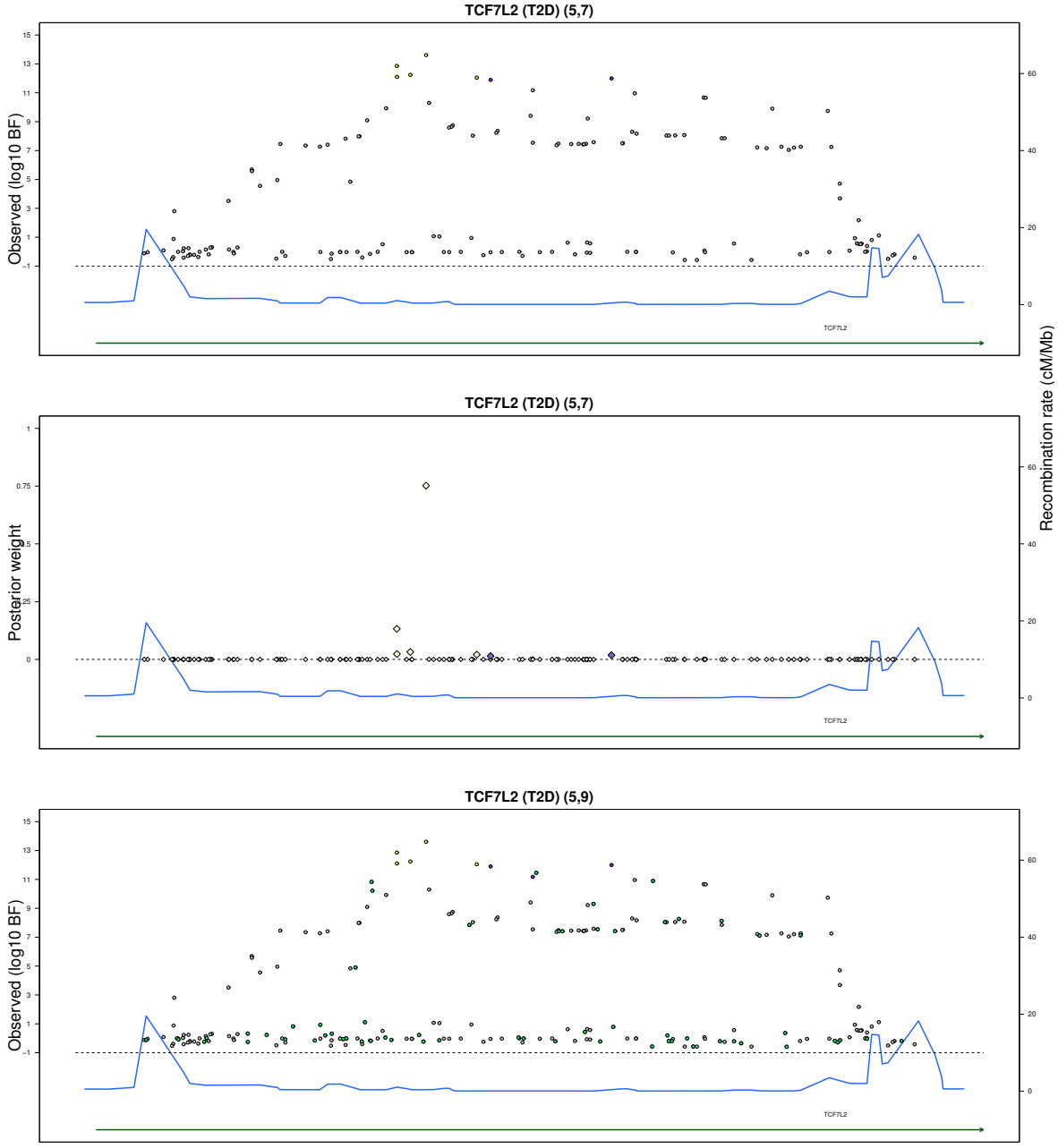


Figure 2.14: Regional results. All panels have genomic position on the x-axis. Recombination rate (cM/Mb) is shown in blue, with the scale on the right y-axis. SNP counts for 95% and 99% sets are in the title of each panel. a) Additive log_{10} Bayes Factors for each SNP in the *TCF7L2* region for T2D. SNPs in the 95% credible set are coloured yellow and SNPs exclusively in the 99% set are coloured purple. The left y-axis is log_{10} BF scale. b) The posterior probability associated with each SNP in the *TCF7L2* region for T2D. SNPs in the 95% credible set are coloured yellow and SNPs exclusively in the 99% set are coloured purple. c) Single marker additive log_{10} Bayes Factors across the *TCF7L2* region for T2D including imputed SNPs. Imputed SNPs are coloured green. The left y-axis is log_{10} BF scale.

JAZF1

Variants around the *JAZF1* gene were not detected in the initial WTCCC effort but were implicated later in T2D-susceptibility through a meta-analysis of three European GWA studies with subsequent replication (combined results from all individuals: rs864745 at position 27,953,796 bp, risk allele T with frequency 0.501, $OR(95\%CI) = 1.10(1.07 - 1.13)$, $p = 5.0 \times 10^{-14}$) (Zeggini et al., 2008). This association has since been confirmed in studies from non-European populations (Omori et al., 2009).

JAZF1 is a transcriptional repressor of *NR2C2* and variants at this locus have also been described with associations to both height (Johansson et al., 2009, Soranzo et al., 2009) and prostate cancer (Frayling et al., 2008, Thomas et al., 2008). There is evidence that supports the effect of *JAZF1* mediation via reduced insulin secretion (Grarup et al., 2008).

Our fine mapping data failed to refine the signal in this region. The top SNP (rs12531540) accounted for 9% of the posterior weight, and 252 SNPs were included in the 95% credible set.

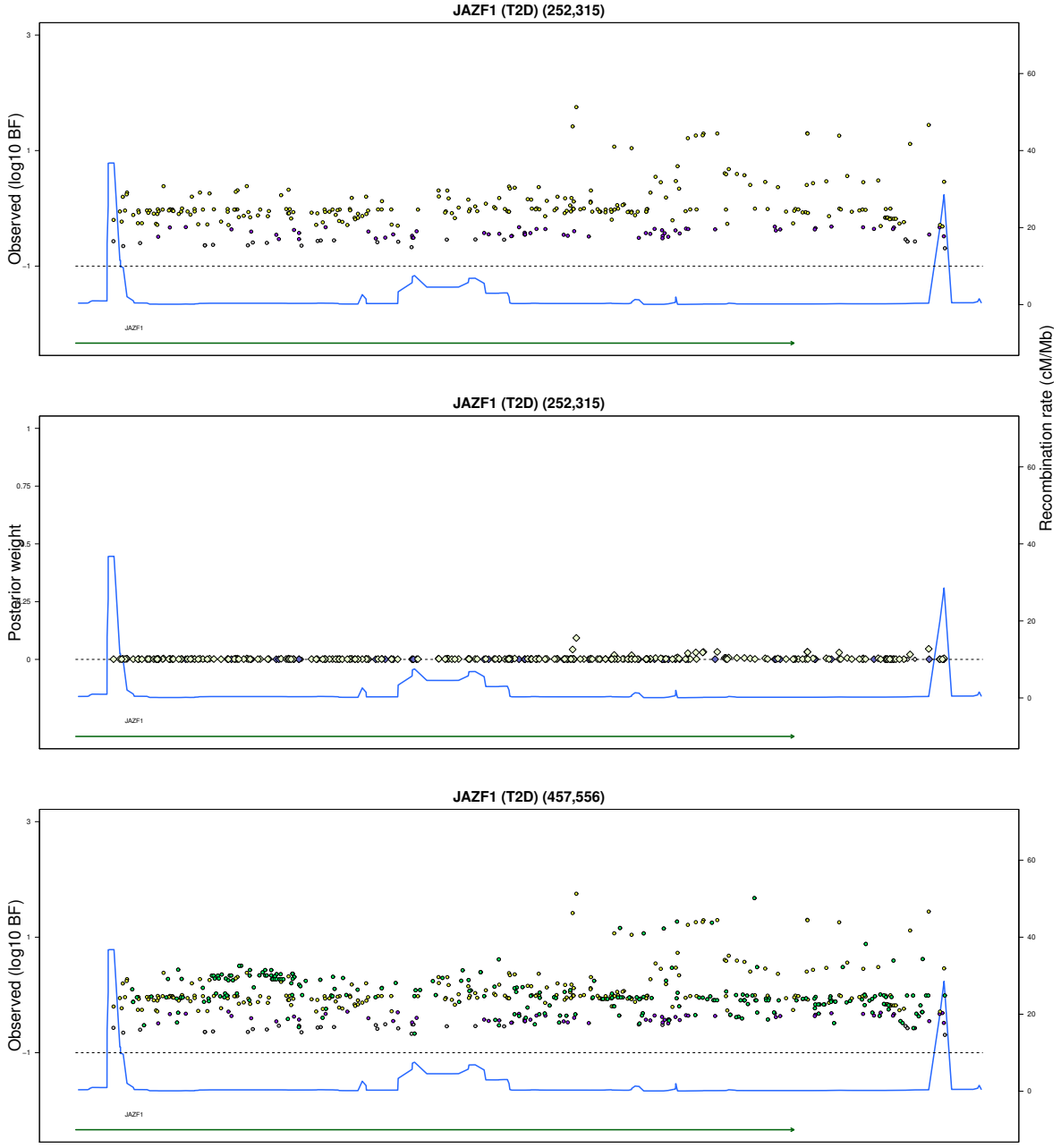


Figure 2.15: Regional results. All panels have genomic position on the x-axis. Recombination rate (cM/Mb) is shown in blue, with the scale on the right y-axis. SNP counts for 95% and 99% sets are in the title of each panel. a) Additive log_{10} Bayes Factors for each SNP in the *JAZF1* region for T2D. SNPs in the 95% credible set are coloured yellow and SNPs exclusively in the 99% set are coloured purple. The left y-axis is log_{10} BF scale. b) The posterior probability associated with each SNP in the *JAZF1* region for T2D. SNPs in the 95% credible set are coloured yellow and SNPs exclusively in the 99% set are coloured purple. c) Single marker additive log_{10} Bayes Factors across the *JAZF1* region for T2D including imputed SNPs. Imputed SNPs are coloured green. The left y-axis is log_{10} BF scale.

2.10 Discussion

The regional signal plots reveal several features. Only two of the regions have most of the cumulative posterior probability resting on a single SNP: rs7903146 for *TCF7L2* in T2D (posterior probability = 0.75) and rs11571297 for *CTLA4* in GD (posterior probability = 0.77). Seven of the 14 regions (*CDKN2A/B* CAD, *CDKN2A/B* T2D, *CDKAL1*, *CTLA-4*, *FTO*, *HHEX*, *TCF7L2*) have relatively small credible sets. This suggests that the fine mapping experiment has provided useful information, at least in excluding large numbers of SNPs from being causal. In the other remaining 7 regions, (*1q41, 2q36, CD25, CXCL12, FCRL3, JAZF1, SORT1*) the fine mapping experiment seems to have added little. Even with extensive new genotyping data, dozens or even hundreds of the SNPs in the region cannot be excluded as being causal.

2.10.1 Comparison with recently published data

Ranking of SNPs by posterior probability is in general different from ranking based on p-values, complicating direct comparisons of our findings with other studies. In comparing our top SNPs with those in a recent large follow-up study of T2D (the DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium et al., 2012), they were identical for three regions (*CDKN2A/B*, *CDKAL1*, *TCF7L2*), and almost perfect proxies in the other two regions for which our experiment was informative (*FTO*, $r^2 = 0.995$; HHEX, $r^2 = 0.967$) (Table 2.3, which also compares our findings with a recent review of the GWAS results for CAD (Peden and Farrall, 2011), for which comparison is complicated because our top SNPs may not have been typed in the studies reviewed.)

Phenotype	Region	Reported lead SNP	Reported effect size	Effect size in our study	Lead SNP(s) in our study	Effect size	r^2
T2D	*FTO*	rs9936385	1.13	1.26	rs17817449	1.26	1.0
	CDKN2A/B	rs10811661			rs10811661		
	HHEX	rs1111875	1.11	1.2	rs10882098	1.21	0.97
	CDKAL1	rs7756992			rs7756992		
	TCF7L2	rs7903146			rs7903146		
	JAZF1	rs849135	1.11	1.13	rs12531540	1.14	0.88
CAD	*SORT1*	rs599839	1.14	1.17	rs3832016	1.22	0.92
	CDKN2A/B	rs4977574	1.29	1.37	rs1537370	1.40	0.85
	CXCL12	rs1746048	1.09	1.12	rs34161818	1.21	0.18
	1q41	rs17465637	1.14	1.12	rs2936023	1.21	0.49
	2q36	not present			rs2673145	1.2	

Table 2.3: Comparison between results of our study and recent publications. T2D results compared with meta-analysis results for 34,840 cases and 114,981 controls from Morris, A.P. et al. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nat Genet (2012). CAD results compared with those listed from the review Peden, J.F. & Farrall, M. Thirty-five common variants for coronary artery disease: the fruits of much collaborative labour. Hum Mol Genet 20, R198-205 (2011). Region headings follow those used in this thesis. Subsequent columns give the rsid of the lead SNP in the relevant recent study followed by the effect size estimate from that study. Next, the table gives the effect size estimate for that SNP in our study, the rsid of our top SNP for the region, and its effect size estimate from our study. The final column gives the r^2 value, calculated in our control data, between the two SNPs. Where the pair of SNPs in a row are identical, other columns are left blank.

2.10.2 Excluding SNPs

We undertook to investigate whether the fine-mapping data and results allow us to effectively *exclude* SNPs as causal. In seven of the 14 regions (*CDKN2A/B* in CAD; *CTLA4* in GD; *CDKN2A/B*, *CDKAL1*, *FTO*, *HHEX*, and *TCF7L2* in T2D), including the three just mentioned, the credible sets are relatively small, fewer than 34 SNPs, and the fine mapping experiment has provided useful information, at least in excluding large numbers of SNPs from being causal (see Table 2.4). In the other seven regions, 1q41, 2q36, *IL2RA*, *CXCL12*, *FCRL3*, *JAZF1*, *SORT1*, the project has failed to exclude many of the SNPs as potentially being causal (or for being good markers) for each region the 95% credible set contains at least 75 and typically over 100 SNPs. Nonetheless, two of these seven regions (*SORT1* and *IL2RA*) have the property that even though large numbers of SNPs cannot be excluded, substantial posterior weight rests on one or a few SNPs, and so typing additional samples in these regions should lead to a much smaller credible set.

Region	Total SNPs	SNPs Excluded	Proportion Excluded
TCF7L2	157	152	0.97
CDKN2A/B (T2D)	519	514	0.99
FTO	207	174	0.84
CDKAL1	497	464	0.93
HHEX	560	546	0.98
CDKN2A/B (CAD)	515	502	0.97
SORT1	231	139	0.60
CTLA4	293	287	0.98
CD25	426	350	0.82
JAZF1	339	87	0.26
1q41	354	114	0.32
2q36	343	257	0.75
CXCL12	363	97	0.27
FCRL3	607	493	0.81

Table 2.4: For each region, the number and proportion of SNPs excluded from the set of potentially causal variants. Any SNP not included in the 95% credible set was considered to be excluded. Total SNPs is the number of genotyped SNPs in each region which pass QC and are polymorphic. SNPs excluded is the number of SNPs in each region not included in the credible set accounting for 95% of the posterior probability. Proportion excluded is the proportion of SNPs in each region not included in the credible set accounting for 95% of the posterior probability.

Chapter 3

Fine Mapping Secondary Analyses

3.1 Functional Annotation

For most of the regions where the fine mapping considerably refined the signal, there were still a set of SNPs which could not be separated based only on the genotype data. One way to further narrow the set of potential causal variants is to look at annotation data to see if any of the top SNPs have functional roles. For example, are there any non-synonymous coding variants, do any correspond to known splice sites, or are any known to be conserved across species? Together with Jake Byrnes, I cross-referenced our top SNPs against all relevant annotation tracks in the UCSC genome browser. The results, for the seven regions where the fine mapping refined the picture, are shown in Table 3.1. The lack of putative functional annotations is striking. Of the 109 SNPs in the 95% credible sets (247 in the 99% credible sets) across the seven regions, there were no SNPs in exons, splice-site variants or enhancers, only 1 SNP (3 in the 99% set) in a putative transcription factor binding site, 5 SNPs (7 in 99% set) in mammalian highly conserved regions, and none are present in the eQTL lists of two recent genome-wide studies (Stranger et al., 2007), (Dixon et al., 2007). Looking further, we find that only 4 SNPs are in predicted high nucleosome occupancy sites, 1 is in a low occupancy site, 4 SNPs are located in DNase hypersensitive sites, and none are in predicted miRNAs or their target sites. One notable exception may be at *TCF7L2*. The top SNP for this locus, rs7903146, which accounts for 75% of the posterior evidence of association, is annotated to be within a predicted *FOX2* binding site (Yeo et al., 2009a). *FOX2* is known to be expressed in muscle and neuronal tissue and plays a role in tissue specific alternative splicing.

Table 3.1 shows the complete biological annotations for the 109 and 247 SNPs

making up the 95% and 99% credible sets across seven fine mapped regions.

Annotation	Proportion of SNPs in 95% (99%) credible set	Proportion of posterior probability in 95% (99%) credible set
refGene exon/intron		
all exons	0 (0.03)	0 (0)
coding exons	0 (0)	0 (0)
largest intron	0.05 (0.09)	0.14 (0.14)
first intron	0.3 (0.19)	0.14 (0.14)
Ensembl introns		
largest intron	0.05 (0.11)	0.14 (0.14)
first intron	0.3 (0.18)	0.14 (0.14)
dbSNP 130 functions		
nonsynonymous	0 (0)	0 (0)
synonymous	0 (0)	0 (0)
intron	0.67 (0.53)	0.43 (0.43)
splicing	0 (0)	0 (0)
5' utr	0 (0)	0 (0)
3' utr	0 (0.02)	0 (0)
near-gene-5' ($\leq$ 2kb)	0.01 (0.01)	0 (0)
near-gene-3' ($\leq$ 2kb)	0.01 (0)	0.01 (0.01)
unknown (intergenic)	0.31 (0.44)	0.56 (0.56)
Other gene prediction methods		
alternative splicing events	0 (0)	0 (0)
noncoding RNA	0 (0)	0 (0)
AceScan	0 (0)	0 (0)
EVOfold secondary structure	0 (0)	0 (0)
miRNA (mirBase)	0 (0)	0 (0)
HGDP Fst and signals of selective sweep		
positive selection on human branch $p < 0.05$	0 (0)	0 (0)
HGDP Fst $p < 0.05$	0.06 (0.07)	0.03 (0.03)
HGDP Bantu iHS $p < 0.05$	0.12 (0.10)	0.14 (0.14)
HGDP Mideast iHS $p < 0.05$	0 (0)	0 (0)
HGDP Europe iHS $p < 0.05$	0 (0)	0 (0)
HGDP S. Asia iHS $p < 0.05$	0 (0)	0 (0)
HGDP E. Asia iHS $p < 0.05$	0.13 (0.42)	0.14 (0.14)
HGDP Oceania iHS $p < 0.05$	0 (0)	0 (0)
HGDP Americas iHS $p < 0.05$	0.05 (0.02)	0.01 (0.01)
HGDP Bantu XP-EHH $p < 0.05$	0.09 (0.08)	0.04 (0.04)
HGDP Mideast XP-EHH $p < 0.05$	0.22 (0.10)	0.06 (0.06)
HGDP Europe XP-EHH $p < 0.05$	0.15 (0.07)	0.04 (0.04)
HGDP S. Asia XP-EHH $p < 0.05$	0.22 (0.10)	0.06 (0.06)
HGDP E. Asia XP-EHH $p < 0.05$	0.12 (0.36)	0.24 (0.24)
HGDP Oceania XP-EHH $p < 0.05$	0.02 (0.01)	0.12 (0.12)
HGDP Americas XP-EHH $p < 0.05$	0.02 (0.01)	0.12 (0.12)
Broad/Uppsala/GIS ChIP-PET histone methylation and acetylation		

Continued on Next Page...

Annotation	Proportion of SNPs in 95% (99%) credible set	Proportion of posterior probability in 95% (99%) credible set
Broad H3K4me1 any $-\log p > 5$	0.71 (0.57)	0.67 (0.67)
Broad H3K4me2 any $-\log p > 5$	0.47 (0.38)	0.54 (0.54)
Broad H3K4me3 any $-\log p > 5$	0.17 (0.14)	0.15 (0.15)
Broad H3K9ac $-\log p > 5$	0.33 (0.24)	0.21 (0.21)
Broad H3K9me1 $-\log p > 5$	0.40 (0.36)	0.31 (0.31)
Broad H3K27ac $-\log p > 5$	0.48 (0.34)	0.37 (0.37)
Broad H3K27me3 $-\log p > 5$	0.38 (0.43)	0.71 (0.71)
Broad H3K36me3 $-\log p > 5$	0.58 (0.71)	0.49 (0.50)
Broad H4K20me1 $-\log p > 5$	0.72 (0.70)	0.48 (0.49)
Uppsala H3ac	0 (0.02)	0 (0)
GIS ChIP-PET H3K4me3	0 (0.01)	0 (0)
GIS ChIP-PET H3K27me3	0.03 (0.03)	0.02 (0.02)
Transcription factor binding sites (ChIP-Seq)		
Uppsala USF1	0 (0)	0 (0)
Uppsala USF2	0 (0)	0 (0)
HAIB NRSF signal > 5	0 (0)	0 (0)
HAIB Pol2 signal > 5	0 (0)	0 (0)
Broad/Duke Pol2 $p < 0.05$	0.03 (0.08)	0.06 (0.06)
Yale Pol2 $q < 0.05$	0 (0.03)	0 (0)
Broad/Duke CTCF $p < 0.05$	0.04 (0.05)	0.04 (0.04)
Duke c-Myc $p < 0.05$	0 (0.01)	0 (0)
GIS ChIP-PET c-Myc score ≥ 800	0.03 (0.02)	0.03 (0.03)
Yale JunD $q < 0.05$	0 (0)	0 (0)
Yale Max $q < 0.05$	0 (0)	0 (0)
Yale SREBP1 $q < 0.05$	0 (0)	0 (0)
Yale SREBP2 $q < 0.05$	0 (0)	0 (0)
Yale c-Fos $q < 0.05$	0.04 (0.04)	0.02 (0.02)
Yale c-Jun $q < 0.05$	0.01 (0)	0.01 (0)
Yale NF-E2 $q < 0.05$	0 (0)	0 (0)
Yale Rad21 $q < 0.05$	0 (0.01)	0 (0)
Yale ZNF263 $q < 0.05$	0 (0)	0 (0)
Yale GATA1 $q < 0.05$	0 (0)	0 (0)
Yale TCF7L2 $q < 0.05$	0.05 (0.05)	0.03 (0.03)
Yale SATA1 $q < 0.05$	0.06 (0.05)	0.05 (0.05)
GIS ChIP-PET p53 score ≥ 800	0 (0)	0 (0)
DNase I hypersensitivity and nucleosome occupancy/accessibility		
Duke/UW DNaseHS $p < 0.05$	0.10 (0.11)	0.07 (0.08)
EIO/JCV Nucleosome Accesibility CD34+	0.03 (0.02)	0.04 (0.04)
EIO/JCV Nucleosome Accesibility CD34-	0.02 (0.02)	0.05 (0.05)
UW Nucleosome Occupancy A375 > 1.0	0.01 (0.03)	0 (0)
UW Nucleosome Occupancy Dennis > 1.0	0.03 (0.04)	0.02 (0.02)
UW Nucleosome Occupancy Mec < -1.0	0.01 (0.01)	0 (0)
Other regulatory predictions		
NHGRI NRE score ≥ 800	0 (0)	0 (0)

Continued on Next Page...

Annotation	Proportion of SNPs in 95% (99%) credible set	Proportion of posterior probability in 95% (99%) credible set
ORegAnno	0.01 (0)	0.01 (0.01)
switchGear TSS	0 (0)	0 (0)
TFBScons	0.01 (0.01)	0 (0)
miRNA target site	0 (0)	0 (0)
Vista Enhancer	0 (0)	0 (0)
7Xreg score > 0.1	0.14 (0.1)	0.07 (0.07)
FOX2 binding sites	0.14 (0.07)	0.15 (0.15)
LI TAF1 sites	0 (0)	0 (0)
NKI LADs	0.06 (0.06)	0.14 (0.14)
eponine TSS	0 (0)	0 (0)
firstEF	0 (0.01)	0 (0)
NHGRI BiPro score ≥ 800	0.30 (0.16)	0.14 (0.14)
Sequence conservation		
17-way most conserved Vertebrate	0.05 (0.04)	0.02 (0.02)
28-way most conserved Mammals	0.05 (0.03)	0.02 (0.02)

Table 3.1: Complete biological annotations for the 109 and 247 SNPs making up the 95% and 99% credible sets across seven fine mapped regions.

3.1.1 Methods

We used the Galaxy web platform to parse relevant SNP annotations found in the UCSC genome browser for the human genome release hg18 (Giardine et al., 2005). All 247 SNPs contained in the 99% credible sets for the seven regions in which fine mapping was informative, along with SNPs contributing significant secondary signals to these regions, were cross-referenced for annotations in the following tracks (details to follow below):

- Genes and Gene Prediction Tracks
  - RefSeq Genes
    - refGene (extracted exons and introns)
  - Ensembl Genes

        * ensGene (extracted exons and introns)
    - Alt Events
    - RNA Genes
    - ACEScan
    - EvoFold
    - sno/miRNA
    - Pos Sel Genes
- Variation and Repeats Tracks
    - SNPs (130)
    - HGDP Smoothd Fst
    - HGDP iHS
    - HGDP XP-EHH
- Regulation Tracks
    - Broad Histone (peaks tables only)
    - EIO/JCVI NAS
    - Eponine TSS
    - First EF
    - GIS ChIP-PET (peaks tables only)
    - HAIB Methyl-seq
    - HAIB Methyl27
    - HAIB TFBS (peaks tables only)
    - NHGRI Bi-Pro
    - NHGRI NRE
    - Open Chromatin
    - ORegAnno
        * oreganno
    - SUNY RBP
    - SwitchGear TSS
    - TFBS Conserved
        * tfbsConsSites

  - TS miRNA sites
  - UW DNaseI HS
  - Vista Enhancers
  - Yale TFBS
  - 7X Reg Potential
  - FOX2 CLIP-seq
    - fox2ClipSeq
  - LI TAF1 Sites
  - NKI LADs (Tig3)
  - Nucl Occ: A375
  - Nucl Occ: Dennis
  - Nucl Occ: MEC
  - UU ChIP Sites
- Comparative Genomics Tracks
  - phastConst28wayPlacMammal
  - phastConst17way

Unless otherwise specified, all available tables for each annotation track were used.

Among the **Genes and Gene Prediction** track group, **RefSeq Genes** (Kent, 2002, Pruitt et al., 2006) and **Ensembl Genes** (Hubbard et al., 2002) tracks were used to annotate SNPs to coding and non-coding (UTR) exons (appearing as **refGene all exons** and **refGene coding exons** in annotation tables). Since both the first and the largest intron in each transcript often contains regulatory information, we wanted to quantify which of our intronic SNPs were in either the first and/or largest introns. A SNP was annotated as **refGene largest intron** or **refGene first intron** if it was contained in largest or first intron of any annotated transcript from the refGene gene set respectively. The same criteria was

used for the **ensGene largest intron** and **ensGene first intron** annotations using the Ensembl gene set. The **Alt Events** track (appearing as **alternative splicing events** in annotation tables), based on analysis from the `txgAnalyse` program by Jim Kent, contains annotations of various alternative splicing events. The **RNA Genes** track (appearing as **noncoding RNA** in annotation tables) contains all non-protein coding RNA genes including tRNAs, rRNAs, miRNAs, etc. (Pasquinelli et al., 2000, Mourelatos et al., 2002). The **ACEScan** track contains predicted alternative conserved exons (Yeo et al., 2005). The **EvoFold** track (appearing as **EvoFold secondary structure** in the Table 3.1) contains the predictions of the EvoFold program, a method which uses evolutionary conservation to help locate regions with RNA secondary structure (Pedersen et al., 2006). The **sno/miRNA** track (appearing as **miRNA** in annotation tables) includes annotations for snoRNAs from snoRNABase and miRNAs from miRBase (formerly the miRNA Registry) (Griffiths-Jones, 2004, Weber, 2005). The **Pos Sel Genes** track (appearing as **positive selection on human branch p $<$ 0.05** in Table ()) shows the results of a genome-wide scan for positively selected genes from a six species sequence alignment. We focused on genes showing significant evidence of positive selection on the human branch of the phylogenetic tree (Kosiol et al., 2008, Yang and Nielsen, 2002).

Among the **Variation and Repeat** track group, we used **SNPs (130)** to further annotate SNPs to genic functional positions (nonsynonymous, synonymous, intron, splicing, 5' utr, 3' utr, near-gene-5' and near-gene-3', unknown) (Sherry et al., 2001b). Human Genome Diversity Panel (HGDP) tracks were used to look for population differentiation and population-specific selection (Pickrell et al., 2009, Li et al., 2008b, Cann et al., 2002, Sabeti et al., 2007, Voight et al., 2006).

Among the **Regulation** track group, many tracks were used to search for higher-order functional effects of SNPs. The **Broad Histone** track (appearing, for example, as **Broad H3K4me1 any -logp** > 5 in Table 3.1) contains regions with significant enrichment for histones with specific modification tags that affect chromatin conformation and thus local gene expresion (Bernstein et al., 2005, Bernstein et al., 2006, Mikkelsen et al., 2007). The data are provided as part of the ENCODE project and each annotation is available for numerous cell types. For this and all other ENCODE annotation data, we used only data available for unrestricted usage as of February 2010. For our analysis, annotations were combined across cell types. A SNP was scored with an annotation if, for any cell type, it was within a peak with significant ($-\log_{10} p > 5$) ChIP-Seq enrichment. The **EIO/JCVI NAS** track (appearing as **EIO/JCV Nucleosome Accessibility CD34** in Table 3.1) provides maps of nuclease hypersensitive regions, known to correlate with functionally active cis-regulatory elements (Boyle et al., 2008a). The **Eponine TSS** track contains annotations of predicted transcription start sites (Down and Hubbard, 2002). The **firstEF** annotation track contains the predictions of the first Exon Finder program for the locations of first exons, promoters, and CpG windows (Davuluri et al., 2001). The **GIS ChIP-PET** track provides transcription factor binding site maps for *p53*, *STAT1*, *c-Myc*, along with histone modifications H3K4me3 and H3K27me3 (Ng et al., 2005, Chiu et al., 2006, Wei et al., 2006, Choo et al., 2006, Zeller et al., 2006). The three **HAIB** tracks provide Hudson Alpha's ENCODE data of transcription factor binding sites mapped using ChIP-Seq (Johnson et al., 2007, Valouev et al., 2008). **NHGRI BiPro** (Yang et al., 2008, Piontkivska et al., 2009) and **NHGRI NRE** (Petrykowska et al., 2008) tracks contain the National Human

Genome Research Institute's maps of bidirectional promoters and negative regulatory elements respectively. The **Open Chromatin** track contains the data from the collaborative Duke/UNC/UT-Austin/EBI ENCODE group mapping accessible chromatin by DNaseI hypersensitivity (appearing as **Duke/UW DNaseHS p $<$ 0.05**), Formaldehyde-Assisted Isolation of Regulatory Elements, and ChIP-chip of transcription factors (appearing combined with Broad ENCODE annotations, for example, **Broad/Duke CTCF** $p < 0.05$) (Bhinge et al., 2007, Boyle et al., 2008a, Boyle et al., 2008b, Buck et al., 2005, Crawford et al., 2006b) (Crawford et al., 2006a, The ENCODE Project Consortium, 2007) (Giresi et al., 2007, Giresi and Lieb, 2009,Li et al., 2008a). The **ORegAnno** track provides literature-curated regulatory regions (Montgomery et al., 2006). **SwitchGear TSS**, created by Nathan Trinklein and Shelley Force Aldred, maps transcription start sites. The **TFBS Conserved** (appearing as **TFBScons**) track contains computationally predicted conserved transcription factor binding sites from a human, mouse, and rat genome alignment. It was created by Matt Weirauch and Brian Raney of UCSC using the Transfac Matrix and Factor databases (Wingender, 2008). **TS miRNA sites** (appearing as **miRNA target site**) contains conserved miRNA target sites in refSeq genes predicted by TargetScanS (Lewis and Shih, 2003, Lewis et al., 2005, Grimson et al., 2007). **UW DNaseI HS** (appearing as **Duke/UW DNaseHS p $<$ 0.05**) provides the University of Washington ENCODE data on DNase hypersensitivity sites (Sabo et al., 2006, Sabo et al., 2004). The track **Vista Enhancers** contains conserved non-coding sequences that have shown reproducible enhancer activity in a transgenic mouse reporter assay (Pennacchio et al., 2006). The **Yale TFBS** track contains the data from the collaborative Yale/UC-Davis/Harvard ENCODE group mapping binding sites for multiple transcription factors across mul-

tiple cell types (Euskirchen et al., 2004, Euskirchen et al., 2007, Martone et al., 2003, Robertson et al., 2007, Rozowsky et al., 2009). Again, a SNP was scored with an annotation if, for any cell type, it was within a peak with significant ($q < 0.05$) ChIP-Seq enrichment. The **7X Reg Potential** track shows regulatory potential (RP) scores derived from an alignment of human, chimp, macaque, mouse, rat, dog, and cow genomes (King et al., 2005, Kolbe et al., 2004). The score compares the frequency of short regulatory motifs in a region to that of neutral DNA. We include only regions with a score > 0.1, which is the suggested cutoff for regions with high regulatory potential. The **FOX2 CLIP-seq** (appearing as **FOX2 binding sites** in Table 3.1) track contains FOX2 CLIP-seq (cross-linking immunoprecipitation combined with high-throughput sequencing) reads from human embryonic stem cells uniquely mapped to the reference genome (Yeo et al., 2009b). **LI TAF1 Sites** shows TAF1 binding sites (Kim et al., 2005). **NKI LADs** shows lamina associated domains. The three **Nucl Occ** (A375, Dennis, and MEC) tracks (appearing as **UW Nucleosome Occupancy**) contain nucleosome occupancy scores determined by MNase digestion assays (Dennis et al., 2007, Ozsolak et al., 2007, Gupta et al., 2008). A375 and Dennis tracks are tuned to locate regions of high occupancy and MEC is tuned to locate regions of low occupancy. The threshold of 1.0 (-1.0 for MEC) corresponds to a confident prediction. Finally, the **phastCons17way** and **phastConst28wayPlacMammal** tracks (appearing as **17-way most conserved Vertebrate** and **28-way most conserved Mammals**) contain predicted conserved elements from a 17-way vertebrate genome alignment and a 28-way placental mammal genome alignment respectively (Siepel et al., 2005). **UU ChIP Sites** (appearing as **Uppsala USF1**, **Uppsala USF2**, and **Uppsala H3ac**) contains a map of *USF1*, *USF2*, and H3ac sites (Rada-Iglesias

et al., 2005, Rada-Iglesias et al., 2008).

3.1.2 Annotations of interest by region

T2D:*TCF7L2*

All 5 SNPs in the 95% posterior set (all 7 in 99% set) are within the largest intron of *TCF7L2*. The best SNP after the fine mapping, rs7903146 (accounting for 75% of the posterior weight) maps into a putative *FOX2* binding site identified by cross-linking immunoprecipitation coupled with high-throughput sequencing (CLIP-seq) (Yeo et al., 2009b). *FOX2* is known to be expressed in muscle and neuronal tissue and plays a role in tissue specific alternative splicing. One of the four other SNPs in the 95% credible set, rs4132670 (accounting for 1% of the posterior), is in a conserved domain with regulatory potential and apparent nucleosome occupancy. This domain also appears to be DNase/nuclease hypersensitive in the CD34- neuroblastoma cell line SK-N-MC, and in CD34- maturating myeloid cells. There is also evidence for extended haplotype homozygosity in East Asia samples from the Human Genome Diversity Panel (HGDP) in this intron which is suggestive of a selective sweep.

T2D:*CDKN2A*

Looking in to SNP annotations for this region, we find that two SNPs in the 99% credible set (rs7866783 and rs564398 with posterior probabilities of 5% and 2% respectively) are in exons of the non-coding antisense RNA (*CDKN2BAS*, also known as *ANRIL*).

T2D:*CDKAL1*

All SNPs in this region are located in the third and fourth introns of *CDKAL1*. 17 SNPs in the 95% credible set are within a block with evidence of extended haplotype homozygosity in the HGDP samples from the Mideast, Europe, and South Asia. Four SNPs (rs9460544, rs9460545, rs4712526, rs35456723 accounting for 2%, 2%, 2%, 1% of the posterior respectively) in the 95% set show conservation and regulatory potential, with one (rs35456723) among the most conserved regions in a 28-way alignment of mammalian genomes.

We find that 1 SNP in the 99% credible set (rs35612982 with a posterior probability of 0.4%) shows high Fst in the HGDP.

T2D:*HHEX*

Two SNPs among the 95% set (rs10882099 and rs10882106 accounting for 9% and 4% of the posterior respectively) appear to be in polymerase II binding sites. rs10882106 is also in a region of high conservation and suggestive DNase hypersensitivity. Looking for signals of selection and population differentiation, we find that 1 SNP (rs7923866 accounting for 4% of the posterior) in the 95% credible set (11 in 99% set) shows high Fst in HGDP. Also, the entire associated region, including *HHEX*, *IDE*, and *KIF11* shows extended haplotype homozygosity among East Asian samples, suggesting the effect of a selective sweep.

5 SNPs in the 99% credible set are also in untranslated gene regions (1 in the 5 UTR of *IDE* and 4 in the 3 UTR of *KIF11*).

CAD:*CDKN2A*

All SNPs in the 99% credible set for this region are within the 3‘ half of *ANRIL* and within a region showing weak evidence of selective sweep in the Bantu samples of HGDP. Our top SNP (rs1537370 accounting for 26% of the posterior) appears to be in a nuclease accessible domain and has evidence for polymerase II binding. Perhaps of interest is the fact that four SNPs (rs1333045, rs10757278, rs10757279, rs10757277 accounting for 5%, 4%, 4%, and 3% of the posterior respectively) in the 95% credible set appear to be in sites that bind *TCF7L2* and one shows some evidence of conservation (rs10757278).

3.2 Secondary Signals

It has been demonstrated a number of times that a region identified as associated to a phenotype can harbor further associated variants (Speliotes et al., 2010) (Maller et al., 2006) (the DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium et al., 2012) (Yang et al., 2012). These can represent completely independent effects on the phenotype, or modify the main effect. When two associated variants are within the same region, they will often be correlated as a result of linkage disequilibrium. In such circumstances, a single marker association test - such as the one described earlier in this chapter - can in effect mask the presence of the weaker signal. In the case of two (or more) independent signals, the signal at the variants with weaker associations may appear to be solely due to the LD to the strongly associated variant. In the case of a modifying effect, the single marker association test we've used will not have any statistical power to observe the effect. One powerful and effective way to find secondary associ-

ations in a region is by stratifying on the genotype at the main SNP in a region. For a comprehensive discussion of this and related approaches, see (Yang et al., 2012) (Purcell et al., 2007a). The conditional test used here is implemented in the PLINK package (Purcell et al., 2007b). As the regions in this study were previously implicated we adopt a less stringent threshold of $p < 10^{-3}$ as a filter for investigating secondary signals.

The conditional ("independent effect") test in PLINK is implemented via haplotypes. In short, say we have SNPs X and Y, with X being the best in a region, and Y being the SNP we wish to test for a secondary signal, with alleles A, B and C, D respectively. The null and alternative models are constructed as follows: the haplotypes in the null model are grouped based on genotype at SNP Y (thus allowing only an effect at snp X, whereas the alternative model allows each haplotype to have an individual effect. Thus, for SNPs X and Y, the two haplotype groups are:

Null model: {AC, AD} {BC,BD}

Alternative model: {AC} {AD} {BC} {BD}

These two models are then compared using a likelihood ratio test (Purcell et al., 2007b) (Purcell et al., 2007a).

Following the conditional test, we used PLINK to examine the estimated frequencies and risks conferred by individual haplotype configurations.

FTO T2D

Conditional analysis on rs17817449 reveals an additional signal at SNP rs8063946 with MAF in controls of 0.06 and RR 1.37 (95% $CI = 1.14 - 1.64$; $P = 5.9 \times 10^{-4}$) comparing the model with both the primary signal at SNP rs17817449 and

rs8063946 vs. the model with just the primary signal). The three haplotypes at rs17817449 and rs8063946 are GC (ancestral), TC, and TT, with respective relative risks of 1.68 (95% $CI = 1.40 - 2.02$), 1.37 (95% $CI = 1.15 - 1.65$), and 1. Note that the relative risk of the high risk haplotype, at 1.68, is high by GWAS standards, although the low frequency (5% in CEU Hapmap) of the low-risk TT haplotype makes precise estimation of this RR difficult. The frequencies of these haplotypes differ substantially between populations. Most obviously, the derived, protective TT haplotype emerges as the most common haplotype in JPT/CHB (47%) and YRI (41%). Four SNPs in the 95% set (rs62033408, rs10468280, rs7202296, rs7202116) have high Fst in the Human Genome Diversity Panel (HGDP), adding further weight to the evidence for population differentiation at this locus.

CDKN2A/B T2D

Previous suggestions that this region contains multiple signals (Zeggini et al., 2007) are confirmed following this denser fine-mapping. There are two key sets of SNPs, (rs10811661, rs2383208, rs10965250, rs10811660) and (rs10217762, rs10757283, rs7019778), the SNPs within each set displaying strong correlations. The model including rs12555274 alone fits the data significantly less well than that based around these pairs of SNPs. The best SNP on the conditional tests here is rs10965250 with a p-value of $p = 4.1 \times 10^{-5}$.

As expected in a region of very low combination, only three of the four possible haplotypes involving these two sets of SNPs occur with any substantial frequency in the data, and haplotype phase at the pair of SNPs is uniquely determined by the SNP genotypes. As a result, statistical models in which disease risk at this

locus depends on the two-SNP haplotypes cannot be distinguished from those in which it depends on the pair of genotypes at the two SNPs. For convenience we describe results in terms of the haplotypes, described explicitly by the pair of SNPs rs10811661 and rs10217762. The three haplotypes present are CC, TT, and TC with frequencies 16%, 59%, and 25%, respectively, listed in increasing order of disease risk, with the unobserved CT haplotype being ancestral. Estimated relative risks for the three haplotypes are 1, 1.19 (95% $CI = 1.06 - 1.34$), and 1.49 (95% $CI = 1.31 - 1.69$). We note that the relative risk for the high-risk haplotype, namely 1.49, is at the high end for recent GWAS findings for common complex diseases.

The best single SNP in our data (rs12555274) has a substantially elevated BF compared to those previously typed including those that define these haplotypes. However, whilst rs12555274 tags the high-risk haplotype reasonably well, it fails to distinguish between the two lower-risk haplotypes. Performing a conditional analysis with this top SNP identifies a secondary signal at rs10965250 which is highly significant. This pair of SNPs capture essentially the same effects as those described in terms of haplotypes above, although not quite as well.

Although the best single SNP in our data does not explain the data as well as the pair of SNPs rs1081161 and rs10217762, we cannot in principle exclude the possibility that there is another, untyped, single SNP which explains the data better than this pair of SNPs. None of the SNPs described has compelling annotations.

CDKAL1

Conditional analysis reveals an additional signal at SNP rs6456360 ($RR = 1.16$; 95% $CI = 1.08 - 1.25$), with MAF in controls of 0.47 ($p = 8.7 \times 10^{-5}$) for the con-

Table 3.2: Two-SNP additive disease model *CDKAL1* (T2D)

SNP	MAF	Relative risk (CI)
rs7756992	0.28	1.28 (1.18–1.40)
rs6456360	0.49	1.17 (1.08–1.26)

ditional test comparing the model with both the primary signal at SNP rs7756992 and rs6456360 vs. the model with just the primary signal). The data are consistent with a model in which risks at the primary SNP(s) and this second SNP combine multiplicatively. Analyses of the joint risks at rs7756992 and rs6456360 estimate the relative risk for each additional copy of both risk alleles as 1.50 (95% $CI = 1.34 - 1.68$), large by GWAS standards, and considerably larger than the effect considering the top SNP alone. Details of the model for the two SNPs are shown in Table 3.2

3.3 Disease Models

We assessed the best model for relating SNP genotype to disease risk. For most reported GWAS associations there is no significant evidence for departure from the simple model in which each additional copy of the risk allele increases disease risk by the same multiplicative factor. In this simple model, the log-odds of disease increase in an additive manner with each additional copy of the risk allele, so the model is sometimes referred to as the "additive" model. But, as has been noted elsewhere (WTCCC, 2007), the power to detect departures from this model is limited unless the true causal variant, or a SNP highly correlated with it, is typed in the study. With the much more extensive genotyping in this study, it is natural to revisit this question. We found no compelling evidence for departure from the

additive model in any region in our study.

In order to test for deviation for additivity, we extend the logistic model we introduced in Section 1.8.2:

$$logit(p) \sim \beta_0 + \beta_1 X_1$$

We achieve this by adding an additional term allowing the heterozygote to have an additional effect beyond that of the log-linear in the additive model. Specifically, we add a variable X_2 , coded 0,1,0 (corresponding to the hetorozygous genotype of AB from the three genotype categories AA,AB,BB), and a corresponding β_2. The logistic equation now becomes

$$logit(p) \sim \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

The value of β_2 corresponds to the deviation from the predicted effect of the additive model observed in samples with a heterozygous genotype. The p-value for this term can be used to compare the models. That is to say that a significant p-value suggests that the data may be better modeled in a non-additive framework.

3.4 Proportion of heritability explained

We also attempted to determine how much we have learned through fine mapping beyond what was known from the original GWAS efforts. Table 3.5 (CAD), Ta-

Region	Proportion SNPs on 500k assay (original study)	Proportion of SNPs in 95% credible set on 500k assay	Contribution to 95% credible set posterior
SORT1 (CAD)	0.07	0.04	0.04
CDKN2A/B (CAD)	0.07	0.31	0.28
CXCL12 (CAD)	0.11	0.12	0.14
1q41 (CAD)	0.07	0.05	0.06
2q36 (CAD)	0.07	0.13	0.08
FTO (T2D)	0.07	0.06	0.06
CDKN2A/B (T2D)	0.07	0.20	0.04
HHEX (T2D)	0.03	0.21	0.21
CDKAL1 (T2D)	0.06	0.12	0.07
TCF7L2 (T2D)	0.09	0.40	0.06
JAZF1 (T2D)	0.10	0.13	0.11
CTLA4 (GD)	0.11	0.33	0.07
CD25 (GD)	0.08	0.17	0.50
FCRL3 (GD)	0.07	0.11	0.10

Table 3.3: Contribution of SNPs on the original Affymetrix 500k Genotyping chip to the 95% posterior set. For each region, the table shows the proportion of 500k SNPs in the 95% set, as well as the proportion of the 95% weight placed on SNPs from the 500k Chip.

ble 3.6 (T2D) and Table 3.7 (GD) compare the top SNP after the fine mapping experiment with the top SNP in the region on each of three commercial genotyping chips (Affymetrix Human 500K SNP Array, Illumina 550K and 1.2M SNP Arrays). The "BF Ratio" column of the tables allows a direct comparison of the relative evidence for the top SNP from a particular chip to the top SNP from the FM experiment: for example, a BF ratio of 0.22 means that the posterior probability for the top SNP from the chip is smaller by a factor of 0.22 than that for the top FM SNP. The top FM SNP improved on the top SNP from Affymetrix 500K, the Illumina 550k and 1.2M arrays respectively in 13, 12, and 8 of the 14 regions, with comparisons amongst SNP-chips differing across regions. As has been widely expected, the fine mapping experiment thus indicates that typically the top GWAS SNPs will not be causal; even if the top FM SNP is not itself causal, the fact that it has a higher BF strongly suggests the top chip-SNP is not causal. Table 3.3 gives another comparison of GWAS and FM findings. It shows, for the Affymetrix 500K array, how many of the SNPs in the credible set after the fine mapping experiment were typed in the original GWAS. In most regions, the vast majority of the SNPs which were important after the fine mapping experiment were not typed in the original association study.

These results show that there can be strong evidence in favour of one SNP over another, even when there are only relatively small differences in effect sizes (most strikingly *CTLA4* for GD, but several other SNP comparisons), and hence only slight differences in heritability explained by the locus. The intuition is that in large samples, small differences in effect sizes can result in strong evidence in favour of one SNP over another in fine mapping. Thus while the fine mapping experiment does substantially change the picture as to potential causal SNPs com-

pared to the original association study, the new top SNPs do not greatly change the proportion of heritability explained by the locus. One caveat is that this requires a high level of accuracy in genotyping, and can be affected by even relatively small population differences.

3.5 Proportion of SNPs captured

One limitation with our study is that we were restricted to genotyping only those SNPs known at the time of design and our specific sequencing efforts, and so we could simply have missed the causative variants or the most associated SNP markers, particularly at lower MAFs. However, we used data from the 1,000 Genomes Project to assess our coverage of variants in the regions studied (Table 3.4). For example, in the June 2011 1,000 Genomes release, there were 1920 and 1601 SNPs in Caucasian samples in our FM regions with MAF in the range 1%-2% and 2%-4% respectively. Of these, we directly typed 92 (4.8%) and 240 (15%) respectively, and in addition obtained good quality imputation data on a further 746 and 1065 SNPs respectively. Our project typed a much denser set of SNPs than GWAS chips, leading to improved imputation. Thus in these frequency ranges, we obtained actual, or good imputed, data on 44% and 82% of these variants. For all higher MAF ranges, the proportion captured or well-imputed is at least 85% (Table 3.4). Coverage via genotyping or imputation of variants with MAF $< 1\%$ is much lower, but we do not believe this undermines our main conclusions.

3.6 Contribution of imputed SNPs

Table 3.8 gives an overview of what we have learned from the 1000 genomes imputation and shows the proportion of the imputed 95% (and 99%) posterior weight that is placed on imputed SNPs.

This table has several interesting attributes. First, in regions such as *CDKN2A/B* (both CAD and T2D), *CDKAL1*, there are a substantial number of imputed SNPs in the 95% posterior set, but these SNPs have a relatively low contribution to the overall weight. These are regions where the highest weighted genotyped SNPs clearly stand out, but not so much that they are the overwhelming contributors to the 95% posterior weight. In certain other regions such as *TCF7L2* and *CTLA4*, there are individual SNPs which contribute the majority of the weight to the 95% posterior weight, and this is reflected in the very small number of imputed SNPs in the 95% set, as well as even smaller contributions to the overall weight. For regions where we do not believe we have learned anything in the fine mapping, the imputed results reflect that as well. In *JAZF1*, the imputed data corresponds to nearly half of the 95% set, as well as nearly half of the posterior weight. Similar results are observed for *1q41*, though the proportions are closer to 0.33. Again, for *FCRL3*, the proportion of imputed SNPs in the 95% causal is about 0.3 which is virtually identical to the overall proportion of imputed SNPs, and the proportion of the posterior weight is about 0.24.

3.7 Discussion

This chapter has covered a number of more in depth analyses of the fine-mapping data. First, we assessed the extent to which genome annotation provides insight

into the underlying mechanisms of putative causal variants. While these annotation databases help shed light on potential avenues for further functional investigation, it is clear from the work presented here that genome-scale annotation databases are not likely (at present) to provide understanding of the mechanism of action of a particular associated variant. Specifically in the context of fine-mapping, it would be ideal to use such annotations towards differentiating between highly (or perfectly) correlated variants. However, this is also challenging at the moment given the variability in coverage and depth of currently available annotation datasets.

We then described assesments of secondary signals and deviation from additivity of putative causal variants. While we observed several regions with evidence for secondary signals, in the majority regions we did not observe substantial evidence of complexity in signal beyond the single-snp of additive effect model of regional association. However, this is likely due, at least in part, to the issue of limited power.

We then moved on to aggregate analyses of heritability and an assessment of SNP coverage and the contribution of imputation. The issue of coverage is a particularly important one in the context of fine mapping. This experiment utilized a genotyping platform that was (at the time) able to provide the best available coverage in terms of proportion of type-able SNPs. Imputation was still required to capture the remainder of SNPs, and was able to do so very successfully with currently available reference panels. The obvious question resulting from these overall analyses is how well would we do if we skipped the extra step of direct genotyping for fine-mapping and simply attempted to fine map using imputation directly from the original GWAS array? This question will be covered in depth by

the next chapter.

Minor Allele Frequency Range	Genotyped in FM	Genotyped or well imputed[a]	1000 Genomes	Proportion Captured
MAF $< .01$	125	1665	20306 (12263[b])	.082 (0.14[b])
$.01 \leq$ MAF $< .02$	92	838	1920	0.44
$.02 \leq$ MAF $< .04$	240	1305	1601	0.82
$.04 \leq$ MAF $< .06$	256	675	794	0.85
$.06 \leq$ MAF $< .10$	260	808	901	0.9
$.10 \leq$ MAF $< .20$	614	972	1083	0.9
$.20 \leq$ MAF	1784	2824	3196	0.88

Table 3.4: Imputation and minor allele frequency based on a reference panel of 143 European Haplotypes in the 1,000 Genomes Project June 2011 data release.
(a) SNPs were deemed to be well imputed if average maximum posterior (as returned by IMPUTE1) was greater than 0.98, and the IMPUTE1 info score above 0.8.
(b) Numbers in parentheses refer to the SNP count (column 4) and proportion (column 5) restricted to SNPs with at least two observations of the minor allele in the reference panel.

Region	SNP Set	SNP ID	Risk allele frequency (controls)	Risk Allele	Effect Size (RR)	log BF	BF Ratio	λ_s
SORT1	Genotyped in FM	rs3832016	0.79	A	1.22 (1.11-1.35)	2.68	1.00	1.006
	Affy 500k	rs599839	0.78	A	1.17 (1.07-1.29)	1.59	0.08	1.004
	Ilmn 550k	rs611917	0.69	A	1.17 (1.07-1.27)	1.96	0.19	1.005
	Ilmn 1M	rs660240	0.79	G	1.22 (1.10-1.34)	2.54	0.72	1.006
CDKN2A/B	Genotyped in FM	rs1537370	0.47	A	1.40 (1.29-1.51)	14.03	1.00	1.028
	Affy 500k	rs6475606	0.47	A	1.40 (1.29-1.51)	13.95	0.83	1.028
	Ilmn 550k	rs10116277	0.47	A	1.40 (1.29-1.51)	13.99	0.92	1.028
	Ilmn 1M	rs1537370	0.47	A	1.40 (1.29-1.51)	14.03	1.00	1.028
CXCL12	Genotyped in FM	rs34161818	0.84	A	1.21 (1.08-1.35)	1.56	1.00	1.004
	Affy 500k	rs977754	0.85	A	1.18 (1.06-1.33)	1.16	0.39	1.003
	Ilmn 550k	rs977754	0.85	A	1.18 (1.06-1.33)	1.16	0.39	1.003
	Ilmn 1M	rs977754	0.85	A	1.18 (1.06-1.33)	1.16	0.39	1.003
1*q*41	Genotyped in FM	rs2936023	0.85	T	1.21 (1.08-1.36)	1.58	1.00	1.004
	Affy 500k	rs17464857	0.85	A	1.17 (1.05-1.31)	0.93	0.22	1.003
	Ilmn 550k	rs17464857	0.85	A	1.17 (1.05-1.31)	0.93	0.22	1.003
	Ilmn 1M	rs11485123	0.85	G	1.17 (1.05-1.31)	0.93	0.22	1.003
2*q*36	Genotyped in FM	rs2673145	0.41	A	1.20 (1.11-1.30)	3.76	1.00	1.008
	Affy 500k	rs2943646	0.64	G	1.19 (1.10-1.29)	3.01	0.18	1.007
	Ilmn 550k	rs2972153	0.67	G	1.21 (1.11-1.32)	3.45	0.49	1.008
	Ilmn 1M	rs2972153	0.67	G	1.21 (1.11-1.32)	3.45	0.49	1.008

Table 3.5: For each CAD region the table shows the top SNP (based on Bayes Factors) after the fine mapping experiment amongst various sets of SNPs, those: passing QC in the FM experiment; from the Affymetrix Human 500K SNP Array; from the Illumina 550K SNP array; and from the Illumina 1.2M SNP Array. Successive columns of the table then show: risk allele frequency; risk allele; relative risk (and 95% CI); the log_10 of the Bayes Factor for the SNP; the ratio of the Bayes Factor for the SNP to the Bayes Factor for the top ranked FM SNP, and the contribution of that SNP to the sibling relative risk (λ_s). For regions with evidence for a second associated SNP the table shows the combined contribution of the pair of SNPs to the sibling relative risk.

Region	SNP Set	SNP ID	Risk allele frequency (controls)	Risk Allele	Effect Size (RR)	log BF	BF Ratio	λ_s
FTO	Genotyped in FM	rs17817449	0.4	C	1.26 (1.17-1.36)	6.69	1	1.013
**	Combined	rs17817449,rs8063946						1.018
	Affy 500k	rs8050136	0.4	A	1.26 (1.17-1.36)	6.49	0.62	1.013
	Ilmn 550k	rs8050136	0.4	A	1.26 (1.17-1.36)	6.49	0.62	1.013
	Ilmn 1M	rs17817449	0.4	C	1.26 (1.17-1.36)	6.69	1	1.013
CDKN2A/B	Genotyped in FM	rs12555274	0.23	C	1.26 (1.15-1.37)	4.75	1	1.011
**	Combined	rs10811661,rs10217762						1.012
	Affy 500k	rs10811661	0.83	A	1.27 (1.14-1.41)	3.46	0.05	1.007
	Ilmn 550k	rs2383208	0.82	A	1.26 (1.13-1.40)	3.2	0.03	1.007
	Ilmn 1M	rs2383208	0.82	A	1.26 (1.13-1.40)	3.2	0.03	1.007
HHEX	Genotyped in FM	rs10882098	0.59	G	1.21 (1.12-1.31)	4.01	1	1.009
	Affy 500k	rs5015480	0.59	G	1.20 (1.11-1.30)	3.8	0.61	1.008
	Ilmn 550k	rs5015480	0.59	G	1.20 (1.11-1.30)	3.8	0.61	1.008
	Ilmn 1M	rs5015480	0.59	G	1.20 (1.11-1.30)	3.8	0.61	1.008
CDKAL1	Genotyped in FM	rs7756992	0.27	G	1.29 (1.19-1.40)	6.71	1	1.014
**	Combined	rs7756992,rs6456360						1.021
	Affy 500k	rs9460546	0.31	C	1.25 (1.15-1.35)	5.38	0.05	1.012
	Ilmn 550k	rs7756992	0.27	G	1.29 (1.19-1.40)	6.71	1	1.014
	Ilmn 1M	rs7756992	0.27	G	1.29 (1.19-1.40)	6.71	1	1.014
TCF7L2	Genotyped in FM	rs7903146	0.3	A	1.40 (1.29-1.52)	13.61	1	1.027
	Affy 500k	rs4506565	0.32	T	1.37 (1.27-1.49)	12.24	0.04	1.024
	Ilmn 550k	rs7903146	0.3	A	1.40 (1.29-1.52)	13.61	1	1.027
	Ilmn 1M	rs7903146	0.3	A	1.40 (1.29-1.52)	13.61	1	1.027
JAZF1	Genotyped in FM	rs12531540	0.51	G	1.14 (1.06-1.23)	1.75	1	1.004
	Affy 500k	rs1859687	0.06	C	1.25 (1.08-1.45)	1.12	0.23	1.003
	Ilmn 550k	rs498475	0.37	G	1.14 (1.05-1.23)	1.44	0.49	1.004
	Ilmn 1M	rs498475	0.37	G	1.14 (1.05-1.23)	1.44	0.49	1.004

Table 3.6: For each T2D region the table shows the top SNP (based on Bayes Factors) after the fine mapping experiment amongst various sets of SNPs, those: passing QC in the FM experiment; from the Affymetrix Human 500K SNP Array; from the Illumina 550K SNP array; and from the Illumina 1.2M SNP Array. Successive columns of the table then show: risk allele frequency; risk allele; relative risk (and 95% CI); the log_10 of the Bayes Factor for the SNP; the ratio of the Bayes Factor for the SNP to the Bayes Factor for the top ranked FM SNP, and the contribution of that SNP to the sibling relative risk (λ_s). For regions with evidence for a second associated SNP the table shows the combined contribution of the pair of SNPs to the sibling relative risk.

Region	SNP Set	SNP ID	Risk allele frequency (controls)	Risk Allele	Effect Size (RR)	log BF	BF Ratio	λ_s
CTLA4	Genotyped in FM	rs11571297	0.51	A	1.39 (1.29-1.50)	16.08	1.00	1.027
	Affy 500k	rs3087243	0.55	G	1.38 (1.28-1.49)	14.89	0.06	1.025
	Ilmn 550k	rs231804	0.57	A	1.37 (1.27-1.48)	13.58	0.00	1.023
	Ilmn 1M	rs11571291	0.58	A	1.38 (1.28-1.48)	13.87	0.01	1.024
CD25/IL2RA	Genotyped in FM	rs10905669	0.23	A	1.20 (1.10-1.30)	3.10	1.00	1.007
	Affy 500k	rs10905669	0.23	A	1.20 (1.10-1.30)	3.10	1.00	1.007
	Ilmn 550k	rs7090530	0.60	A	1.16 (1.08-1.25)	2.49	0.25	1.005
	Ilmn 1M	rs10905669	0.23	A	1.20 (1.10-1.30)	3.10	1.00	1.007
FCRL3	Genotyped in FM	rs11264798	0.52	C	1.17 (1.09-1.26)	3.00	1.00	1.006
	Affy 500k	rs2785663	0.58	C	1.17 (1.08-1.26)	2.80	0.63	1.006
	Ilmn 550k	rs2785665	0.58	A	1.17 (1.08-1.26)	2.76	0.58	1.006
	Ilmn 1M	rs11264798	0.52	C	1.17 (1.09-1.26)	3.00	1.00	1.006

Table 3.7: For each GD region the table shows the top SNP (based on Bayes Factors) after the fine mapping experiment amongst various sets of SNPs, those: passing QC in the FM experiment; from the Affymetrix Human 500K SNP Array; from the Illumina 550K SNP array; and from the Illumina 1.2M SNP Array. Successive columns of the table then show: risk allele frequency; risk allele; relative risk (and 95% CI); the log_10 of the Bayes Factor for the SNP; the ratio of the Bayes Factor for the SNP to the Bayes Factor for the top ranked FM SNP, and the contribution of that SNP to the sibling relative risk (λ_s). For regions with evidence for a second associated SNP the table shows the combined contribution of the pair of SNPs to the sibling relative risk.

Phenotype	Region	Genotyped SNPs	Imputed SNPs	Proportion Imputed	Proportion 95% SNPs imputed	Proportion 95% posterior imputed	Proportion 99% SNPs imputed	Proportion 99% posterior imputed
CAD	*SORT1*	230	65	0.22	0.17	0.13	0.23	0.14
	CDKN2A/B	518	92	0.15	0.29	0.06	0.15	0.06
	CXCL12	362	196	0.35	0.36	0.28	0.37	0.28
	1q41	353	214	0.38	0.37	0.28	0.39	0.31
	2q36	342	83	0.20	0.18	0.96	0.18	0.93
T2D	*CDKN2A/B*	514	89	0.15	0.21	0.08	0.26	0.09
	FTO	206	46	0.18	0.15	0.15	0.19	0.16
	HHEX	559	204	0.27	0.46	0.36	0.31	0.33
	CDKAL1	496	188	0.28	0.10	0.09	0.20	0.11
	TCF7L2	156	62	0.28	0	0	0.11	0.01
	JAZF1	338	256	0.43	0.46	0.41	0.45	0.41
GD	*CTLA4*	292	84	0.22	0	0	0.06	0.01
	CD25	425	119	0.22	0.26	0.29	0.21	0.29
	FCRL3	606	245	0.29	0.29	0.24	0.30	0.24

Table 3.8: Contribution of imputed data to results and proportion of the imputed 95% posterior weight that is placed on non-imputed SNPs. For each region, the number of Genotyped SNPs is shown, followed by the number of QC-passing imputed SNPs. The remaining columns are all based on the combined genotyped/imputed set of SNPs. Proportion imputed is, for each region, the proportion of SNPs contributed by imputation. Proportion 95% SNPs imputed is the proportion of imputed SNPs in the 95% credible set, followed by the proportion of regional posterior placed on imputed SNPs in the 95% credible set. The final two columns show the proportion of imputed SNPs in the 99% credible set and the proportion of posterior placed on imputed SNPs in the 99% credible set.

Chapter 4

Imputation

The new generation of sequencing technologies has enabled the cost-effective deep-resequencing of thousands of samples, which provides a more complete catalogue of variation in regions of interest prior to fine mapping. Additionally, genome-wide resequencing data for hundreds of samples is available from the 1000 genomes project. If this sort of sequence data could be used to accurately infer genotypes at both known and novel SNPs in an original disease-study sample, it has the potential to greatly improve fine mapping efficiency. In this chapter we attempt to answer the question of whether genotype imputation from GWAS data can be used to fine map associated regions and identify which variant(s) are likely to be causal. We also investigate whether imputed data can be used to identify sets of variants which are very unlikely to be causal.

We attempted to assess the accuracy of imputing from resequencing and genotype data, as well as the effectiveness of using such imputation to improve fine mapping efficiency.

4.1 Imputation of the Regions of Interest Using Four Reference Panels

Genotype imputation has been widely used for several years, particularly for increasing coverage in GWAS and for meta-analyses (Marchini and Howie, 2010) . One area in which imputation could potentially be useful is that of fine mapping of associated regions. With large and highly accurate publicly available genone-wide reference data sets, it may be possible to avoid the time, effort and expense of dense genotyping in many samples by simply imputing into a GWAS dataset. A second way that imputation might prove useful for fine mapping is in limiting the number of SNPs to be genotyped. Even if imputed genotype datasets are inadequate for direct use for fine mapping, they may be sufficiently accurate to be used to identify sets of SNPs so unlikely to be causal that they need not be genotyped. These are the two possible approaches to using imputation in the context of fine mapping that I will be discussing in this chapter.

There have been many evaluations of the accuracy of imputed genotype data (Nothnagel et al., 2009) (Marchini and Howie, 2010) (Huang et al., 2009), but the focus of these has been on "how accurately do you impute this SNP" on a marker-by-marker basis. While this type of accuracy is critical to the utility of imputed genotype data, we are mainly concerned with a number of bigger picture questions. We want to get a sense, in a number of different ways, of how useful imputation can be in fine mapping. We want to know how accurately can we get a sense of what's going on in a particular region, which is a fundamentally different question. We are interested in relative differences between associated SNPs in each region, which is one of the primary goals of fine mapping. These SNPs

will frequently be highly correlated and difficult to differentiate. With enough samples genotyped, a fine mapping study can differentiate between SNPs with very small differences between then. It is asking a lot of imputation to achieve such high accuracy, especially given the density of SNPs in GWAS target samples, and evaluations of imputation to date have not generally dealt with these questions.

We sought to address these questions as follows, using our fine mapping data (as described in Chapter 2) and publicly available reference panels. We treated the fine mapping dataset as the "gold standard", containing a "complete" catalogue of SNPs for our regions. We took the FM data for all 8000 FM samples at all SNPs, and thinned the SNPs such that we only look at SNPs on the Affymetrix 500K genotyping array. We used this thinned dataset as the target for imputation (eg. these are the samples for which genotypes are imputed). We then imputed data for all the other SNPs which actually have genotype data (but which were hidden by the thinning procedure). This forms the basis for our comparisons. We repeated this using several reference panels. After generating these datasets, we analysed the data in the same way as in the FM experiment (see Chapter 2). Having done that, we assessed these results in several ways: visual inspection of the signal plots, assessment of how much of each regions posterior probability is captured, how this posterior probability is distributed across SNPs, and how well we can exclude SNPs as likely to be causal. Our main focus across these analyses was determining whether any of the reference panels perform adequately for fine mapping, and how much of a difference the choice of reference panel will make. There have been numerous evaluations of the SNP-by-SNP accuracy of genotype imputation (Nothnagel et al., 2009) (Marchini and Howie, 2010) (Huang et al., 2009), and the results and assessments in this chapter focus on the aspects

particular to fine mapping.

There are a number of prominent imputation methods that are broadly similar in performance (Marchini et al., 2007) (Li et al., 2010) (Browning and Browning, 2007). In order to simplify our analyses we're going to focus on one of these methods. We used the IMPUTE package to perform the imputation (Marchini et al., 2007).

4.1.1 Samples

We used the dataset from the fine mapping experiment described in Chapter 2 as the target samples for imputation. The complete dataset consists of approximately 8000 samples, densely genotyped across 13 regions with a total of 3668 QC-passing SNPs. These samples are comprised of approximately 2000 case samples for each of T2D, CAD and GD, and a further 2000 58C/NBS control samples (see Chapter 2 for details).

4.1.2 SNPs

We treated the set of QC-passing SNPs from the fine mapping as the complete set of SNPs for the fine mapping regions we are looking at. In order to mimic a GWAS, we thinned the data to include only SNPs that are on the Affymetrix 500K genotyping array. We then attempted to impute data for all of the non-500k SNPs in our regions.

4.1.3 Reference Panels

We compared imputation using several different reference panels.

Fine-mapping Controls

We sought to create a "best case" reference panel using our fine mapping genotype data. We phased the complete set of 8000 samples using fastPHASE (Scheet and Stephens, 2006). We then randomly selected 100 of the 58C/NBS control samples, and used their haplotypes as a reference panel. These samples were excluded from the set of imputation target samples. These haplotypes are generated from high quality genotypes, as part of the same genotyping experiment as the imputation targets, with data at a very high density and at all target sites. The data was generated in the same experiment as the target samples, removing most or all experimental variation between the reference and the target panels. Therefore, this reference panel is effectively a best case scenario, with the high accuracy of a genotyped reference panel and the density of a resequenced reference panel. This reference panel gives us an approximate upper bound of how well imputation is likely to do for the particular set of samples and SNPs we are imputing into.

Regional Resequencing

Data from the WTCCC regional resequencing project formed the second imputation reference panel. This data was generated in parallel with the fine mapping data, in order to assess the performance of a high-throughput sequencing platform for discovery of variants in GWAS association regions (see Chapter 2 for a detailed description). This dataset contained 32 samples, taken from the CEPH samples included in the HapMap project (Altshuler et al., 2005). Five (of 16) resequenced regions overlapped with the regions included in the fine mapping project; *SORT1*, *CDKN2A/B*, *FTO*, *HHEX* 2*q*36. Only QC passing, polymorphic

SNPs were included, as described in Chapter 2. This data was phased using the PHASE package (Stephens et al., 2001b).

HapMap2

Phased haplotype data in 60 individuals from the CEU population from the HapMap2 project made up the third reference panel. This panel is generated from directly genotyped data, which will generally be more accurate than the reference panels generated from resequenced data. However, the SNP density is lower than in the resequenced panels.

1000 Genomes

Haplotype data for 143 Caucasian samples (samples from the CEPH population (CEU), individuals from England and Scotland (GBR), Iberian populations from Spain (IBS) and Tuscan individuals (TSI)) from the June 2010 release of the 1000 Genomes project form the final reference panel. These haplotypes are generated from genotypes called from low-pass resequencing data. While these genotypes has proven to be quite accurate when compared to direct genotyping (1000 Genomes Project Consortium, 2010), this type of data will still be more error-prone than directly genotyped datasets such as HapMap2. However, the SNP density for this panel should be substantially higher than direct genotyping.

4.1.4 Association Testing

The imputed data was filtered to include only SNPs which were well imputed, as determined by confidence measures provided by IMPUTE and SNPTEST (Mar-

chini et al., 2007). SNPs with info $< .4$ or certainty $< .9$ were filtered out from the imputed data.

4.2 Results

Figure 4.1 through Figure 4.14 show association signal for the imputed data for each of the imputation reference panels, as well as for the full fine mapping genotype data. These are similar to the plots shown earlier in Chapter 2. For regions included in the WTCCC resequencing reference panel there are five signal plots shown, while four are shown for the rest of the regions.

These plots reveal several features of the imputed data. It is immediately apparent that while general patterns are similar across the imputation sources, there is non-trivial variation in signal patterns. This is true not just for the high quality 1000 genomes panel, but even our "gold standard" genotyped reference panel is substantially different from the real data in many of the regions. For the more strongly associated regions, there is particular variability in the association peak, such that the location of the peak shifts, contains noticeably fewer SNPs or even disappears completely. There is also substantial variability in the size of 95%/99% credible sets, though for regions showing stronger signal this is less pronounced (e.g. *TCF7L2*).

We will now focus on the imputed results for the *FTO*, *HHEX* and *TCF7L2* regions. All three regions are robustly associated to T2D in our real FM data, but represent three distinct patterns of association and will provide a broad sense of how well imputation performs.

4.2.1 *FTO*

In the real FM results, the 95% credible set for *FTO* contains 33 SNPs which are all highly correlated, and have very similar posterior weights. For the "best case" genotyped reference panel, the results (Figure 4.6) closely resemble the real FM results, both visually and in terms of the posterior weight distribution and the 95% credible set size. The 1000 genomes performs quite well as well, though the log Bayes Factors spread a bit in both directions for the top SNPs, and more SNPs are included in both the 95% and 99% sets. All reference panels retain the general signal shape across the region, but the close correlation between the top SNPs is not retained except in the imputation with the genotyped reference panel.

4.2.2 *HHEX*

In the real results for *HHEX*,there is a relatively small 95% credible set with 14 SNPs, but the strength of signal in our data is only just strong enough to differentiate these SNPs. In this region, both the genotyped reference panels and 1000 genomes perform reasonably well, though for the 1000 genomes there is a substantial increase in the size of the 99% credible set, which more than doubles in size (Figure 4.5). For both of these panels, the general pattern of signal is retained in the area to the right where the signal is strong. The genotyped reference panel retains the best SNP, as does the HapMap imputation, while this SNP is not present in both the QC-passing 1000 genomes imputed data and the resequenced reference panel.

4.2.3 *TCF7L2*

There is a sharp peak in the real data for *TCF7L2* with a clearly established and replicated best SNP, accounting for 75% of the posterior weight. While the imputed results (Figure 4.11) for the genotyped reference panel in this region visually appear to be broadly consistent with the real results, as will be seen in Figure 4.17(c) in the next section, there is substantial variability among the top SNPs. The best SNP in the real results (with 75% of posterior weight) has about 10% of the weight in the genotyped imputation, 35% in the 1000 genomes and a mere 3% in the HapMap imputed data. The results across these three regions, especially in *TCF7L2*, a region with a simple, clear picture of association in the real results illustrate that fine mapping via imputation as attempted here will present many challenges.

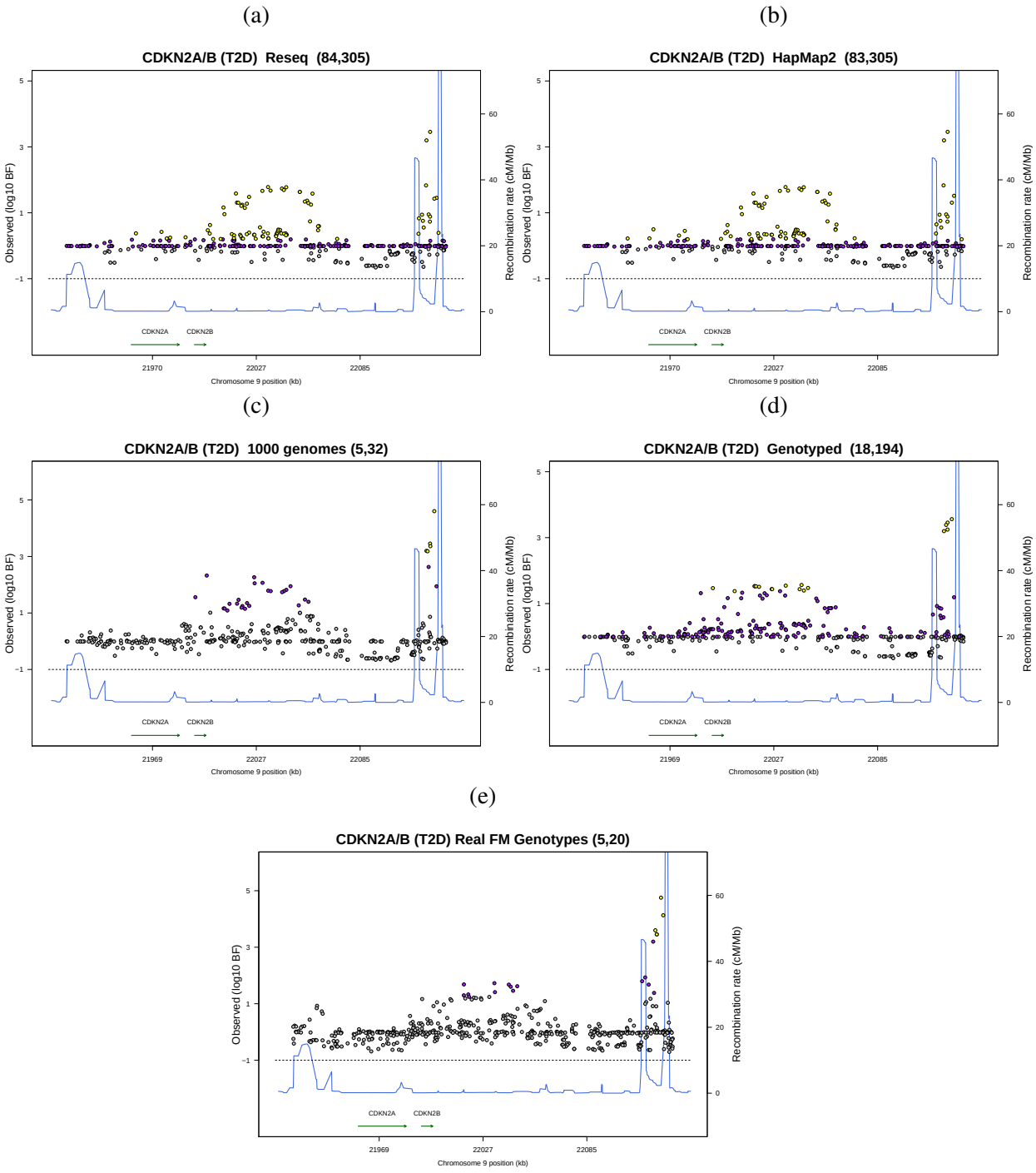


Figure 4.1: Imputed signal for the region *CDKN2A/B* T2D for each of 4 reference panels, as well as in the FM genotype data. The x-axis corresponds to genomic position, and the y-axis corresponds to log_{10}BF. Recombination is shown in blue on the right y-axis, and genes are shown in green at the bottom. SNPs are coloured by membership in the 95% (yellow) and 99% (purple) credible sets or grey if neither. The number of SNPs in each of the 95 (99) credible sets is shown in the title of each subfigure. Subfigure a-d correspond to: (a) WTCCC resequenced reference panel, (b) HapMap2 reference panel, (c) 1000 genomes reference panel, (d) genotyped reference panel. Subfigure (e) shows the results for the full fine mapping genotyped data.

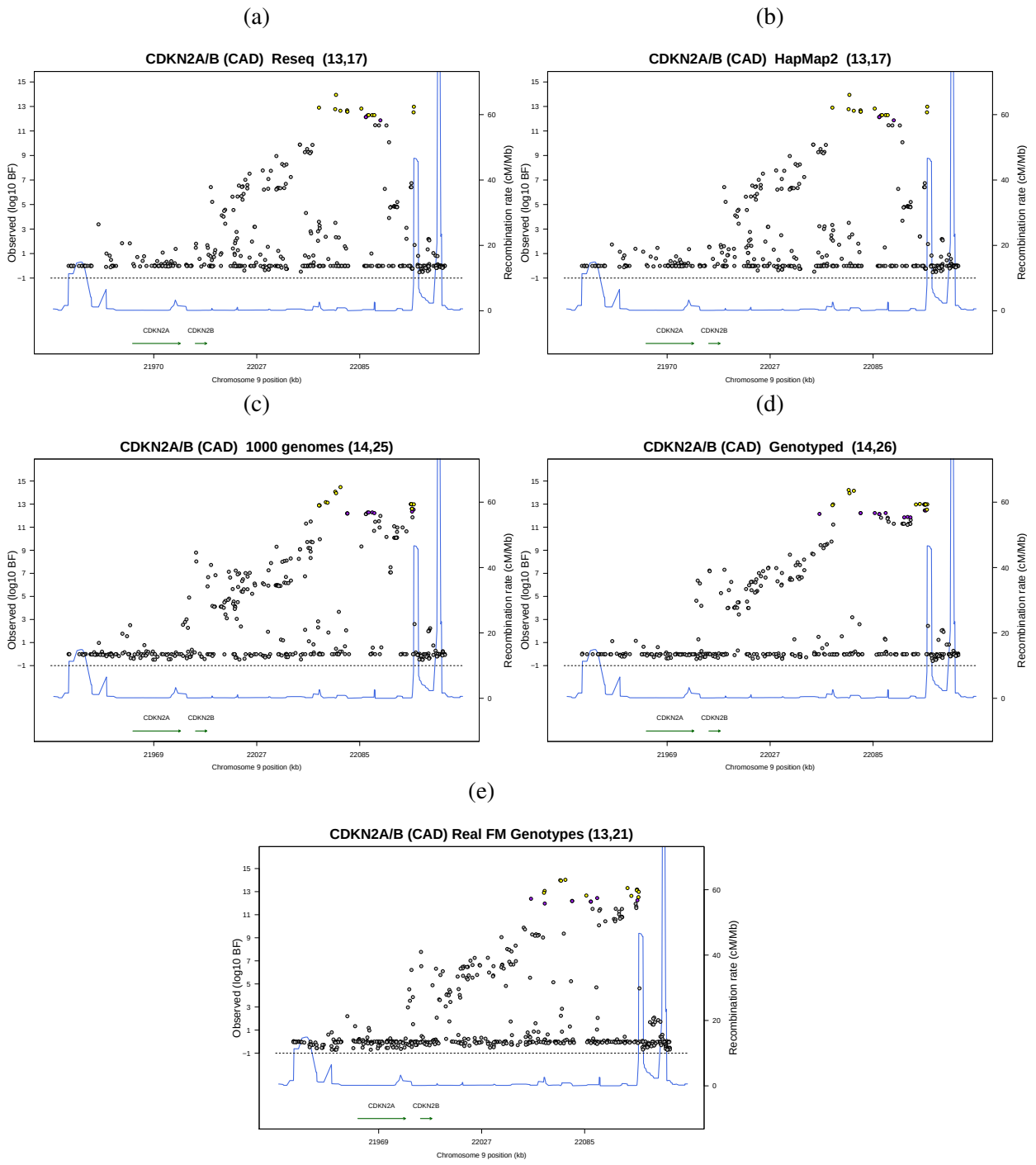


Figure 4.2: Imputed signal for the region *CDKN2A/B* CAD for each of 4 reference panels, as well as in the FM genotype data. The x-axis corresponds to genomic position, and the y-axis corresponds to log_{10}BF. Recombination is shown in blue on the right y-axis, and genes are shown in green at the bottom. SNPs are coloured by membership in the 95% (yellow) and 99% (purple) credible sets or grey if neither. The number of SNPs in each of the 95 (99) credible sets is shown in the title of each subfigure. Subfigure a-d correspond to: (a) WTCCC resequenced reference panel, (b) HapMap2 reference panel, (c) 1000 genomes reference panel, (d) genotyped reference panel. Subfigure (e) shows the results for the full fine mapping genotyped data.

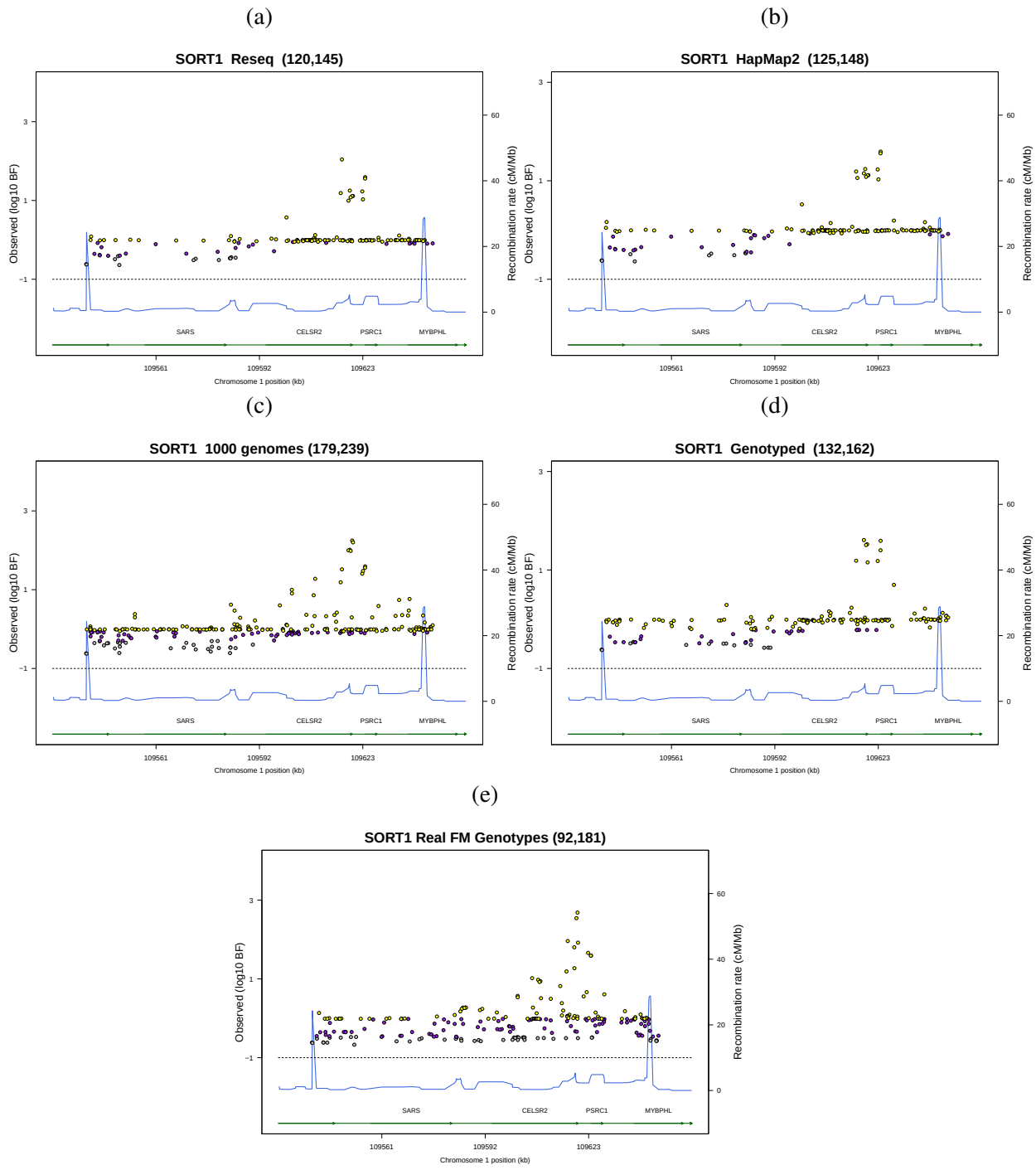


Figure 4.3: Imputed signal for the region *SORT1* CAD for each of 4 reference panels, as well as in the FM genotype data. The x-axis corresponds to genomic position, and the y-axis corresponds to log_{10}BF. Recombination is shown in blue on the right y-axis, and genes are shown in green at the bottom. SNPs are coloured by membership in the 95% (yellow) and 99% (purple) credible sets or grey if neither. The number of SNPs in each of the 95 (99) credible sets is shown in the title of each subfigure. Subfigure a-d correspond to: (a) WTCCC resequenced reference panel, (b) HapMap2 reference panel, (c) 1000 genomes reference panel, (d) genotyped reference panel. Subfigure (e) shows the results for the full fine mapping genotyped data.

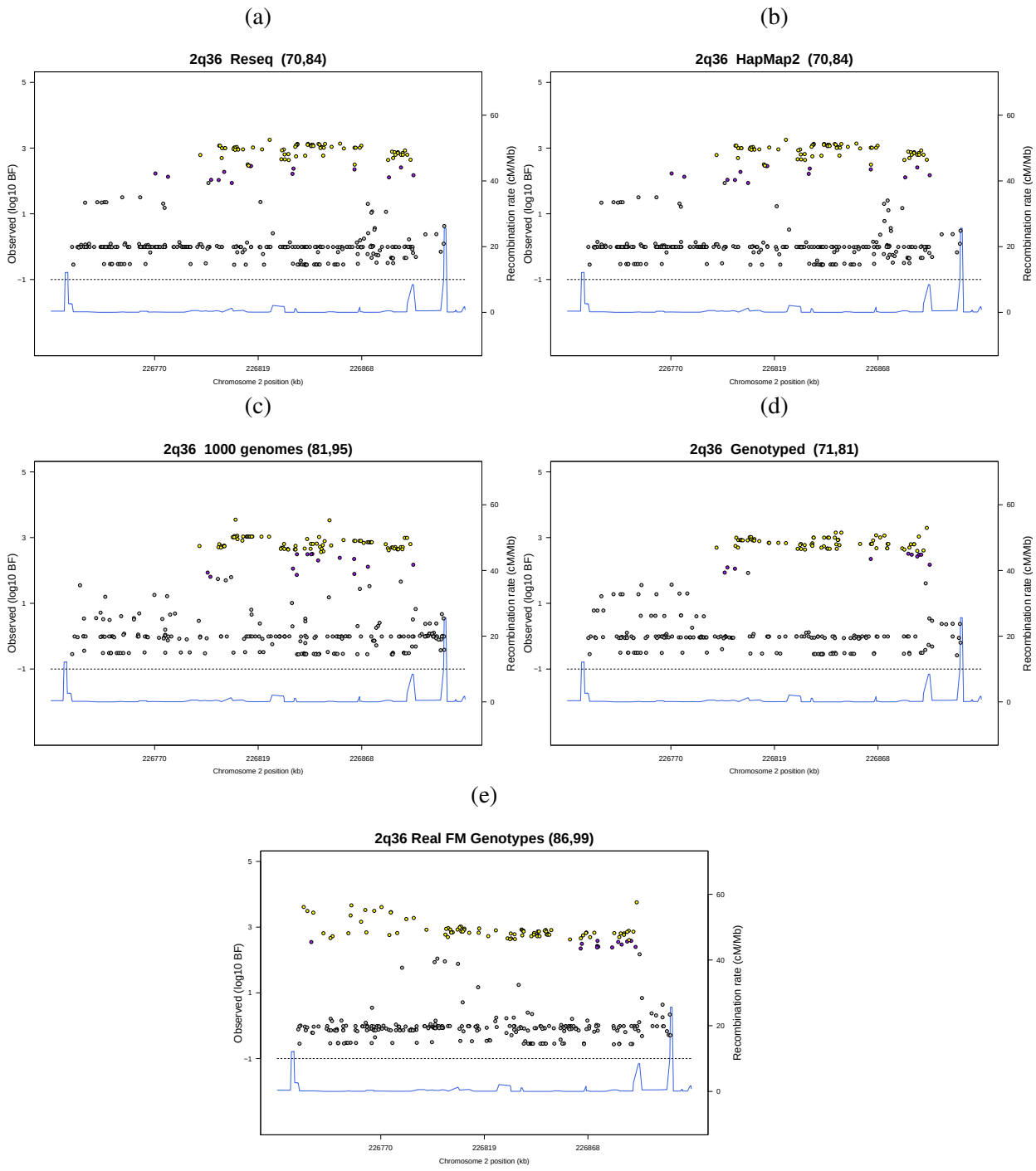


Figure 4.4: Imputed signal for the region 2q36 CAD for each of 4 reference panels, as well as in the FM genotype data. The x-axis corresponds to genomic position, and the y-axis corresponds to log_{10}BF. Recombination is shown in blue on the right y-axis, and genes are shown in green at the bottom. SNPs are coloured by membership in the 95% (yellow) and 99% (purple) credible sets or grey if neither. The number of SNPs in each of the 95 (99) credible sets is shown in the title of each subfigure. Subfigure a-d correspond to: (a) WTCCC resequenced reference panel, (b) HapMap2 reference panel, (c) 1000 genomes reference panel, (d) genotyped reference panel. Subfigure (e) shows the results for the full fine mapping genotyped data.

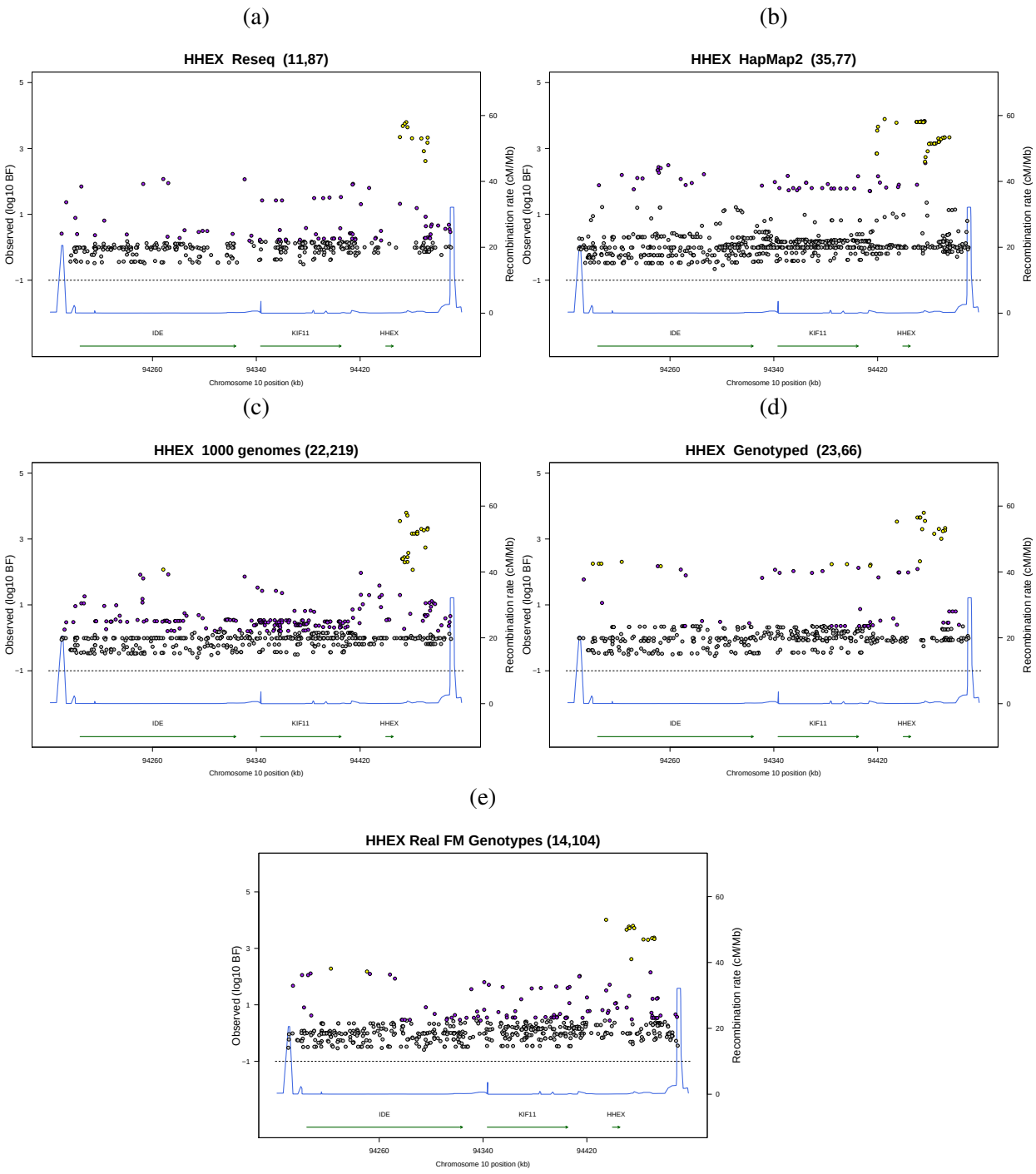


Figure 4.5: Imputed signal for the region *HHEX* T2D for each of 4 reference panels, as well as in the FM genotype data. The x-axis corresponds to genomic position, and the y-axis corresponds to log_{10}BF. Recombination is shown in blue on the right y-axis, and genes are shown in green at the bottom. SNPs are coloured by membership in the 95% (yellow) and 99% (purple) credible sets or grey if neither. The number of SNPs in each of the 95 (99) credible sets is shown in the title of each subfigure. Subfigure a-d correspond to: (a) WTCCC resequenced reference panel, (b) HapMap2 reference panel, (c) 1000 genomes reference panel, (d) genotyped reference panel. Subfigure (e) shows the results for the full fine mapping genotyped data.

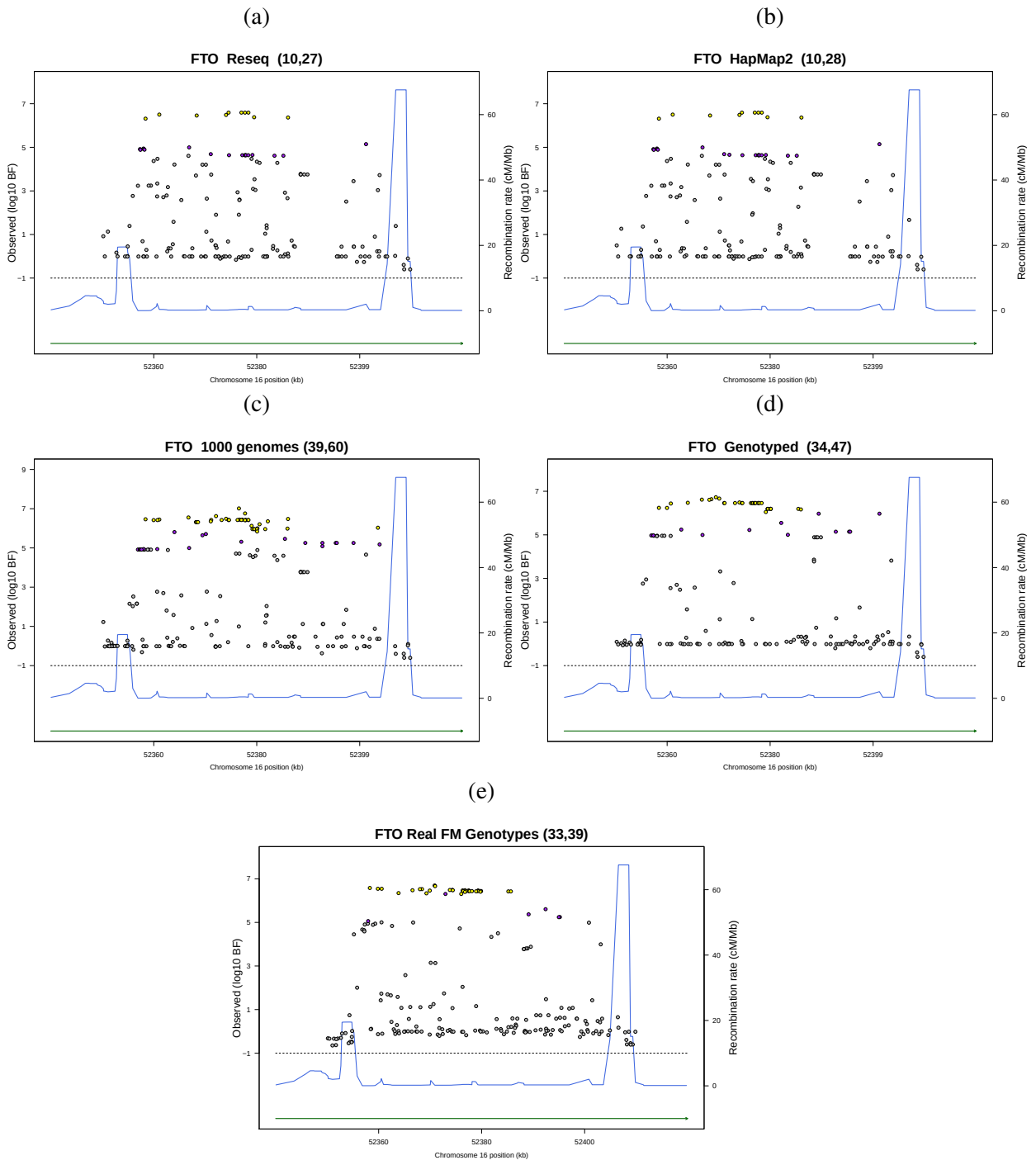


Figure 4.6: Imputed signal for the region *FTO* T2D for each of 4 reference panels, as well as in the FM genotype data. The x-axis corresponds to genomic position, and the y-axis corresponds to log_{10}BF. Recombination is shown in blue on the right y-axis, and genes are shown in green at the bottom. SNPs are coloured by membership in the 95% (yellow) and 99% (purple) credible sets or grey if neither. The number of SNPs in each of the 95 (99) credible sets is shown in the title of each subfigure. Subfigure a-d correspond to: (a) WTCCC resequenced reference panel, (b) HapMap2 reference panel, (c) 1000 genomes reference panel, (d) genotyped reference panel. Subfigure (e) shows the results for the full fine mapping genotyped data.

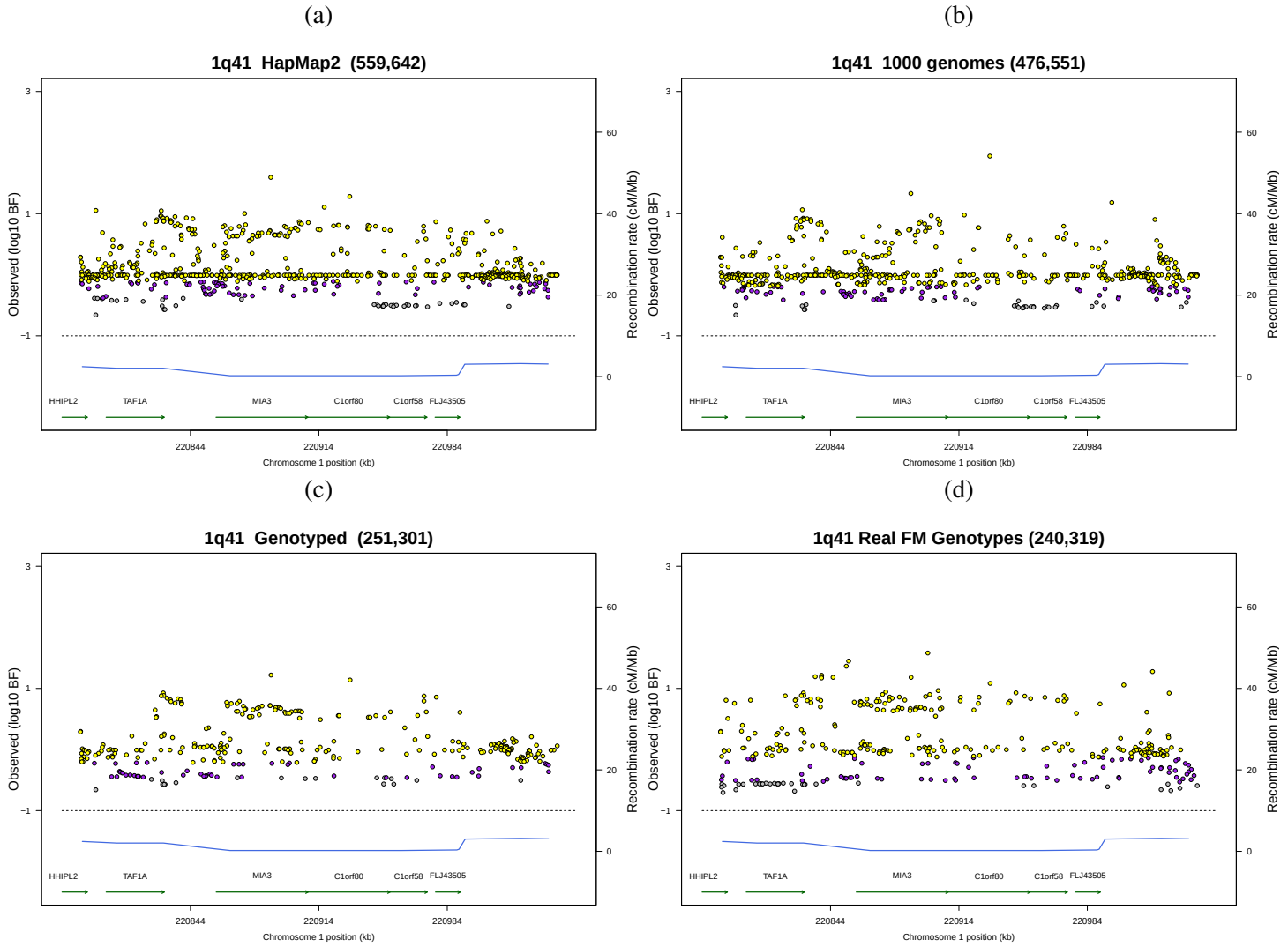


Figure 4.7: Imputed signal for the region 1q41 CAD for each of 3 reference panels, as well as in the FM genotype data. The x-axis corresponds to genomic position, and the y-axis corresponds to log_{10}BF. Recombination is shown in blue on the right y-axis, and genes are shown in green at the bottom. SNPs are coloured by membership in the 95% (yellow) and 99% (purple) credible sets or grey if neither. The number of SNPs in each of the 95 (99) credible sets is shown in the title of each subfigure. Subfigure a-c correspond to: (a) HapMap2 reference panel, (b) 1000 genomes reference panel, (c) genotyped reference panel. Subfigure (d) shows the results for the full fine mapping genotyped data.

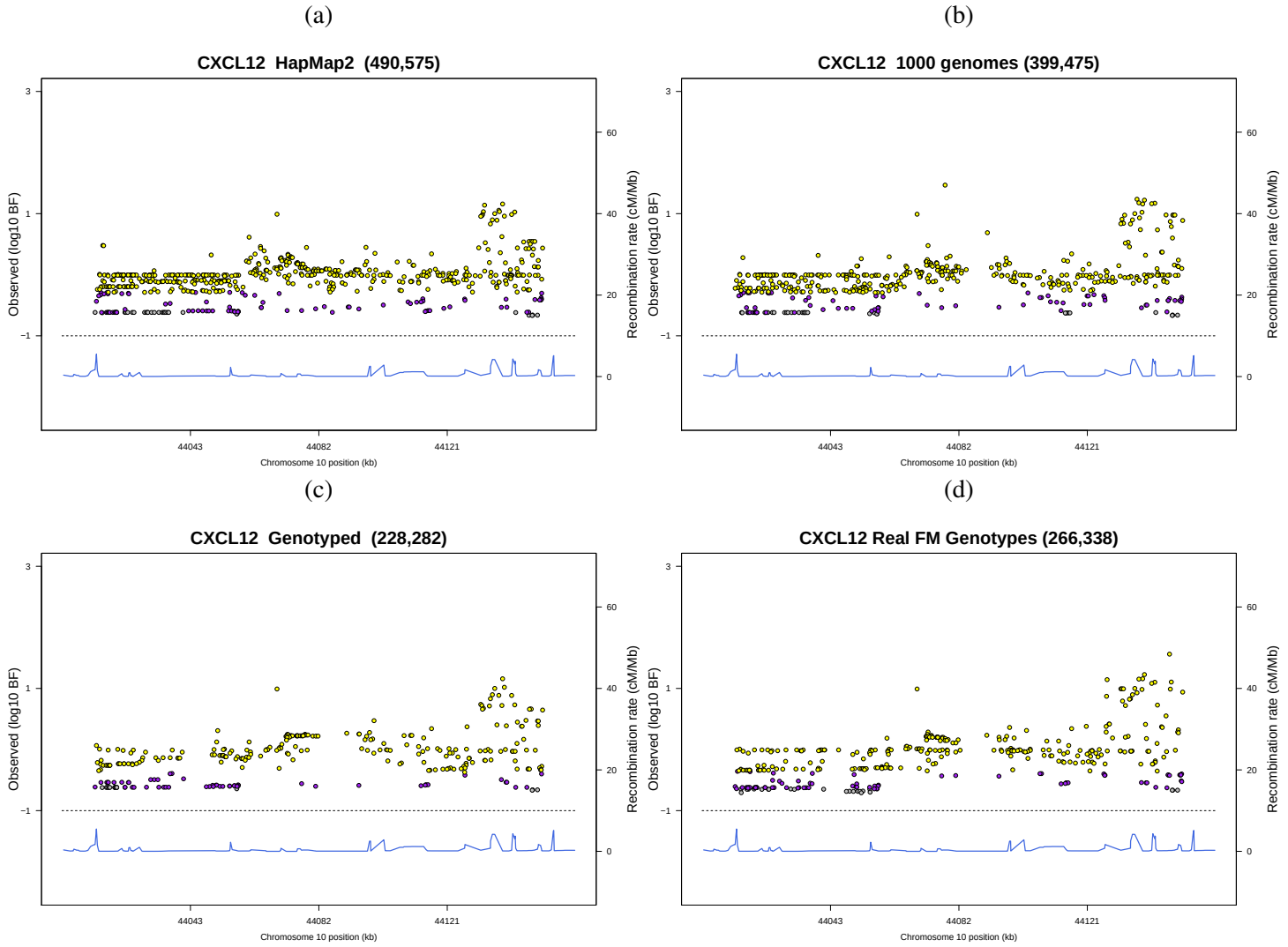


Figure 4.8: Imputed signal for the region *CXCL12* CAD for each of 3 reference panels, as well as in the FM genotype data. The x-axis corresponds to genomic position, and the y-axis corresponds to log_{10}BF. Recombination is shown in blue on the right y-axis, and genes are shown in green at the bottom. SNPs are coloured by membership in the 95% (yellow) and 99% (purple) credible sets or grey if neither. The number of SNPs in each of the 95 (99) credible sets is shown in the title of each subfigure. Subfigure a-c correspond to: (a) HapMap2 reference panel, (b) 1000 genomes reference panel, (c) genotyped reference panel. Subfigure (d) shows the results for the full fine mapping genotyped data.

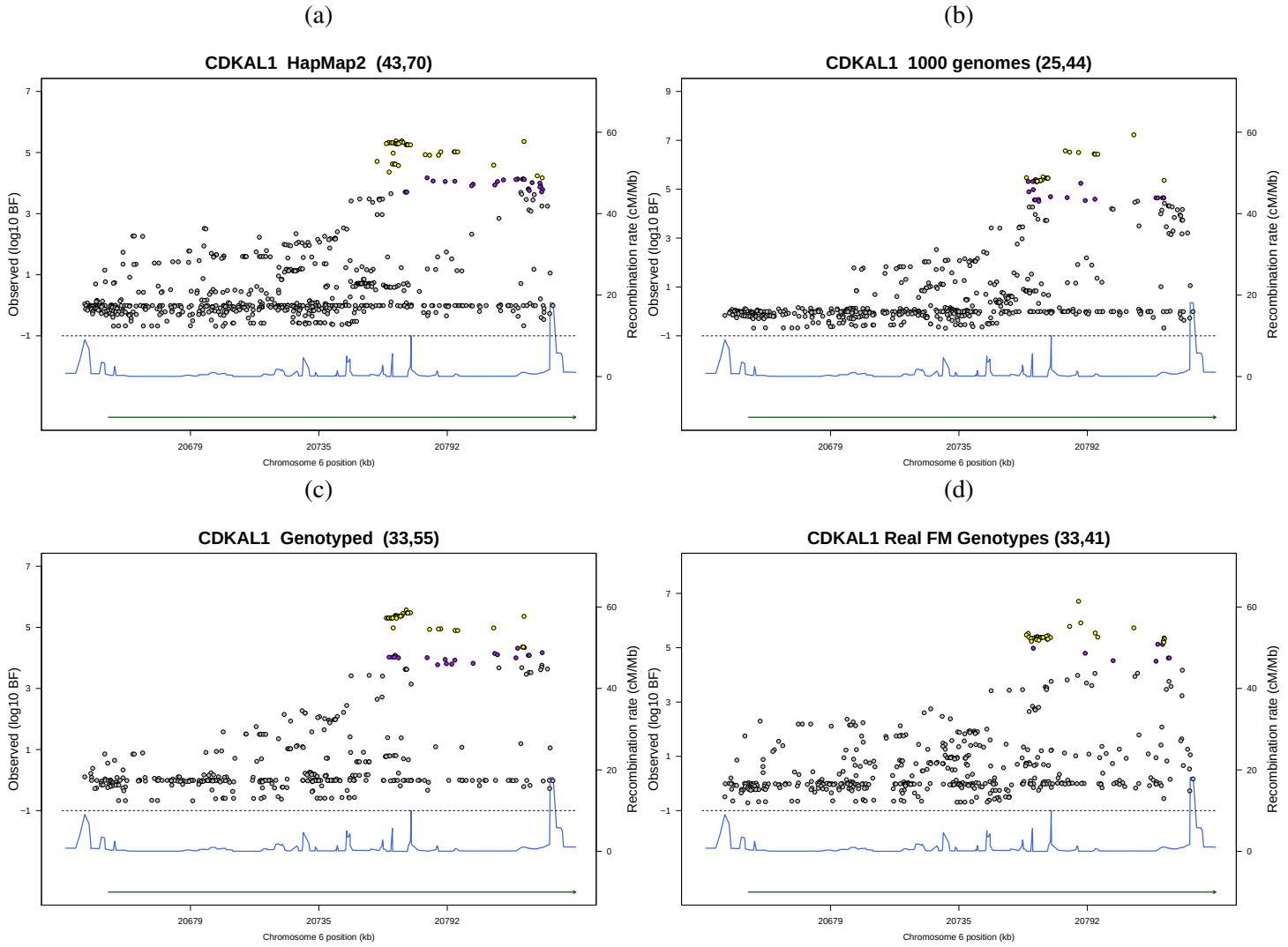


Figure 4.9: Imputed signal for the region *CDKAL1* T2D for each of 3 reference panels, as well as in the FM genotype data. The x-axis corresponds to genomic position, and the y-axis corresponds to log_{10}BF. Recombination is shown in blue on the right y-axis, and genes are shown in green at the bottom. SNPs are coloured by membership in the 95% (yellow) and 99% (purple) credible sets or grey if neither. The number of SNPs in each of the 95 (99) credible sets is shown in the title of each subfigure. Subfigure a-c correspond to: (a) HapMap2 reference panel, (b) 1000 genomes reference panel, (c) genotyped reference panel. Subfigure (d) shows the results for the full fine mapping genotyped data.

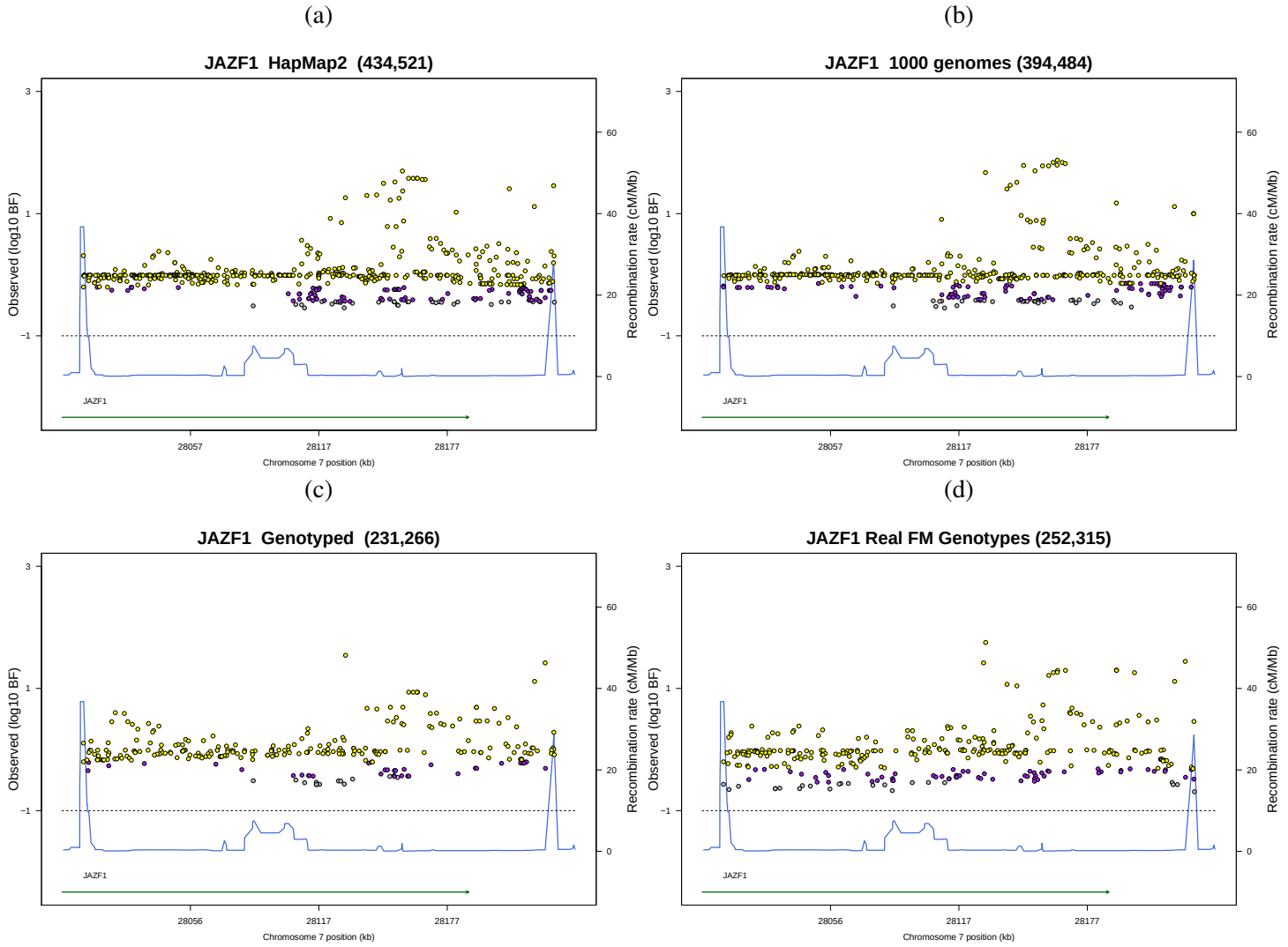


Figure 4.10: Imputed signal for the region *JAZF1* T2D for each of 3 reference panels, as well as in the FM genotype data. The x-axis corresponds to genomic position, and the y-axis corresponds to log_{10}BF. Recombination is shown in blue on the right y-axis, and genes are shown in green at the bottom. SNPs are coloured by membership in the 95% (yellow) and 99% (purple) credible sets or grey if neither. The number of SNPs in each of the 95 (99) credible sets is shown in the title of each subfigure. Subfigure a-c correspond to: (a) HapMap2 reference panel, (b) 1000 genomes reference panel, (c) genotyped reference panel. Subfigure (d) shows the results for the full fine mapping genotyped data.

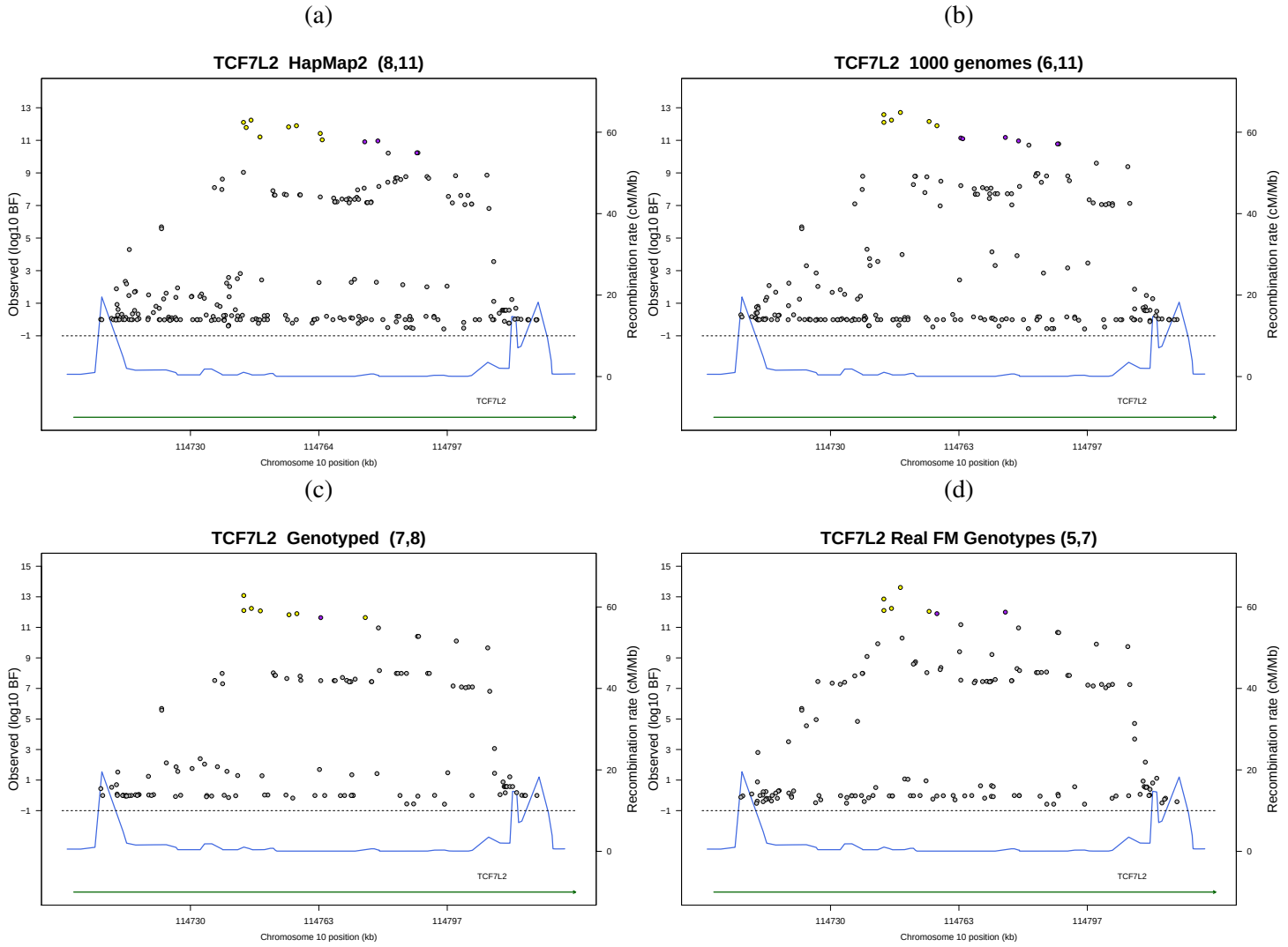


Figure 4.11: Imputed signal for the region *TCF7L2* T2D for each of 3 reference panels, as well as in the FM genotype data. The x-axis corresponds to genomic position, and the y-axis corresponds to log_{10}BF. Recombination is shown in blue on the right y-axis, and genes are shown in green at the bottom. SNPs are coloured by membership in the 95% (yellow) and 99% (purple) credible sets or grey if neither. The number of SNPs in each of the 95 (99) credible sets is shown in the title of each subfigure. Subfigure a-c correspond to: (a) HapMap2 reference panel, (b) 1000 genomes reference panel, (c) genotyped reference panel. Subfigure (d) shows the results for the full fine mapping genotyped data.

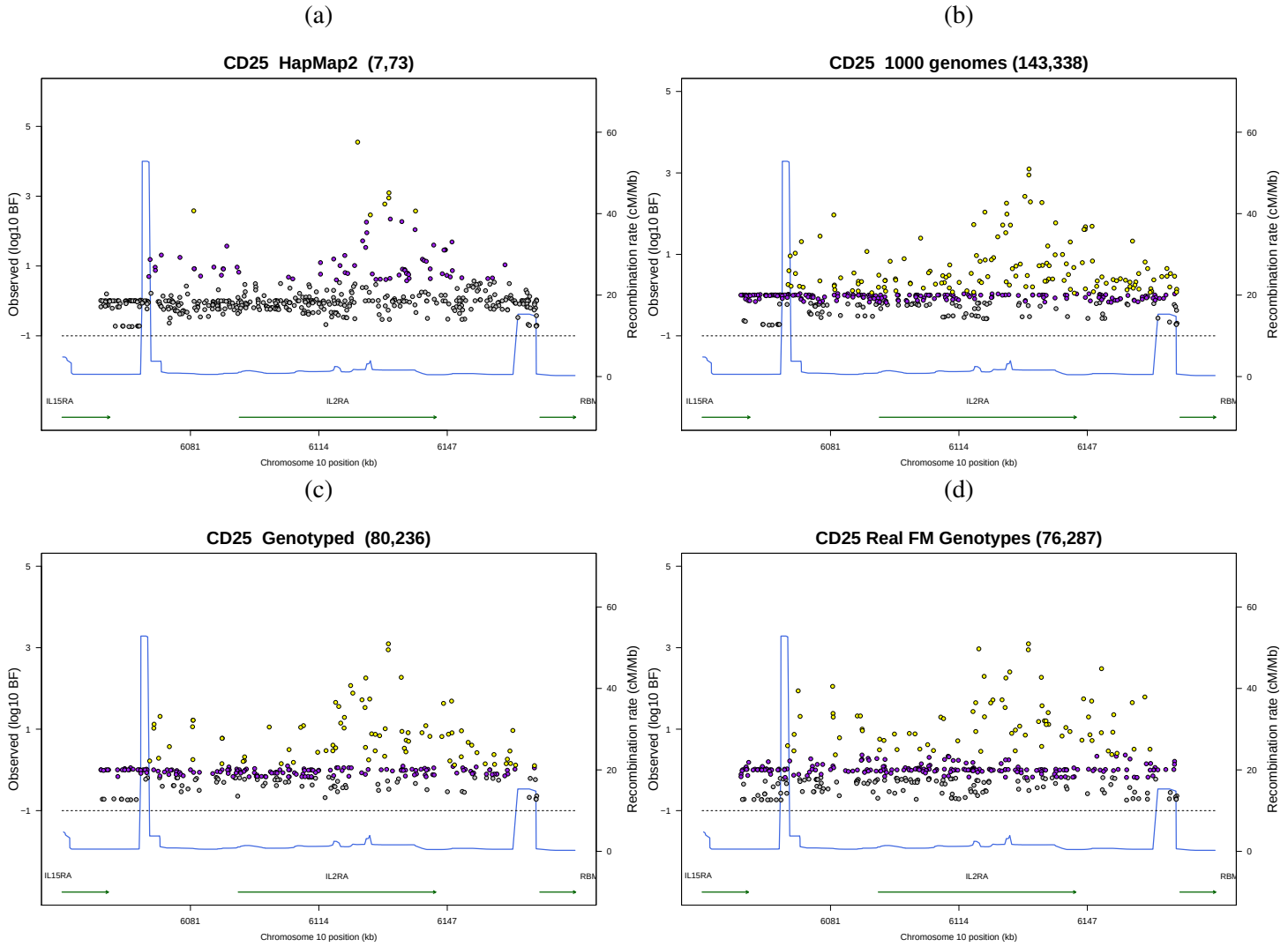


Figure 4.12: Imputed signal for the region *CD25* GD for each of 3 reference panels, as well as in the FM genotype data. The x-axis corresponds to genomic position, and the y-axis corresponds to log_{10}BF. Recombination is shown in blue on the right y-axis, and genes are shown in green at the bottom. SNPs are coloured by membership in the 95% (yellow) and 99% (purple) credible sets or grey if neither. The number of SNPs in each of the 95 (99) credible sets is shown in the title of each subfigure. Subfigure a-c correspond to: (a) HapMap2 reference panel, (b) 1000 genomes reference panel, (c) genotyped reference panel. Subfigure (d) shows the results for the full fine mapping genotyped data.

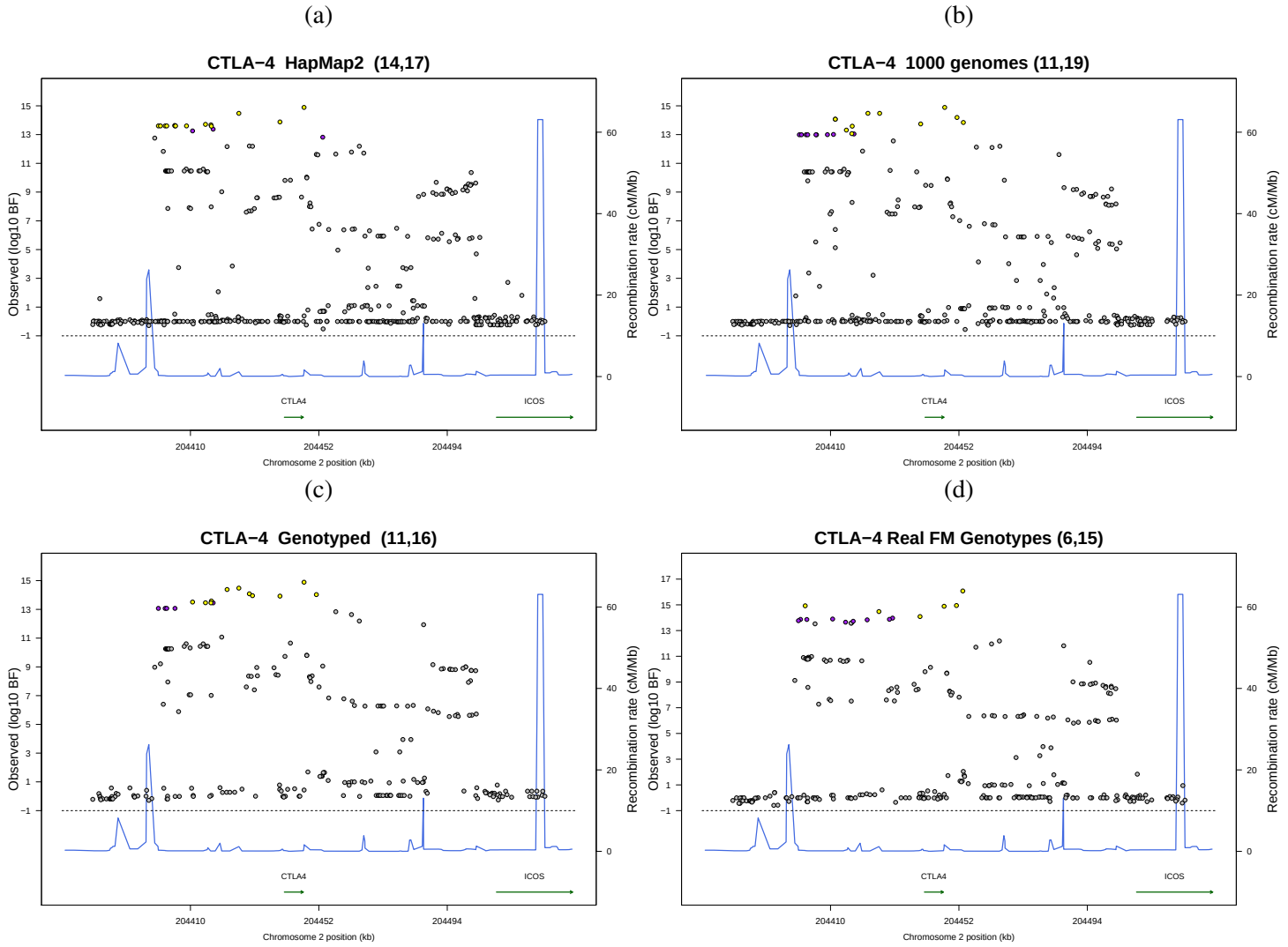


Figure 4.13: Imputed signal for the region *CTLA4* GD for each of 3 reference panels, as well as in the FM genotype data. The x-axis corresponds to genomic position, and the y-axis corresponds to log_{10}BF. Recombination is shown in blue on the right y-axis, and genes are shown in green at the bottom. SNPs are coloured by membership in the 95% (yellow) and 99% (purple) credible sets or grey if neither. The number of SNPs in each of the 95 (99) credible sets is shown in the title of each subfigure. Subfigure a-c correspond to: (a) HapMap2 reference panel, (b) 1000 genomes reference panel, (c) genotyped reference panel. Subfigure (d) shows the results for the full fine mapping genotyped data.

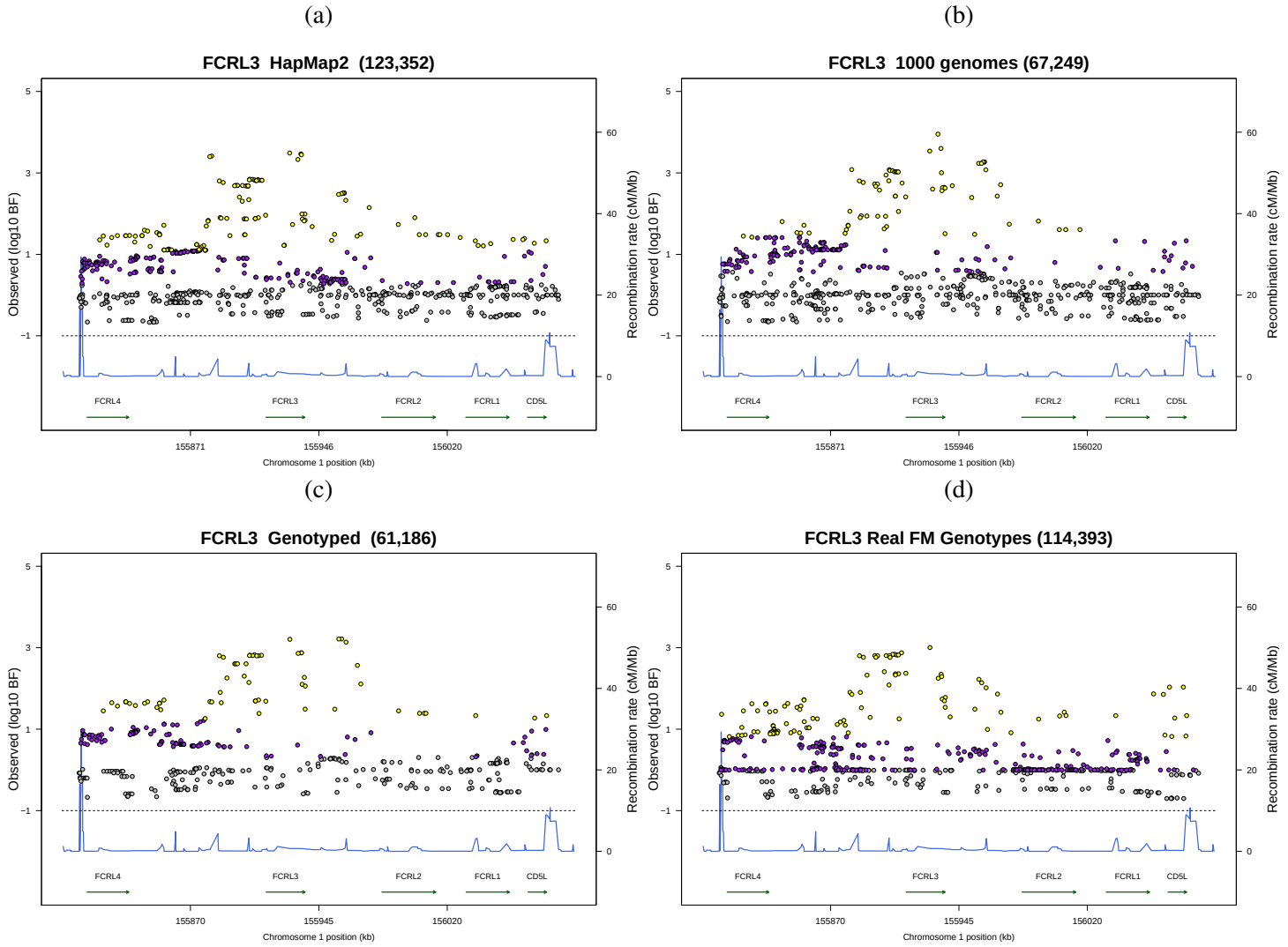


Figure 4.14: Imputed signal for the region *FCRL3* GD for each of 3 reference panels, as well as in the FM genotype data. The x-axis corresponds to genomic position, and the y-axis corresponds to log_{10}BF. Recombination is shown in blue on the right y-axis, and genes are shown in green at the bottom. SNPs are coloured by membership in the 95% (yellow) and 99% (purple) credible sets or grey if neither. The number of SNPs in each of the 95 (99) credible sets is shown in the title of each subfigure. Subfigure a-c correspond to: (a) HapMap2 reference panel, (b) 1000 genomes reference panel, (c) genotyped reference panel. Subfigure (d) shows the results for the full fine mapping genotyped data.

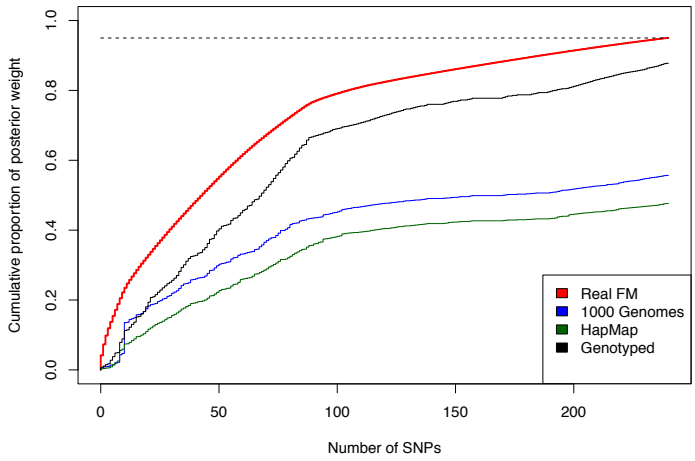
4.3 Step Plots

The distribution of posterior probability across SNPs in a region is of fundamental interest in fine mapping. Metrics such as the size of 95%/99% credible sets provide a general sense of differences between imputed results and results from our real FM and between reference panels. We now look at the distribution of posterior probability across SNPs in each region, and how this differs between reference panels. This will provide substantial insight into how differences between reference panels affect patterns of association among the most interesting SNPs. In order to assess this we plot the cumulative posterior captured in the imputed results, at the actual top SNPs (SNPs in the 95% credible set) from the fine mapping experiment. We compare this distribution between the genotyped and imputed results, as well as between the various reference panel.

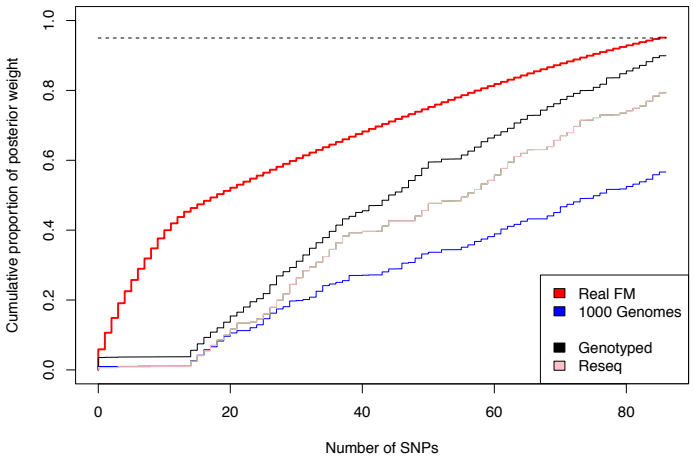
Figure 4.15 through Figure 4.18 show, for each region, the imputed cumulative posterior across reference panels, among SNPs in the 95% credible set as determined by the real fine mapping genotype results.

4.4 Average Posterior Captured

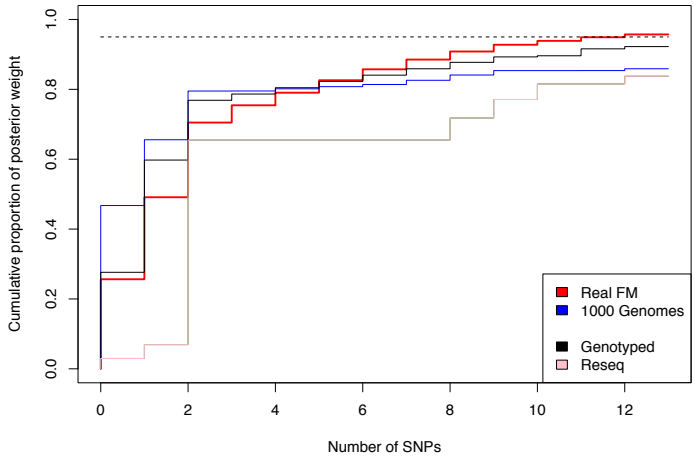
Figure 4.19 shows the proportion of the real posterior captured by the imputed data, across all regions and reference panels. The vertical position corresponds, for each region, to the proportion of the 95% credible set in the real genotypes that is captured at the same set of SNPs in imputed data. In other words, for each region, for each reference panel, we calculated the proportion of imputed posterior captured by the set of SNPs in the 95% credible set from the real genotyped data. In the figure, the Position on the horizontal indicates the reference panel. The



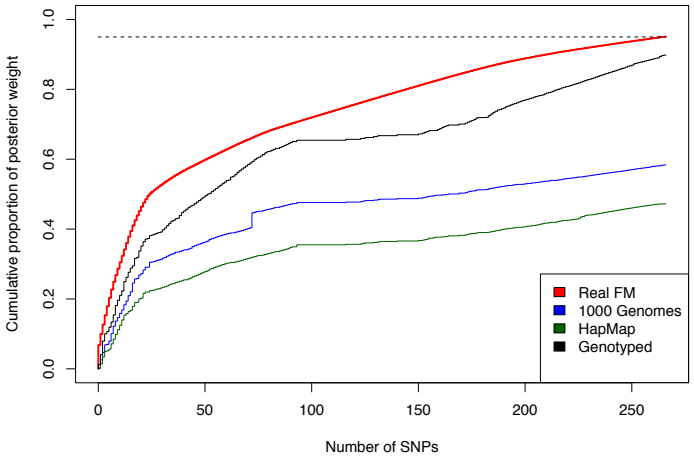
(a) 1*q*41 CAD



(b) 2*q*36 CAD

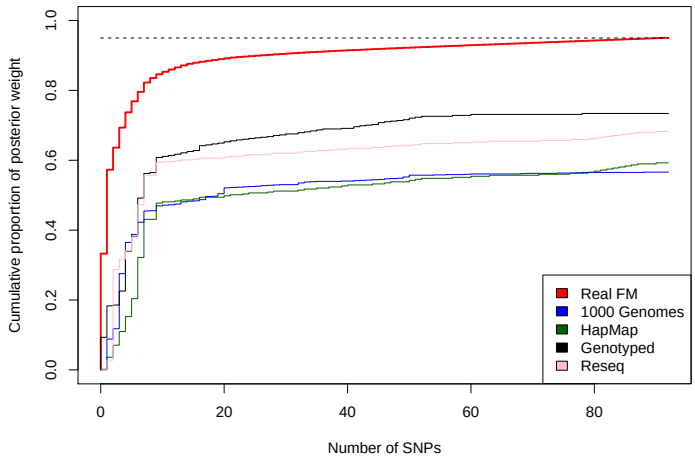


(c) *CDKN2A/B* CAD

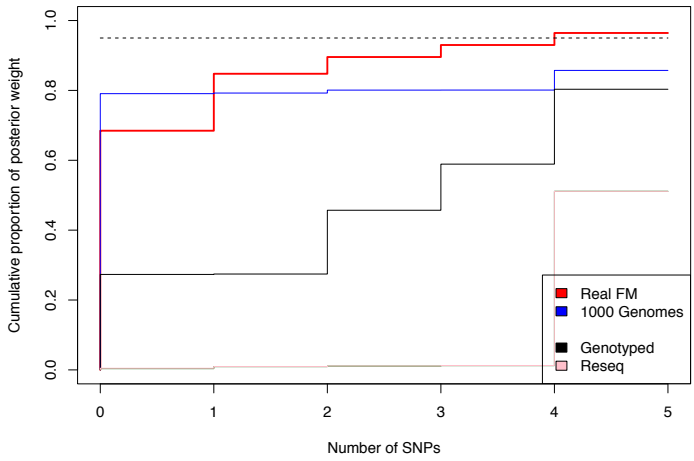


(d) *CXCL12* CAD

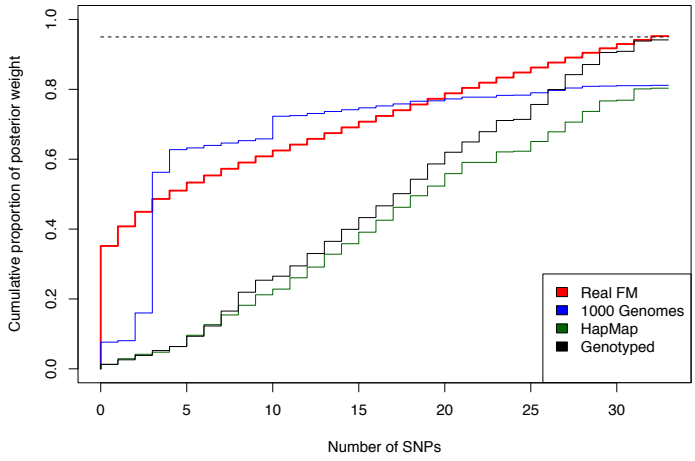
Figure 4.15: Step plots for imputed posterior weight at fine mapping SNPs. In each region, for each SNP (x-axis) in the 95% credible set in the real fine mapping results, the imputed posterior (y-axis) is shown cumulatively. Each region contains a series for each of the reference panels (Genotyped, 1000 genomes, HapMap and where applicable, Resequencing).



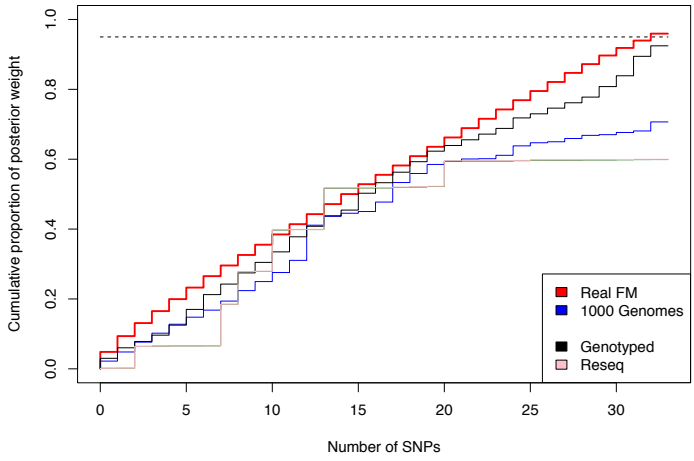
(a) *SORT1* CAD



(b) *CDKN2A/B* T2D

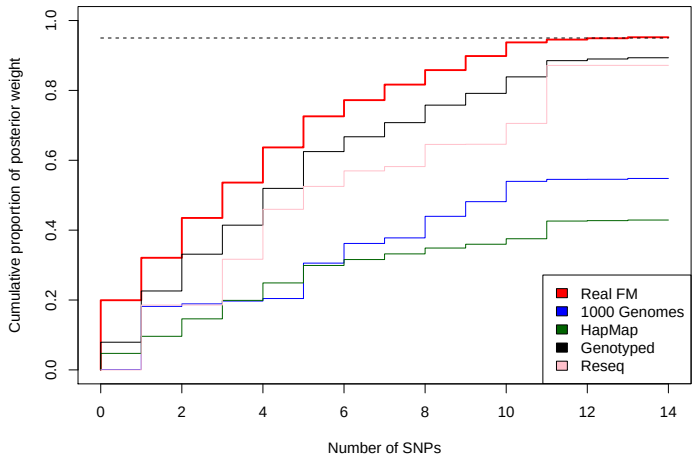


(c) *CDKAL1* T2D

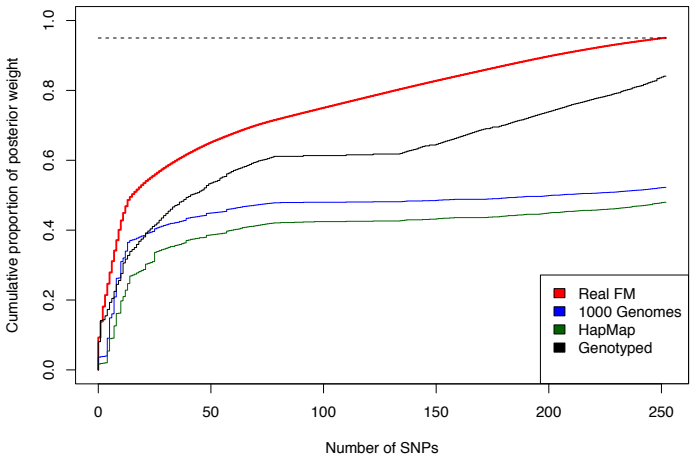


(d) *FTO* T2D

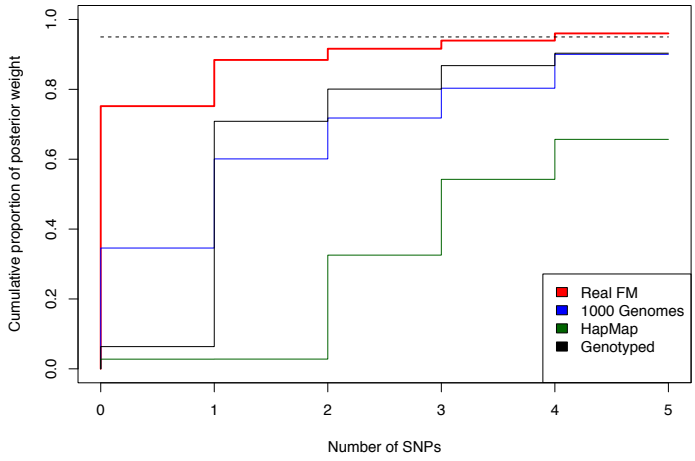
Figure 4.16: Step plots for imputed posterior weight at fine mapping SNPs. In each region, for each SNP (x-axis) in the 95% credible set in the real fine mapping results, the imputed posterior (y-axis) is shown cumulatively. Each region contains a series for each of the reference panels (Genotyped, 1000 genomes, HapMap and where applicable, Resequencing).



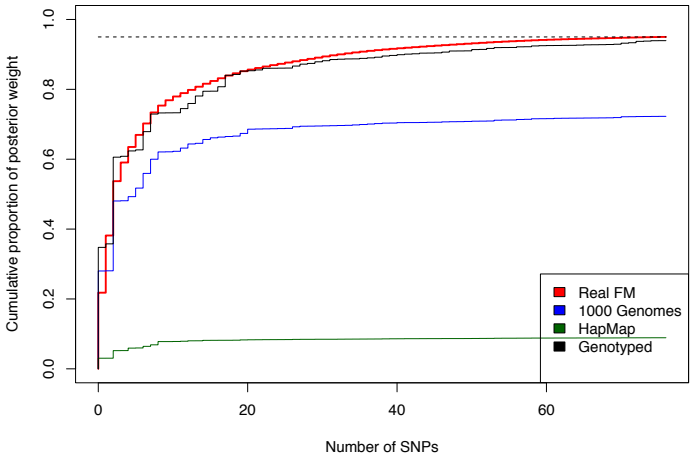
(a) *HHEX* T2D



(b) *JAZF1* T2D



(c) *TCF7L2* T2D



(d) *CD25* GD

Figure 4.17: Step plots for imputed posterior weight at fine mapping SNPs. In each region, for each SNP (x-axis) in the 95% credible set in the real fine mapping results, the imputed posterior (y-axis) is shown cumulatively. Each region contains a series for each of the reference panels (Genotyped, 1000 genomes, HapMap and where applicable, Resequencing).

vertical position indicates the proportion of posterior captured. Therefore, the maximum value is 0.95 (the dashed line). The red diamonds correspond to the average proportion of posterior captured for each of the reference panels. There are 14 regions shown for the HapMap2, genotyped and 1000 genomes reference panels, and five regions for the resequenced reference panel.

Figure 4.20 is the same as Figure 4.19, but with the addition of lines connecting the same regions between reference panels. This illustrates the variability of imputation performance between reference panels for particular regions.

4.5 Evaluating and Improving Fine Mapping Efficiency

The WTCCC fine-mapping and resequencing pilot projects in combination with publicly available data provide a unique opportunity to empirically evaluate strategies for following up GWAS results.

One question of particular interest is whether it is possible to use statistical methods such as imputation to cut down on the number of SNPs that need to be genotyped. For example, a GWAS follow-up experiment involving regional resequencing of 100 samples followed by large scale genotyping in thousands of samples. This is a labor intensive and expensive undertaking, and so improving the efficiency of genotyping for fine mapping is of obvious interest. If it were possible to use the resequenced samples as a reference panel for imputation, it might be possible to confidently exclude the need to genotype certain SNPs. In many of the fine-mapping regions we genotyped, the large majority of SNPs showed essentially no signal. While we don't expect that imputation could very accurately

identify the most associated 1% of SNPs, a more modest task such as marking $1/3$ of SNPs in the region as extremely unlikely to be in the most associated 10% of SNPs, seems plausible and worth investigating. The data from the fine-mapping pilot provides a great resource for empirically comparing approaches in this context, as the 14 regions span the spectrum of allele frequency, effect size, region size, region complexity, and also span multiple disorders. We have undertaken an detailed investigation of imputation for increasing fine-mapping efficiency, looking at several reference panels including regional resequencing, HapMap, 1000 Genomes project data, and fine-mapping genotype data.

4.5.1 Excluding SNPs From Credible Sets

We undertook to evaluate the possibility of excluding SNPs based on GWAS imputation with dense reference panels. Our primary strategy was to filter the imputed data based on various posterior thresholds, and then look at how accurately SNPs were excluded. After imputing the data, we used SNPTEST (Marchini et al., 2007) to generate $\log_{10}$ Bayes Factors (as described previously), and then, for each region, calculated the proportion of posterior assigned to each SNP. We then tried excluding SNPs at various thresholds. For each threshold, we then calculated the proportion of these excluded SNPs present in the 95% credible set from the real fine mapping genotypes. The closer this proportion gets to zero, the better the exclusion method is performing.

One challenge in this evaluation is in interpreting the results across regions. As discussed in the fine mapping chapter, we had limited ability to observe clear signal in some of the regions. For regions where we have sufficient samples and

the effect size is relatively large (eg. *FTO*, *TCF7L2*; See Chapter 2 for a detailed discussion), the 95% credible sets are fairly small ($<$ 10% of SNPs in the region). In these regions, the SNPs in the 95% credible set will generally be strongly associated and thus clearly differentiated from the non-associated SNPs. However, in regions where the signal is weak and/or there is substantial LD, the 95% credible set will include some or many SNPs with a very small proportion of posterior.

Tables 4.1, 4.2 and 4.3 show the results of this analysis. Table 4.1 shows the results for excluding in the 1000 genomes imputed data, Table 4.2 for the genotype reference panel and Table 4.3 for the HapMap2 reference panel.We did not include the locally resequenced reference panel in this evaluation due to the limited number of regions available.

The results of this analysis for the genotyped reference panel are striking. We considered imputation with this reference as our "best case" scenario, and in this evaluation, the performance appears to be quite impressive. The regions are divided almost exactly along the lines of our ability to observe/refine the signal (along the lines of the discussion in Chapters 2 and 3). For the regions where we clearly observe the association signal, in seven out of eight regions we exclude at least 55% of imputed SNPs at a posterior weight threshold of 0.005, with only a single mistake in *CTLA4* (i.e. excluded SNP in the 95% credible set in the real FM results). These seven regions are *CDKN2A/B* CAD/T2D, *FTO, HHEX, CDKAL1, TCF7L2, CTLA4*. This corresponds to a total of 1877 SNPs excluded for these regions, out of a total of 2140, which corresponds to 88% of imputed SNPs excluded. For the 1000 genomes reference panel, the numbers are nearly as encouraging. For the same posterior weight threshold, 1616 SNPs are excluded out of 1805, corresponding to 90% of imputed SNPs excluded. The error rate

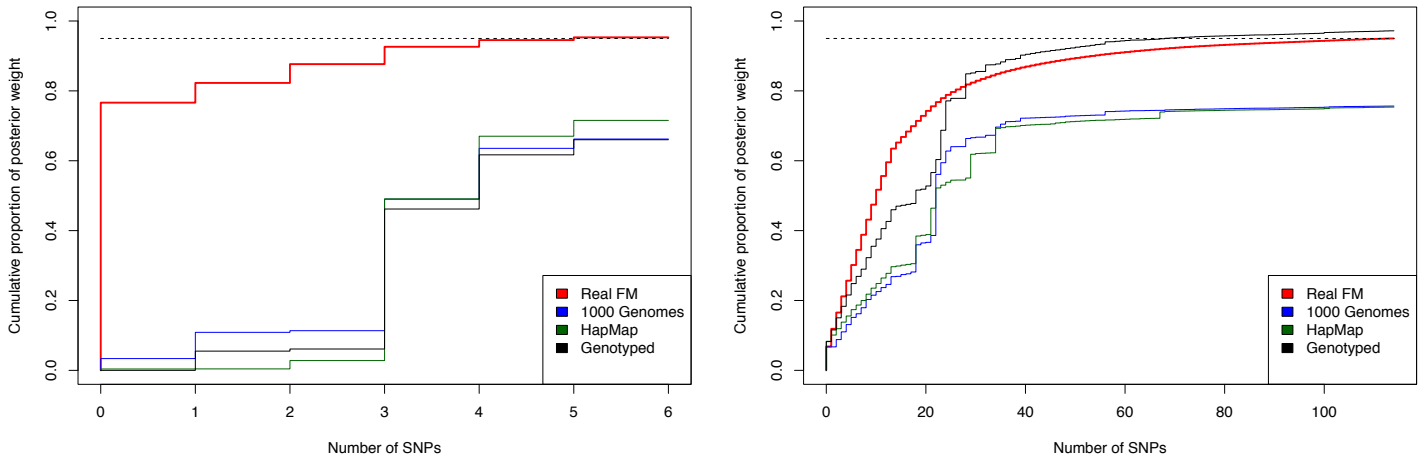
is very slightly higher, with seven of the excluded SNPs appearing in the 95% credible sets in the real data. This result is arguably much more encouraging than that obtained via the genotyped reference panel, as the 1000 genomes imputation more accurately reflects how this approach would be applied in practice.

4.6 Discussion

The results for the genotyped reference panel imputation suggest that using imputation to fine map associated regions will present considerable challenges, if it is at all possible. From the signal plots and particularly from the step plots, it becomes clear that there is considerable variability in terms of accuracy and coverage among the most strongly associated SNPs. The average posterior captured across regions is relatively low, even for the best-case genotyped reference panel. For the other reference panels the results are even less encouraging, both in terms of coverage and accuracy for potentially causal SNPs. While some regions do appear substantially better imputed than others, an investigator attempting to use imputed data for fine mapping has no good a-priori way of predicting the performance for a particular region. This may improve for denser imputation targets as genotyping density increases, and as resequencing data is rapidly improving in accuracy while the cost is also rapidly dropping.

The attempt to improve efficiency via SNP exclusion seems to hold more promise, for both the genotyped reference panel as well as the 1000 genomes and HapMap references. Our results suggest that fine mapping genotyping experiments could realistically substantially limit the number of SNPs per region. Among the benefits of such an approach would be reduced cost for the experiment,

an order-of-magnitude increase in the number of regions that could be surveyed as well as the possibility of targeting the SNPs identified with multiple proxies that might not have otherwise been included for cost or other reasons. It would also be possible to substantially increase the number of samples involved in a fine mapping study. Given the results of our fine mapping efforts (described in previous chapters), this possibility is perhaps the most exciting.



(a) *CTLA4* GD

(b) *FCRL3* GD

Figure 4.18: Step plots for imputed posterior weight at fine mapping SNPs. In each region, for each SNP (x-axis) in the 95% credible set in the real fine mapping results, the imputed posterior (y-axis) is shown cumulatively. Each region contains a series for each of the reference panels (Genotyped, 1000 genomes, HapMap and where applicable, Resequencing).

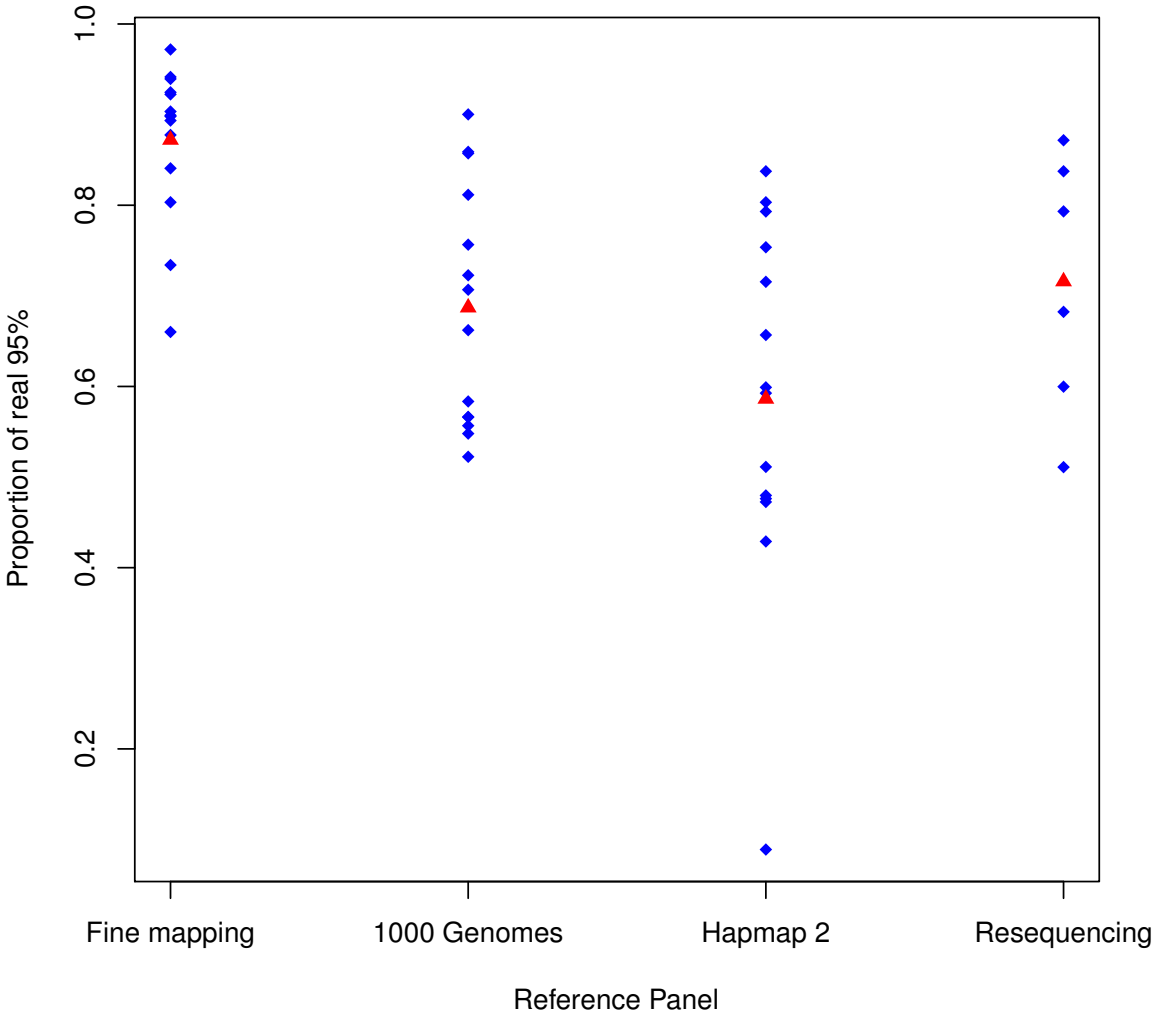


Figure 4.19: For each region and reference panel, the proportion of imputed posterior assigned to SNPs in the 95 % credible set in the real FM genotype data, shown on the left vertical axis. Each blue diamond corresponds to a particular region imputed from a particular reference panel. Reference panel is indicated by the horizontal axis. For each reference panel, the average value across regions is indicated by the red triangle.

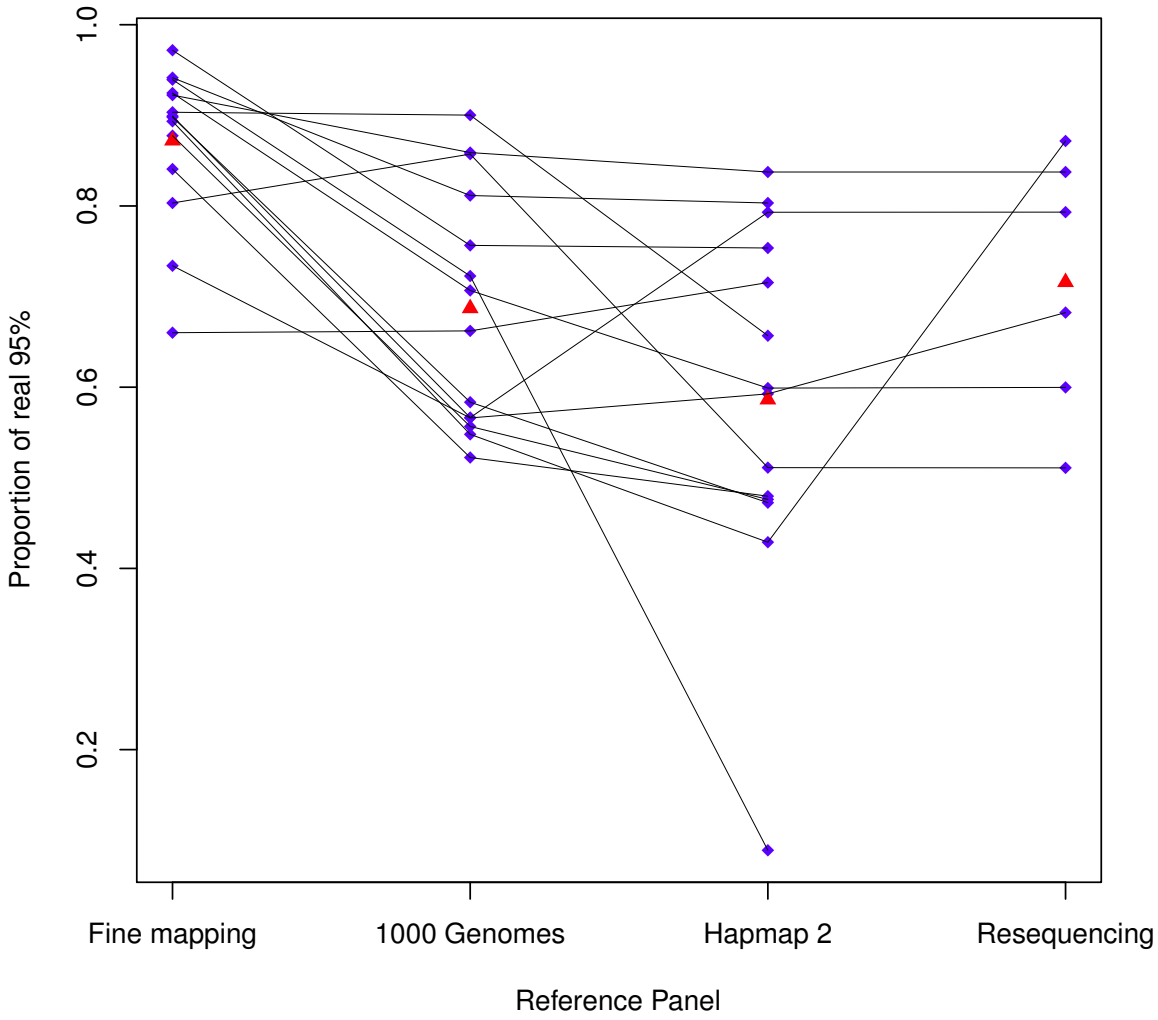


Figure 4.20: For each region and reference panel, the proportion of imputed posterior assigned to SNPs in the 95 % credible set in the real FM genotype data, shown on the left vertical axis. Each blue diamond corresponds to a particular region imputed from a particular reference panel. Reference panel is indicated by the horizontal axis. For each reference panel, the average value across regions is indicated by the red triangle. Black lines link individual regions across reference panels.

Region	Exclusion threshold	SNP count	Imputed SNPs	Excluded SNPs	Wrongly excluded	Proportion excluded	Prop. wrongly excluded
SORT1	0.001	230	162	0	0	0	0
	0.005	230	162	26	0	0.11	0
	$1/nSNPs$	230	162	148	41	0.64	0.28
CDKN2A/B	0.001	518	279	246	0	0.48	0
	0.005	518	279	263	1	0.51	0.01
	$1/nSNPs$	518	279	271	2	0.52	0.01
CDKN2A/B	0.001	514	279	252	2	0.49	0.01
	0.005	514	279	257	3	0.5	0.01
	$1/nSNPs$	514	279	259	3	0.50	0.01
CXCL12	0.001	362	282	0	0	0	0
	0.005	362	282	44	1	0.12	0.02
	$1/nSNPs$	362	282	252	157	0.70	0.62
$1q41$	0.001	353	305	0	0	0	0
	0.005	353	305	39	5	0.11	0.13
	$1/nSNPs$	353	305	253	146	0.72	0.58
$2q36$	0.001	342	187	103	15	0.30	0.15
	0.005	342	187	108	18	0.32	0.17
	$1/nSNPs$	342	187	119	24	0.35	0.20
FTO	0.001	206	134	80	0	0.39	0
	0.005	206	134	82	0	0.40	0
	$1/nSNPs$	206	134	102	4	0.50	0.04
HHEX	0.001	559	380	319	0	0.57	0
	0.005	559	380	355	1	0.64	0
	$1/nSNPs$	559	380	363	2	0.65	0.01
CDKAL1	0.001	496	352	297	2	0.60	0.01
	0.005	496	352	309	2	0.62	0.01
	$1/nSNPs$	496	352	326	8	0.66	0.03
TCF7L2	0.001	156	120	107	0	0.69	0
	0.005	156	120	109	0	0.70	0
	$1/nSNPs$	156	120	112	0	0.72	0
JAZF1	0.001	338	235	0	0	0	0
	0.005	338	235	55	26	0.16	0.47
	$1/nSNPs$	338	235	216	133	0.64	0.62
CTLA4	0.001	292	261	240	0	0.82	0
	0.005	292	261	241	0	0.83	0
	$1/nSNPs$	292	261	245	0	0.84	0
CD25	0.001	425	282	32	0	0.08	0
	0.005	425	282	222	35	0.52	0.16
	$1/nSNPs$	425	282	267	60	0.63	0.23
FCRL3	0.001	606	377	229	12	0.38	0.05
	0.005	606	377	322	61	0.53	0.19
	$1/nSNPs$	606	377	344	81	0.57	0.24

Table 4.1: Exclusion from 95% causal set (1000 Genomes). For each region, the total number of genotyped SNPs is shown, followed by the number of QC-passing imputed SNPs. Posterior weights are generated for the imputed SNPs, which are then filtered based on the specified exclusion threshold. The Excluded SNPs column shows how many SNPs have a posterior weight below the threshold (and thus excluded). Wrongly excluded shows the number of these excluded SNPs which appear in the 95% credible set in the real FM results. Proportion excluded is the proportion of Imputed SNPs that are excluded. Proportion wrongly excluded is the proportion of excluded SNPs which are present in the real 95% credible set.

Region	Exclusion threshold	SNP count	Imputed SNPs	Excluded SNPs	Wrongly excluded	Proportion excluded	Prop. wrongly excluded
SORT1	0.001	230	174	0	0	0	0
	0.005	230	174	0	0	0	0
	$1/nSNPs$	230	174	162	40	0.70	0.25
CDKN2A/B	0.001	518	373	259	0	0.5	0
	0.005	518	373	333	0	0.64	0
	$1/nSNPs$	518	373	357	1	0.69	0
CDKN2A/B	0.001	514	373	332	0	0.65	0
	0.005	514	373	342	0	0.67	0
	$1/nSNPs$	514	373	350	0	0.68	0
CXCL12	0.001	362	297	0	0	0	0
	0.005	362	297	0	0	0	0
	$1/nSNPs$	362	297	207	114	0.57	0.55
$1q41$	0.001	353	319	0	0	0	0
	0.005	353	319	6	0	0.02	0
	$1/nSNPs$	353	319	239	126	0.68	0.53
$2q36$	0.001	342	264	168	8	0.49	0.05
	0.005	342	264	179	18	0.52	0.10
	$1/nSNPs$	342	264	187	20	0.55	0.11
FTO	0.001	206	171	114	0	0.55	0
	0.005	206	171	114	0	0.55	0
	$1/nSNPs$	206	171	135	0	0.66	0
HHEX	0.001	559	456	412	0	0.74	0
	0.005	559	456	417	0	0.75	0
	$1/nSNPs$	559	456	420	0	0.75	0
CDKAL1	0.001	496	399	327	0	0.66	0
	0.005	496	399	334	0	0.67	0
	$1/nSNPs$	496	399	361	0	0.73	0
TCF7L2	0.001	156	130	117	0	0.75	0
	0.005	156	130	118	0	0.76	0
	$1/nSNPs$	156	130	122	0	0.78	0
JAZF1	0.001	338	280	0	0	0	0
	0.005	338	280	0	0	0	0
	$1/nSNPs$	338	280	220	140	0.65	0.64
CTLA4	0.001	292	238	218	1	0.75	0.01
	0.005	292	238	219	1	0.75	0.01
	$1/nSNPs$	292	238	221	1	0.76	0.01
CD25	0.001	425	315	25	0	0.06	0
	0.005	425	315	251	22	0.59	0.09
	$1/nSNPs$	425	315	285	47	0.67	0.17
FCRL3	0.001	606	410	220	5	0.36	0.02
	0.005	606	410	338	43	0.56	0.13
	$1/nSNPs$	606	410	362	66	0.60	0.18

Table 4.2: Exclusion from 95% causal set (Genotyped Reference Panel). For each region, the total number of genotyped SNPs is shown, followed by the number of QC-passing imputed SNPs. Posterior weights are generated for the imputed SNPs, which are then filtered based on the specified exclusion threshold. The Excluded SNPs column shows how many SNPs have a posterior weight below the threshold (and thus excluded). Wrongly excluded shows the number of these excluded SNPs which appear in the 95% credible set in the real FM results. Proportion excluded is the proportion of Imputed SNPs that are excluded. Proportion wrongly excluded is the proportion of excluded SNPs which are present in the real 95% credible set.

Region	Exclusion threshold	SNP count	Imputed SNPs	Excluded SNPs	Wrongly excluded	Proportion excluded	Prop. wrongly excluded
SORT1	0.001	230	126	0	0	0	0
	0.005	230	126	0	0	0	0
	$1/nSNPs$	230	126	115	36	0.5	0.31
CDKN2A/B	0.001	518	296	39	0	0.08	0
	0.005	518	296	261	0	0.50	0
	$1/nSNPs$	518	296	274	2	0.53	0.01
CDKN2A/B	0.001	514	295	279	6	0.54	0.02
	0.005	514	295	279	6	0.54	0.02
	$1/nSNPs$	514	295	279	6	0.54	0.02
CXCL12	0.001	362	281	0	0	0	0
	0.005	362	281	45	1	0.12	0.02
	$1/nSNPs$	362	281	254	158	0.70	0.62
$1q41$	0.001	353	305	0	0	0	0
	0.005	353	305	48	5	0.14	0.10
	$1/nSNPs$	353	305	235	129	0.67	0.55
$2q36$	0.001	342	246	164	13	0.48	0.08
	0.005	342	246	175	20	0.51	0.11
	$1/nSNPs$	342	246	183	22	0.54	0.12
FTO	0.001	206	128	93	5	0.45	0.05
	0.005	206	128	105	12	0.51	0.11
	$1/nSNPs$	206	128	122	23	0.59	0.19
HHEX	0.001	559	383	341	0	0.61	0
	0.005	559	383	356	0	0.64	0
	$1/nSNPs$	559	383	369	2	0.66	0.01
CDKAL1	0.001	496	348	280	1	0.57	0
	0.005	496	348	291	1	0.59	0
	$1/nSNPs$	496	348	304	1	0.61	0
TCF7L2	0.001	156	119	104	0	0.67	0
	0.005	156	119	109	1	0.70	0.01
	$1/nSNPs$	156	119	111	1	0.71	0.01
JAZF1	0.001	338	238	0	0	0	0
	0.005	338	238	35	9	0.10	0.26
	$1/nSNPs$	338	238	219	135	0.65	0.62
CTLA4	0.001	292	267	244	0	0.84	0
	0.005	292	267	248	1	0.85	0
	$1/nSNPs$	292	267	253	1	0.87	0
CD25	0.001	425	284	240	35	0.57	0.15
	0.005	425	284	273	65	0.64	0.24
	$1/nSNPs$	425	284	278	70	0.65	0.25
FCRL3	0.001	606	377	205	6	0.34	0.03
	0.005	606	377	311	48	0.51	0.15
	$1/nSNPs$	606	377	340	75	0.56	0.22

Table 4.3: Exclusion from 95% causal set (HapMap2). For each region, the total number of genotyped SNPs is shown, followed by the number of QC-passing imputed SNPs. Posterior weights are generated for the imputed SNPs, which are then filtered based on the specified exclusion threshold. The Excluded SNPs column shows how many SNPs have a posterior weight below the threshold (and thus excluded). Wrongly excluded shows the number of these excluded SNPs which appear in the 95% credible set in the real FM results. Proportion excluded is the proportion of Imputed SNPs that are excluded. Proportion wrongly excluded is the proportion of excluded SNPs which are present in the real 95% credible set.

Chapter 5

Discussion and next steps

5.1 Fine Mapping Results

Our fine mapping of GWAS loci that I have described in this thesis suggest that follow-up of GWAS hits will typically not be straightforward, but that, at least in some cases, genetic approaches can help to refine the signal from the original GWAS, prior to functional studies. The Bayesian statistical approaches for assigning posterior probabilities and credible sets of causal SNPs were efficient in interpreting the data for these regions. In addition to the detailed analyses of the particular fine mapping regions described in Chapters 2 and 3, there are several general conclusions to be made from the overall fine mapping experiment. Although it is obvious that the sample size for such an experiment should depend on the true effect size, the experience in this study is that in all cases where the original GWAS signal was relatively weak, fine mapping failed to add greatly to the resolution of the signal for experiments of our size (2000 cases and in effect 4000 or 6000 controls). Had the modest GWAS signal in these regions resulted from

poor tagging of a variant with a larger effect size, this fine mapping experiment would have been powered to find the larger-effect variant (if it were genotyped or tagged well), but this did not occur in any region. Given this I conclude that for FM experiments at GWAS loci for complex disease with typical effect sizes (say RR of 1.2 and below) much larger experiments than the ones described here would be prudent. In half the investigated regions, fine mapping eliminated the vast majority of SNPs as potential causal variants underlying the association study. In these regions, it typically remained the case that there was a collection of SNPs which might still be causal. In virtually all of these cases, the top SNPs after fine mapping were different from those from the GWAS experiment, so that the fine mapping experiment added substantial new information. In only two cases, corresponding to the two strongest association signals in the original study did the post-fine mapping evidence concentrate on a single SNP. It is striking that amongst the 109 SNPs in the 95% credible regions of the seven regions where fine mapping eliminated many of the SNPs, none had obviously compelling conventional functional annotations. This adds to evidence that the mechanisms underlying GWAS signals may often be different from those associated with standard annotations in genome browsers. In their current status, annotations as to histone modifications or chromatin accessibility did not seem to be particularly helpful in separating interesting SNPs, but this position could change as these annotations are developed and refined. It is also important to highlight that functional consequence of specific variation is still not characterized on a very fine level. The functional annotation is often based on computer-based algorithms, which is a great first step towards a better understanding of the functional impact of variation but cannot replace actual investigations. It is possible, though not very likely, that we miss or

under appreciate functional consequence of the queried variants due to this.

5.2 Next steps in fine mapping

5.2.1 Custom designed genotyping arrays

Metabochip, Immunochip and similar custom designed genotyping arrays present a cost-effective strategy to follow-up and fine-map large-scale association studies (Voight et al., 2012) (Cortes and Brown, 2011). On Metabochip, 122,241 SNPs to fine-map 257 loci which showed genome-wide significant evidence for association with one or more of the 23 traits were selected from the catalogues of the International HapMap Project and the August 2009 release of the 1000 Genomes Project and contributed results for the design of the chip. Immunochip contains 196,524 variants across 186 known autoimmunity risk loci (Cortes and Brown, 2011) (Liu et al., 2012). While success using these custom arrays for fine mapping is dependent on the presence of genetic variants, especially the causal ones, on the array. One weakness is that many rare variants have yet to be identified, they might be population specific, and are thus not represented on these chips. Genotyping rare variants is a non trivial process and a proportion of the rarer variants will inevitably not be accurately genotyped by the chip and fail QC and thus not be queried for fine mapping purpose and cause false negative results. Despite these cautions these arrays have started to yield interesting results and first steps towards identifying causal variants (Liu et al., 2012).

5.2.2 1000 Genomes imputation efforts in large consortia

With the nearing completion of the 1000Genomes sequencing, the catalogue of variants (both common and low frequency) is substantially expanded. Using the 1000G reference panels will allow the imputation of these variants into large sample sizes for various traits and first stage fine mapping, often paired with conditional analysis, is starting to emerge from these efforts. Also, methods for conditional analysis that can use summary-level statistics from a meta-analysis of GWAS in combination with estimated LD data from individual-level genotype data in an independent reference sample make these types of analysis significantly easier (Yang et al., 2012). While a useful tool in many cases (Liu et al., 2010), imputation has limitations in regions of low LD and for variants with lower frequency. While this is an obvious cost effective first step but will not provide any an answer in every associated region.

5.2.3 Trans-ethnic finemapping

As it is difficult to differentiate functional variation when there is high LD across a region there is much attention focusing on trans-ethnic fine mapping as a possible solution. Populations with lower LD are ideal here as the associated variants are likely to be closer to the true causal variant in these populations (Teo et al., 2010). But even here it can be hard to narrow down the number of possible causative markers to a reasonable number. Success is susceptible to regions where risk variants differ in occurrence across populations, where diseases and traits have differences in frequencies between populations or where risk variants contribute different effect sizes in populations (Rosenberg et al., 2010). As a

proof of principal, a recent study of serum protein concentrations European meta-analysis was combined with data from a GWAS of Japanese ancestry (Franceschini et al., 2012). When looking at the evidence of heterogeneity in allelic effects between ethnic groups, through comparison of association BF under a Bayesian partition model, no large differences for these nine loci were found between the groups of European and Japanese ancestry. The authors used our methods to evaluate "credible sets" with the strongest signals of association and therefore the most likely to be causal (or tagging an unobserved causal variant), on the basis of European-ancestry GWAS only and then after inclusion of the Japanese study. They observed that the resolution of fine mapping of potential causal variants was increased by leveraging transethnic differences in the distribution of LD.

5.2.4 Functional fine mapping

Another, parallel and complementary, avenue to try to identify the causal variant(s) is to integrate the genetic variants in the fine mapped region with gene expression data and other disease relevant intermediate phenotypes (non coding RNA, mRNA, methylation levels etc.). This could potentially help prioritize fine-mapping efforts and provide a shortcut to biology. However, this is a non-trivial task as the statistics involved are challenging and the success of this is dependent of having all markers from the fine mapping study typed in the samples that also have the genomic data and intermediate phenotypes. This requires access to the right tissue in large enough sample sizes to perform a well-powered study, which is usually the biggest challenge, in parallel with relatively high costs of these studies (Nica and Dermitzakis, 2008).

5.3 Conclusion

In conclusion, in this thesis I have described my work and relevant literature around fine mapping and imputation. I outline and discuss the possibilities and challenges to identify causal variants for complex traits through fine mapping. The work included, together with emerging data from other research groups, shows that fine mapping is feasible and informative to move robust GWAS signals closer to a causative culprit.

Bibliography

[1000 Genomes Project Consortium, 2010] 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–73.

[Altshuler et al., 2005] Altshuler, D., Brooks, L. D., Chakravarti, A., Collins, F. S., Daly, M. J., Donnelly, P., and Int HapMap, C. (2005). A haplotype map of the human genome. *Nature*, 437(7063):1299–1320.

[Altshuler et al., 2008] Altshuler, D., Daly, M. J., and Lander, E. S. (2008). Genetic mapping in human disease. *Science (New York, N.Y.)*, 322(5903):881–888.

[Altshuler et al., 2000] Altshuler, D., Hirschhorn, J., Klannemark, M., Lindgren, C., Vohl, M., Nemesh, J., Lane, C., Schaffner, S., Bolk, S., Brewer, C., et al. (2000). The common PPARγ Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nature Genetics*, 26(1):76–80.

[Ardlie et al., 2002] Ardlie, K., Kruglyak, L., and Seielstad, M. (2002). Patterns of linkage disequilibrium in the human genome. *Nature Reviews Genetics*, 3(4):299–309.

[Askari et al., 2003] Askari, A., Unzek, S., Popovic, Z., Goldman, C., Forudi, F., Kiedrowski, M., Rovner, A., Ellis, S., Thomas, J., and DiCorleto, P. (2003). Effect of stromal-cell-derived factor 1 on stem-cell homing and tissue regeneration in ischaemic cardiomyopathy. *The Lancet*, 362(9385):697–703.

[Auton and McVean, 2007] Auton, A. and McVean, G. (2007). Recombination rate estimation in the presence of hotspots. *Genome Res*, 17(8):1219–27.

[Balding, 2006] Balding, D. J. (2006). A tutorial on statistical methods for population association studies. *Nat Rev Genet*, 7(10):781–91.

[Barrett and Cardon, 2006] Barrett, J. C. and Cardon, L. R. (2006). Evaluating coverage of genome-wide association studies. *Nat Genet*, 38(6):659–62.

[Barrett et al., 2008] 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 Gossum, A., Zelenika, D., Franchimont, D., Hugot, J. P., de Vos, 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., and Daly, M. J. (2008). Genome-wide association defines more than 30 distinct susceptibility loci for crohn's disease. *Nat Genet*, 40(8):955–62.

[Bernstein et al., 2005] Bernstein, B., Kamal, M., Lindblad-Toh, K., Bekiranov, S., Bailey, D., Huebert, D., McMahon, S., Karlsson, E., Kulbokas III, E., Gingeras, T., et al. (2005). Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell*, 120(2):169–181.

[Bernstein et al., 2006] Bernstein, B., Mikkelsen, T., Xie, X., Kamal, M., Huebert, D., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., et al. (2006). A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell*, 125(2):315–326.

[Bhinge et al., 2007] Bhinge, A., Kim, J., Euskirchen, G., Snyder, M., and Iyer, V. (2007). Mapping the chromosomal targets of STAT1 by sequence tag analysis of genomic enrichment (STAGE). *Genome research*, 17(6):910.

[Boissel et al., 2009] Boissel, S., Reish, O., Proulx, K., Kawagoe-Takaki, H., Sedgwick, B., Yeo, G. S., Meyre, D., Golzio, C., Molinari, F., Kadhom, N., Etchevers, H. C., Saudek, V., Farooqi, I. S., Froguel, P., Lindahl, T., O'Rahilly, S., Munnich, A., and Colleaux, L. (2009). Loss-of-function mutation in the dioxygenase-encoding fto gene causes severe growth retardation and multiple malformations. *Am J Hum Genet*, 85(1):106–11.

[Bort et al., 2006] Bort, R., Signore, M., Tremblay, K., Barbera, J., and Zaret, K. (2006). Hex homeobox gene controls the transition of the endoderm to a pseudostratified, cell emergent epithelium for liver bud development. *Developmental biology*, 290(1):44–56.

[Bosserhoff and Buettner, 2002] Bosserhoff, A. and Buettner, R. (2002). Expression, function and clinical relevance of mia (melanoma inhibitory activity). *Histology and histopathology*, 17(1):289–300.

[Boyle et al., 2008a] Boyle, A., Davis, S., Shulha, H., Meltzer, P., Margulies, E., Weng, Z., Furey, T., and Crawford, G. (2008a). High-resolution mapping and characterization of open chromatin across the genome. *Cell*, 132(2):311–322.

[Boyle et al., 2008b] Boyle, A., Guinney, J., Crawford, G., and Furey, T. (2008b). F-Seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics*, 24(21):2537.

[Brand et al., 2007] Brand, O. J., Lowe, C. E., Heward, J. M., Franklyn, J. A., Cooper, J. D., Todd, J. A., and Gough, S. C. L. (2007). Association of the interleukin-2 receptor alpha (il-2r)/cd25 gene region with graves' disease using a multilocus test and tag snps. *Clinical Endocrinology*, 66(4):508–512.

[Braun et al., 1998] Braun, J., Donner, H., Siegmund, T., Walfish, P., Usadel, K., and Badenhoop, K. (1998). CTLA-4 promoter variants in patients with Graves' disease and Hashimoto's thyroiditis. *Tissue Antigens*, 51(5):563.

[Browning and Browning, 2007] Browning, S. R. and Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet*, 81(5):1084–97.

[Buck et al., 2005] Buck, M., Nobel, A., and Lieb, J. (2005). ChIPOTle: a user-friendly tool for the analysis of ChIP-chip data. *Genome Biology*, 6(11):R97.

[Burton et al., 2007] Burton, P. R., Clayton, D. G., Cardon, L. R., Craddock, N., Deloukas, P., Duncanson, A., Kwiatkowski, D. P., McCarthy, M. I., Ouwehand, W. H., Samani, N. J., Todd, J. A., Donnelly, P., Barrett, J. C., Davison, D., Easton, D., Evans, D. M., Leung, H. T., Marchini, J. L., Morris, A. P., Spencer, C. C., Tobin, M. D., Attwood, A. P., Boorman, J. P., Cant, B., Everson, U., Hussey, J. M., Jolley, J. D., Knight, A. S., Koch, K., Meech, E., Nutland, S., Prowse, C. V., Stevens, H. E., Taylor, N. C., Walters, G. R., Walker, N. M., Watkins, N. A., Winzer, T., Jones, R. W., McArdle, W. L., Ring, S. M., Strachan, D. P., Pembrey, M., Breen, G., St Clair, D., Caesar, S., Gordon-Smith, K., Jones, L., Fraser, C., Green, E. K., Grozeva, D., Hamshere, M. L., Holmans, P. A., Jones, I. R., Kirov, G., Moskivina, V., Nikolov, I., O'Donovan, M. C., Owen, M. J., Collier, D. A., Elkin, A., Farmer, A., Williamson, R., McGuffin, P., Young, A. H., Ferrier, I. N., Ball, S. G., Balmforth, A. J., Barrett, J. H., Bishop, T. D., Iles, M. M., Maqbool, A., Yuldasheva, N., Hall, A. S., Braund, P. S., Dixon, R. J., Mangino, M., Stevens, S., Thompson, J. R., Bredin, F., Tremelling, M., Parkes, M., Drummond, H., Lees, C. W., Nimmo, E. R., Satsangi, J., Fisher, S. A., Forbes, A., Lewis, C. M., Onnie, C. M., Prescott, N. J., Sanderson, J., Matthew, C. G., Barbour, J., Mohiuddin, M. K., Todhunter, C. E., Mansfield, J. C., Ahmad, T., Cummings, F. R., Jewell, D. P., et al. (2007). Association scan of 14,500 nonsynonymous snps in four diseases identifies autoimmunity variants. *Nat Genet*, 39(11):1329–37.

[CADC, 2009] CADC (2009). Large scale association analysis of novel genetic loci for coronary artery disease. *Arterioscler Thromb Vasc Biol*, 29(5):774–780.

[Cann et al., 2002] Cann, H., de Toma, C., Cazes, L., Legrand, M., Morel, V., Piouffre, L., Bodmer, J., Bodmer, W., Bonne-Tamir, B., Cambon-Thomsen, A., et al. (2002). A Human Genome Diversity Cell Line Panel. *Science*, 296(5566):261.

[Cardon and Bell, 2001] Cardon, L. and Bell, J. (2001). Association study designs for complex diseases. *Nature Reviews Genetics*, 2(2):91–99.

[Cardon and Palmer, 2003] Cardon, L. and Palmer, L. (2003). Population stratification and spurious allelic association. *The Lancet*, 361(9357):598–604.

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

[Chiu et al., 2006] Chiu, K., Wong, C., Chen, Q., Ariyaratne, P., Ooi, H., Wei, C., Sung, W., and Ruan, Y. (2006). PET-Tool: a software suite for comprehensive processing and managing of Paired-End diTag (PET) sequence data. *BMC Bioinformatics*, 7:390.

[Cho et al., 2012] Cho, Y. S., Chen, C.-H., Hu, C., Long, J., Ong, R. T. H., Sim, X., Takeuchi, F., Wu, Y., Go, M. J., Yamauchi, T., Chang, Y.-C., Kwak, S. H., Ma, R. C. W., Yamamoto, K., Adair, L. S., Aung, T., Cai, Q., Chang, L.-C., Chen, Y.-T., Gao, Y., Hu, F. B., Kim, H.-L., Kim, S., Kim, Y. J., Lee, J. J.-M., Lee, N. R., Li, Y., Liu, J. J., Lu, W., Nakamura, J., Nakashima, E., Ng, D. P.-K., Tay, W. T., Tsai, F.-J., Wong, T. Y., Yokota, M., Zheng, W., Zhang, R., Wang, C., So, W. Y., Ohnaka, K., Ikegami, H., Hara, K., Cho, Y. M., Cho,

N. H., Chang, T.-J., Bao, Y., Hedman, Å. K., Morris, A. P., McCarthy, M. I., DIAGRAM Consortium, MuTHER Consortium, Takayanagi, R., Park, K. S., Jia, W., Chuang, L.-M., Chan, J. C. N., Maeda, S., Kadowaki, T., Lee, J.-Y., Wu, J.-Y., Teo, Y. Y., Tai, E. S., Shu, X. O., Mohlke, K. L., Kato, N., Han, B.-G., and Seielstad, M. (2012). Meta-analysis of genome-wide association studies identifies eight new loci for type 2 diabetes in east asians. *Nat Genet*, 44(1):67–72.

[Choo et al., 2006] Choo, A., Padmanabhan, J., Chin, A., Fong, W., and Oh, S. (2006). Immortalized feeders for the scale-up of human embryonic stem cells in feeder and feeder-free conditions. *Journal of biotechnology*, 122(1):130–141.

[Collins et al., 1997] Collins, F. S., Guyer, M. S., and Charkravarti, A. (1997). Variations on a theme: cataloging human dna sequence variation. *Science*, 278(5343):1580–1.

[Cortes and Brown, 2011] Cortes, A. and Brown, M. A. (2011). Promise and pitfalls of the immunochip. *Arthritis Res Ther*, 13(1):101.

[Crawford et al., 2006a] Crawford, G., Davis, S., Scacheri, P., Renaud, G., Halawi, M., Erdos, M., Green, R., Meltzer, P., Wolfsberg, T., and Collins, F. (2006a). Dnase-chip: a high-resolution method to identify dnase i hypersensitive sites using tiled microarrays. *Nature methods*, 3(7):503–509.

[Crawford et al., 2006b] Crawford, G. E., Holt, I. E., Whittle, J., Webb, B. D., Tai, D., Davis, S., Margulies, E. H., Chen, Y., Bernat, J. A., Ginsburg, D., Zhou, D., Luo, S., Vasicek, T. J., Daly, M. J., Wolfsberg, T. G., and Collins, F. S.

(2006b). Genome-wide mapping of dnase hypersensitive sites using massively parallel signature sequencing (mpss). *Genome Res*, 16(1):123–31.

[Daly et al., 2001] Daly, M., Rioux, J., Schaffner, S., Hudson, T., and Lander, E. (2001). High-resolution haplotype structure in the human genome. *Nature genetics*, 29(2):229–232.

[Davuluri et al., 2001] Davuluri, R., Grosse, I., and Zhang, M. (2001). Computational identification of promoters and first exons in the human genome. *Nature genetics*, 29(4):412–417.

[Dendrou et al., 2009] Dendrou, C. A., Plagnol, V., Fung, E., Yang, J. H. M., Downes, K., Cooper, J. D., Nutland, S., Coleman, G., Himsworth, M., Hardy, M., Burren, O., Healy, B., Walker, N. M., Koch, K., Ouwehand, W. H., Bradley, J. R., Wareham, N. J., Todd, J. A., and Wicker, L. S. (2009). Cell-specific protein phenotypes for the autoimmune locus il2ra using a genotype-selectable human bioresource. *Nat Genet*, 41(9):1011–1015. 10.1038/ng.434.

[Dennis et al., 2007] Dennis, J., Fan, H., Reynolds, S., Yuan, G., Meldrim, J., Richter, D., Peterson, D., Rando, O., Noble, W., and Kingston, R. (2007). Independent and complementary methods for large-scale structural analysis of mammalian chromatin. *Genome research*, 17(6):928.

[Despriet et al., 2006] Despriet, D. D. G., Klaver, C. C. W., Witteman, J. C. M., Bergen, A. A. B., Kardys, I., de Maat, M. P. M., Boekhoorn, S. S., Vingerling, J. R., Hofman, A., Oostra, B. A., Uitterlinden, A. G., Stijnen, T., van Duijn, C. M., and de Jong, P. T. V. M. (2006). Complement factor h polymorphism,

complement activators, and risk of age-related macular degeneration. *JAMA*, 296(3):301–9.

[Dina et al., 2007] Dina, C., Meyre, D., Gallina, S., Durand, E., Korner, A., Jacobson, P., Carlsson, L. M., Kiess, W., Vatin, V., Lecoeur, C., Delplanque, J., Vaillant, E., Pattou, F., Ruiz, J., Weill, J., Levy-Marchal, C., Horber, F., Potoczna, N., Hercberg, S., Le Stunff, C., Bougneres, P., Kovacs, P., Marre, M., Balkau, B., Cauchi, S., Chevre, J. C., and Froguel, P. (2007). Variation in fto contributes to childhood obesity and severe adult obesity. *Nat Genet*, 39(6):724–6.

[Dixon et al., 2007] Dixon, A., Liang, L., Moffatt, M., Chen, W., Heath, S., Wong, K., Taylor, J., Burnett, E., Gut, I., and Farrall, M. (2007). A genome-wide association study of global gene expression. *Nature genetics*, 39(10):1202–1207.

[Down and Hubbard, 2002] Down, T. and Hubbard, T. (2002). Computational Detection and Location of Transcription Start Sites in Mammalian Genomic DNA. *Genome Research*, 12(3):458.

[Dupuis et al., 2010] Dupuis, J., Langenberg, C., Prokopenko, I., Saxena, R., Soranzo, N., Jackson, A. U., Wheeler, E., Glazer, N. L., Bouatia-Naji, N., Gloyn, A. L., Lindgren, C. M., Magi, R., Morris, A. P., Randall, J., Johnson, T., Elliott, P., Rybin, D., Thorleifsson, G., Steinthorsdottir, V., Henneman, P., Grallert, H., Dehghan, A., Hottenga, J. J., Franklin, C. S., Navarro, P., Song, K., Goel, A., Perry, J. R. B., Egan, J. M., Lajunen, T., Grarup, N., Sparso, T., Doney, A., Voight, B. F., Stringham, H. M., Li, M., Kanoni, S., Shrader, P., Cavalcanti-

Proenca, C., Kumari, M., Qi, L., Timpson, N. J., Gieger, C., Zabena, C., Rocheleau, G., Ingelsson, E., An, P., O'Connell, J., Luan, J., Elliott, A., McCarroll, S. A., Payne, F., Roccasecca, R. M., Pattou, F., Sethupathy, P., Ardlie, K., Ariyurek, Y., Balkau, B., Barter, P., Beilby, J. P., Ben-Shlomo, Y., Benediktsson, R., Bennett, A. J., Bergmann, S., Bochud, M., Boerwinkle, E., Bonnefond, A., Bonnycastle, L. L., Borch-Johnsen, K., Böttcher, Y., Brunner, E., Bumpstead, S. J., Charpentier, G., Chen, Y.-D. I., Chines, P., Clarke, R., Coin, L. J. M., Cooper, M. N., Cornelis, M., Crawford, G., Crisponi, L., Day, I. N. M., de Geus, E. J. C., Delplanque, J., Dina, C., Erdos, M. R., Fedson, A. C., Fischer-Rosinsky, A., Forouhi, N. G., Fox, C. S., Frants, R., Franzosi, M. G., Galan, P., Goodarzi, M. O., Graessler, J., Groves, C. J., Grundy, S., Gwilliam, R., Gyllensten, U., Hadjadj, S., Hallmans, G., Hammond, N., Han, X., Hartikainen, A.-L., Hassanali, N., Hayward, C., Heath, S. C., Hercberg, S., Herder, C., Hicks, A. A., Hillman, D. R., Hingorani, A. D., Hofman, A., Hui, J., Hung, J., Isomaa, B., and Johnson (2010). New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nat Genet*, 42(2):105–16.

[Euskirchen et al., 2004] Euskirchen, G., Royce, T., Bertone, P., Martone, R., Rinn, J., Nelson, F., Sayward, F., Luscombe, N., Miller, P., Gerstein, M., et al. (2004). CREB binds to multiple loci on human chromosome 22. *Molecular and cellular biology*, 24(9):3804.

[Euskirchen et al., 2007] Euskirchen, G., Rozowsky, J., Wei, C., Lee, W., Zhang, Z., Hartman, S., Emanuelsson, O., Stolc, V., Weissman, S., Gerstein, M., et al. (2007). Mapping of transcription factor binding regions in mammalian cells

by ChIP: Comparison of array-and sequencing-based technologies. *Genome Research*, 17(6):898.

[Fakhrai-Rad et al., 2000] Fakhrai-Rad, H., Nikoshkov+, A., Kamel, A., Fernstrom, M., Zierath, J., Norgren, S., Luthman, H., and Galli, J. (2000). Insulin-degrading enzyme identified as a candidate diabetes susceptibility gene in gk rats. 9(14):2149–2158.

[Falconer DS, 1996] Falconer DS, M. T. (1996). *Introduction to Quantitative Genetics*. Addison Wesley Longman, Harlow UK.

[Fan et al., 2006] Fan, J.-B., Chee, M. S., and Gunderson, K. L. (2006). Highly parallel genomic assays. *Nat Rev Genet*, 7(8):632–44.

[Farris et al., 2003] Farris, W., Mansourian, S., Chang, Y., Lindsley, L., Eckman, E., Frosch, M., Eckman, C., Tanzi, R., Selkoe, D., and GuÈnette, S. (2003). Insulin-degrading enzyme regulates the levels of insulin, amyloid -protein, and the -amyloid precursor protein intracellular domain in vivo. *Proceedings of the National Academy of Sciences*, 100(7):4162–4167.

[Fischer et al., 2009] Fischer, J., Koch, L., Emmerling, C., Vierkotten, J., Peters, T., Bruning, J. C., and Ruther, U. (2009). Inactivation of the fto gene protects from obesity. *Nature*, 458(7240):894–8.

[Fodor et al., 1991] Fodor, S. P., Read, J. L., Pirrung, M. C., Stryer, L., Lu, A. T., and Solas, D. (1991). Light-directed, spatially addressable parallel chemical synthesis. *Science*, 251(4995):767–73.

[Franceschini et al., 2012] Franceschini, N., van Rooij, F. J. A., Prins, B. P., Feitosa, M. F., Karakas, M., Eckfeldt, J. H., Folsom, A. R., Kopp, J., Vaez, A., Andrews, J. S., Baumert, J., Boraska, V., Broer, L., Hayward, C., Ngwa, J. S., Okada, Y., Polasek, O., Westra, H.-J., Wang, Y. A., Del Greco M, F., Glazer, N. L., Kapur, K., Kema, I. P., Lopez, L. M., Schillert, A., Smith, A. V., Winkler, C. A., Zgaga, L., The LifeLines Cohort Study, Bandinelli, S., Bergmann, S., Boban, M., Bochud, M., Chen, Y. D., Davies, G., Dehghan, A., Ding, J., Doering, A., Durda, J. P., Ferrucci, L., Franco, O. H., Franke, L., Gunjaca, G., Hofman, A., Hsu, F.-C., Kolcic, I., Kraja, A., Kubo, M., Lackner, K. J., Launer, L., Loehr, L. R., Li, G., Meisinger, C., Nakamura, Y., Schwienbacher, C., Starr, J. M., Takahashi, A., Torlak, V., Uitterlinden, A. G., Vitart, V., Waldenberger, M., Wild, P. S., Kirin, M., Zeller, T., Zemunik, T., Zhang, Q., Ziegler, A., Blankenberg, S., Boerwinkle, E., Borecki, I. B., Campbell, H., Deary, I. J., Frayling, T. M., Gieger, C., Harris, T. B., Hicks, A. A., Koenig, W., O' Donnell, C. J., Fox, C. S., Pramstaller, P. P., Psaty, B. M., Reiner, A. P., Rotter, J. I., Rudan, I., Snieder, H., Tanaka, T., van Duijn, C. M., Vollenweider, P., Waeber, G., Wilson, J. F., Witteman, J. C. M., Wolffenbuttel, B. H. R., Wright, A. F., Wu, Q., Liu, Y., Jenny, N. S., North, K. E., Felix, J. F., Alizadeh, B. Z., Cupples, L. A., Perry, J. R. B., and Morris, A. P. (2012). Discovery and fine mapping of serum protein loci through transethnic meta-analysis. *Am J Hum Genet*.

[Franke et al., 2010] Franke, A., McGovern, D. P. B., 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., Büning, C., Cohen, A., Colombel, J.-F., Cottone, M., Stronati, L., Denson, T., De Vos, M., D'Inca, R., Dubinsky, M., Edwards, C., Florin, T., Franchimont, D., Gearry, R., Glas, J., Van Gossum, 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., Panés, J., Phillips, A., Proctor, D. D., Regueiro, M., Russell, R., Rutgeerts, P., Sanderson, J., Sans, M., Seibold, F., Steinhart, A. H., Stokkers, P. C. F., 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., and Parkes, M. (2010). Genome-wide meta-analysis increases to 71 the number of confirmed crohn's disease susceptibility loci. *Nat Genet*, 42(12):1118–25.

[Frayling et al., 2008] Frayling, T., Colhoun, H., and Florez, J. (2008). A genetic link between type 2 diabetes and prostate cancer. *Diabetologia*, 51(10):1757–1760.

[Frazer et al., 2007] Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., Gibbs, R. A., Belmont, J. W., Boudreau, A., Hardenbol, P., Leal, S. M., Pasternak, S., Wheeler, D. A., Willis, T. D., Yu, F. L., Yang, H. M., Zeng, C. Q., Gao, Y., Hu, H. R., Hu, W. T., Li, C. H., Lin, W., Liu, S. Q., Pan,

H., Tang, X. L., Wang, J., Wang, W., Yu, J., Zhang, B., Zhang, Q. R., Zhao, H. B., Zhao, H., Zhou, J., Gabriel, S. B., Barry, R., Blumenstiel, B., Camargo, A., Defelice, M., Faggart, M., Goyette, M., Gupta, S., Moore, J., Nguyen, H., Onofrio, R. C., Parkin, M., Roy, J., Stahl, E., Winchester, E., Ziaugra, L., Altshuler, D., Shen, Y., Yao, Z. J., Huang, W., Chu, X., He, Y. G., Jin, L., Liu, Y. F., Shen, Y. Y., Sun, W. W., Wang, H. F., Wang, Y., Xiong, X. Y., Xu, L., Waye, M. M. Y., Tsui, S. K. W., Wong, J. T. F., Galver, L. M., Fan, J. B., Gunderson, K., Murray, S. S., Oliphant, A. R., Chee, M. S., Montpetit, A., Chagnon, F., Ferretti, V., Leboeuf, M., Olivier, J. F., Phillips, M. S., Roumy, S., Sallee, C., Verner, A., Hudson, T. J., Kwok, P. Y., Cai, D. M., Koboldt, D. C., Miller, R. D., Pawlikowska, L., Taillon-Miller, P., Xiao, M., Tsui, L. C., Mak, W., Song, Y. Q., Tam, P. K. H., Nakamura, Y., Kawaguchi, T., Kitamoto, T., Morizono, T., Nagashima, A., Ohnishi, Y., Sekine, A., Tanaka, T., et al. (2007). A second generation human haplotype map of over 3.1 million snps. *Nature*, 449(7164):851–U3.

[Gabriel et al., 2002] Gabriel, S., Schaffner, S., Nguyen, H., Moore, J., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., et al. (2002). The structure of haplotype blocks in the human genome.

[Gaulton et al., 2010] Gaulton, K. J., Nammo, T., Pasquali, L., Simon, J. M., Giresi, P. G., Fogarty, M. P., Panhuis, T. M., Mieczkowski, P., Secchi, A., Bosco, D., Berney, T., Montanya, E., Mohlke, K. L., Lieb, J. D., and Ferrer, J. (2010). A map of open chromatin in human pancreatic islets. *Nat Genet*, 42(3):255–9.

[Gerken et al., 2007] Gerken, T., Girard, C. A., Tung, Y. C., Webby, C. J.,

Saudek, V., Hewitson, K. S., Yeo, G. S., McDonough, M. A., Cunliffe, S., McNeill, L. A., Galvanovskis, J., Rorsman, P., Robins, P., Prieur, X., Coll, A. P., Ma, M., Jovanovic, Z., Farooqi, I. S., Sedgwick, B., Barroso, I., Lindahl, T., Ponting, C. P., Ashcroft, F. M., O'Rahilly, S., and Schofield, C. J. (2007). The obesity-associated fto gene encodes a 2-oxoglutarate-dependent nucleic acid demethylase. *Science*, 318(5855):1469–72.

[Giardine et al., 2005] Giardine, B., Riemer, C., Hardison, R., Burhans, R., Elnitski, L., Shah, P., Zhang, Y., Blankenberg, D., Albert, I., and Taylor, J. (2005). Galaxy: a platform for interactive large-scale genome analysis. *Genome research*, 15(10):1451.

[Giresi et al., 2007] Giresi, P., Kim, J., McDaniell, R., Iyer, V., and Lieb, J. (2007). FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome research*, 17(6):877.

[Giresi and Lieb, 2009] Giresi, P. and Lieb, J. (2009). Isolation of active regulatory elements from eukaryotic chromatin using FAIRE (formaldehyde assisted isolation of regulatory elements). *Methods*, 48(3):233–239.

[Goddard et al., 2000] Goddard, K. A., Hopkins, P. J., Hall, J. M., and Witte, J. S. (2000). Linkage disequilibrium and allele-frequency distributions for 114 single-nucleotide polymorphisms in five populations. *Am J Hum Genet*, 66(1):216–34.

[Grarup et al., 2008] Grarup, N., Andersen, G., Krarup, N., Albrechtsen, A., Schmitz, O., Jorgensen, T., Borch-Johnsen, K., Hansen, T., and Pedersen, O.

(2008). Association testing of novel type 2 diabetes risk alleles in the jazf1, cdc123/camk1d, tspan8, thada, adamts9, and notch2 loci with insulin release, insulin sensitivity, and obesity in a population-based sample of 4,516 glucose-tolerant middle-aged danes. *Diabetes*, 57(9):2534.

[Griffiths-Jones, 2004] Griffiths-Jones, S. (2004). The microrna registry. *Nucleic acids research*, 32(Database Issue):D109.

[Grimson et al., 2007] Grimson, A., Farh, K., Johnston, W., Garrett-Engele, P., Lim, L., and Bartel, D. (2007). Microrna targeting specificity in mammals: determinants beyond seed pairing. *Molecular cell*, 27(1):91–105.

[Gupta et al., 2008] Gupta, S., Dennis, J., Thurman, R., Kingston, R., Stamatoyannopoulos, J., and Noble, W. (2008). Predicting human nucleosome occupancy from primary sequence. *PLoS Computational Biology*, 4(8).

[Gusella et al., 1983] Gusella, J., Wexler, N., Conneally, P., Naylor, S., Anderson, M., Tanzi, R., Watkins, P., Ottina, K., Wallace, M., Sakaguchi, A., et al. (1983). A polymorphic DNA marker genetically linked to Huntington's disease.

[Haines et al., 2005] Haines, J. L., Hauser, M. A., Schmidt, S., Scott, W. K., Olson, L. M., Gallins, P., Spencer, K. L., Kwan, S. Y., Noureddine, M., Gilbert, J. R., Schnetz-Boutaud, N., Agarwal, A., Postel, E. A., and Pericak-Vance, M. A. (2005). Complement factor h variant increases the risk of age-related macular degeneration. *Science*, 308(5720):419–21.

[Helgadottir et al., 2008] Helgadottir, A., Thorleifsson, G., Magnusson, K., GrÈtarsdottir, S., Steinthorsdottir, V., Manolescu, A., Jones, G., Rinkel, G., Blankensteijn, J., and Ronkainen, A. (2008). The same sequence variant on

9p21 associates with myocardial infarction, abdominal aortic aneurysm and intracranial aneurysm. *Nature genetics*, 40(2):217–224.

[Helgadottir et al., 2007] Helgadottir, A., Thorleifsson, G., Manolescu, A., Gretarsdottir, S., Blondal, T., Jonasdottir, A., Sigurdsson, A., Baker, A., Palsson, A., Masson, G., Gudbjartsson, D. F., Magnusson, K. P., Andersen, K., Levey, A. I., Backman, V. M., Matthiasdottir, S., Jonsdottir, T., Palsson, S., Einarsdottir, H., Gunnarsdottir, S., Gylfason, A., Vaccarino, V., Hooper, W. C., Reilly, M. P., Granger, C. B., Austin, H., Rader, D. J., Shah, S. H., Quyyumi, A. A., Gulcher, J. R., Thorgeirsson, G., Thorsteinsdottir, U., Kong, A., and Stefansson, K. (2007). A common variant on chromosome 9p21 affects the risk of myocardial infarction. *Science*, 316(5830):1491–3.

[Helgason et al., 2007] Helgason, A., Palsson, S., Thorleifsson, G., Grant, S. F., Emilsson, V., Gunnarsdottir, S., Adeyemo, A., Chen, Y., Chen, G., Reynisdottir, I., Benediktsson, R., Hinney, A., Hansen, T., Andersen, G., Borch-Johnsen, K., Jorgensen, T., Schafer, H., Faruque, M., Doumatey, A., Zhou, J., Wilensky, R. L., Reilly, M. P., Rader, D. J., Bagger, Y., Christiansen, C., Sigurdsson, G., Hebebrand, J., Pedersen, O., Thorsteinsdottir, U., Gulcher, J. R., Kong, A., Rotimi, C., and Stefansson, K. (2007). Refining the impact of tcf7l2 gene variants on type 2 diabetes and adaptive evolution. *Nat Genet*, 39(2):218–25.

[Heward et al., 1999] Heward, J., Gordon, C., Allahabadia, A., Barnett, A., Franklyn, J., and Gough, S. (1999). The AG polymorphism in exon 1 of the CTLA-4 gene is not associated with systemic lupus erythematosus.

[Hill and Robertson, 1968] Hill, W. G. and Robertson, A. (1968). Linkage disequilibrium in finite populations. *TAG Theoretical and Applied Genetics*, 38:226–231.

[Hirschhorn and Daly, 2005] Hirschhorn, J. N. and Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, 6(2):95–108.

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

[Hubbard et al., 2002] Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., and Down, T. (2002). The ensembl genome database project. *Nucleic acids research*, 30(1):38–41.

[Ioannidis et al., 2009] Ioannidis, J. P. A., Thomas, G., and Daly, M. J. (2009). Validating, augmenting and refining genome-wide association signals. *Nat Rev Genet*, 10(5):318–29.

[Jeffreys et al., 2001] Jeffreys, A., Kauppi, L., and Neumann, R. (2001). Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nature Genetics*, 29(2):217–222.

[Johansson et al., 2009] Johansson, A., Marroni, F., Hayward, C., Franklin, C., Kirichenko, A., Jonasson, I., Hicks, A., Vitart, V., Isaacs, A., and Axenovich, T. (2009). Common variants in the jazf1 gene associated with height identified by linkage and genome-wide association analysis. *Human Molecular Genetics*, 18(2):373.

[Johnson et al., 2007] Johnson, D., Mortazavi, A., Myers, R., and Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. *Science*, 316(5830):1497.

[Kathiresan et al., 2008] Kathiresan, S., Melander, O., Guiducci, C., Surti, A., Burtt, N., Rieder, M., Cooper, G., Roos, C., Voight, B., and Havulinna, A. (2008). Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nature genetics*, 40(2):189–197.

[Kauppi et al., 2004] Kauppi, L., Jeffreys, A. J., and Keeney, S. (2004). Where the crossovers are: recombination distributions in mammals. *Nat Rev Genet*, 5(6):413–24.

[Kavvoura et al., 2007] Kavvoura, F., Akamizu, T., Awata, T., Ban, Y., Chistiakov, D., Frydecka, I., Ghaderi, A., Gough, S., Hiromatsu, Y., Ploski, R., et al. (2007). Cytotoxic T-lymphocyte associated antigen 4 gene polymorphisms and autoimmune thyroid disease: a meta-analysis. *Journal of Clinical Endocrinology & Metabolism*, 92(8):3162.

[Kent, 2002] Kent, W. (2002). Blat the blast-like alignment tool. *Genome research*, 12(4):656–664.

[Kim et al., 2005] Kim, T., Barrera, L., Zheng, M., Qu, C., Singer, M., Richmond, T., Wu, Y., Green, R., and Ren, B. (2005). A high-resolution map of active promoters in the human genome. *Nature*, 436(7052):876–880.

[King et al., 2005] King, D., Taylor, J., Elnitski, L., Chiaromonte, F., Miller, W., and Hardison, R. (2005). Evaluation of regulatory potential and conservation

scores for detecting cis-regulatory modules in aligned mammalian genome sequences. *Genome research*, 15(8):1051.

[Klein et al., 2005] Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J.-Y., Sackler, R. S., Haynes, C., Henning, A. K., SanGiovanni, J. P., Mane, S. M., Mayne, S. T., Bracken, M. B., Ferris, F. L., Ott, J., Barnstable, C., and Hoh, J. (2005). Complement factor h polymorphism in age-related macular degeneration. *Science*, 308(5720):385–389.

[Kochi et al., 2005] Kochi, Y., Yamada, R., Suzuki, A., Harley, J., Shirasawa, S., Sawada, T., Bae, S., Tokuhiro, S., Chang, X., Sekine, A., et al. (2005). A functional variant in FCRL3, encoding Fc receptor-like 3, is associated with rheumatoid arthritis and several autoimmunities. *Nature genetics*, 37(5):478–485.

[Kolbe et al., 2004] Kolbe, D., Taylor, J., Elnitski, L., Eswara, P., Li, J., Miller, W., Hardison, R., and Chiaromonte, F. (2004). Regulatory potential scores from genome-wide three-way alignments of human, mouse, and rat. *Genome research*, 14(4):700.

[Kong et al., 2009] Kong, A., Steinthorsdottir, V., Masson, G., Thorleifsson, G., Sulem, P., Besenbacher, S., Jonasdottir, A., Sigurdsson, A., Kristinsson, K. T., Jonasdottir, A., Frigge, M. L., Gylfason, A., Olason, P. I., Gudjonsson, S. A., Sverrisson, S., Stacey, S. N., Sigurgeirsson, B., Benediktsdottir, K. R., Sigurdsson, H., Jonsson, T., Benediktsson, R., Olafsson, J. H., Johannsson, O. T., Hreidarsson, A. B., Sigurdsson, G., DIAGRAM Consortium, Ferguson-Smith, A. C., Gudbjartsson, D. F., Thorsteinsdottir, U., and Stefansson, K. (2009).

Parental origin of sequence variants associated with complex diseases. *Nature*, 462(7275):868–74.

[Kooner et al., 2011] Kooner, J. S., Saleheen, D., Sim, X., Sehmi, J., Zhang, W., Frossard, P., Been, L. F., Chia, K.-S., Dimas, A. S., Hassanali, N., Jafar, T., Jowett, J. B. M., Li, X., Radha, V., Rees, S. D., Takeuchi, F., Young, R., Aung, T., Basit, A., Chidambaram, M., Das, D., Grundberg, E., Hedman, A. K., Hydrie, Z. I., Islam, M., Khor, C.-C., Kowlessur, S., Kristensen, M. M., Liju, S., Lim, W.-Y., Matthews, D. R., Liu, J., Morris, A. P., Nica, A. C., Pinidiyapathirage, J. M., Prokopenko, I., Rasheed, A., Samuel, M., Shah, N., Shera, A. S., Small, K. S., Suo, C., Wickremasinghe, A. R., Wong, T. Y., Yang, M., Zhang, F., DIAGRAM, MuTHER, Abecasis, G. R., Barnett, A. H., Caulfield, M., Deloukas, P., Frayling, T. M., Froguel, P., Kato, N., Katulanda, P., Kelly, M. A., Liang, J., Mohan, V., Sanghera, D. K., Scott, J., Seielstad, M., Zimmet, P. Z., Elliott, P., Teo, Y. Y., McCarthy, M. I., Danesh, J., Tai, E. S., and Chambers, J. C. (2011). Genome-wide association study in individuals of south asian ancestry identifies six new type 2 diabetes susceptibility loci. *Nat Genet*, 43(10):984–9.

[Kosiol et al., 2008] Kosiol, C., Vinař, T., da Fonseca, R., Hubisz, M., Bustamante, C., Nielsen, R., and Siepel, A. (2008). Patterns of Positive Selection in Six Mammalian Genomes. *PLoS Genetics*, 4(8).

[Krishnamurthy et al., 2006] Krishnamurthy, J., Ramsey, M., Ligon, K., Torrice, C., Koh, A., Bonner-Weir, S., and Sharpless, N. (2006). p16ink4a induces an age-dependent decline in islet regenerative potential. *Nature*, 443(7110):453–457.

[Kruglyak, 1999] Kruglyak, L. (1999). Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genetics*, 22:139–144.

[Kruglyak and Nickerson, 2001] Kruglyak, L. and Nickerson, D. A. (2001). Variation is the spice of life. *Nat Genet*, 27(3):234–6.

[Lander et al., 2001] Lander, E., Linton, L., Birren, B., Nusbaum, C., Zody, M., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

[Lewis et al., 2005] Lewis, B., Burge, C., and Bartel, D. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microrna targets. *Cell*, 120(1):15–20.

[Lewis and Shih, 2003] Lewis, B. and Shih, I. (2003). Prediction of mammalian microrna targets. *Cell*, 115(7):787–798.

[Lewontin, 1964] Lewontin, R. C. (1964). The interaction of selection and linkage. i. general considerations; heterotic models. *Genetics*, 49(1):49–67.

[Li et al., 2008a] Li, H., Ruan, J., and Durbin, R. (2008a). Mapping short dna sequencing reads and calling variants using mapping quality scores. *Genome Res*, 18(11):1851–8.

[Li et al., 2008b] Li, J., Absher, D., Tang, H., Southwick, A., Casto, A., Ramachandran, S., Cann, H., Barsh, G., Feldman, M., Cavalli-Sforza, L., et al. (2008b). Worldwide human relationships inferred from genome-wide patterns of variation. *Science*, 319(5866):1100.

[Li et al., 2010] Li, Y., Willer, C. J., Ding, J., Scheet, P., and Abecasis, G. R. (2010). Mach: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol*, 34(8):816–34.

[Liu et al., 2012] Liu, J. Z., Almarri, M. A., Gaffney, D. J., Mells, G. F., Jostins, L., Cordell, H. J., Ducker, S. J., Day, D. B., Heneghan, M. A., Neuberger, J. M., Donaldson, P. T., Bathgate, A. J., Burroughs, A., Davies, M. H., Jones, D. E., Alexander, G. J., Barrett, J. C., The UK Primary Biliary Cirrhosis (PBC) Consortium, The Wellcome Trust Case Control Consortium 3, Sandford, R. N., and Anderson, C. A. (2012). Dense fine-mapping study identifies new susceptibility loci for primary biliary cirrhosis. *Nat Genet*, 44(10):1137–1141.

[Liu et al., 2010] Liu, J. Z., Tozzi, F., Waterworth, D. M., Pillai, S. G., Muglia, P., Middleton, L., Berrettini, W., Knouff, C. W., Yuan, X., Waeber, G., Vollenweider, P., Preisig, M., Wareham, N. J., Zhao, J. H., Loos, R. J. F., Barroso, I., Khaw, K.-T., Grundy, S., Barter, P., Mahley, R., Kesaniemi, A., McPherson, R., Vincent, J. B., Strauss, J., Kennedy, J. L., Farmer, A., McGuffin, P., Day, R., Matthews, K., Bakke, P., Gulsvik, A., Lucae, S., Ising, M., Brueckl, T., Horstmann, S., Wichmann, H.-E., Rawal, R., Dahmen, N., Lamina, C., Polasek, O., Zgaga, L., Huffman, J., Campbell, S., Kooner, J., Chambers, J. C., Burnett, M. S., Devaney, J. M., Pichard, A. D., Kent, K. M., Satler, L., Lindsay, J. M., Waksman, R., Epstein, S., Wilson, J. F., Wild, S. H., Campbell, H., Vitart, V., Reilly, M. P., Li, M., Qu, L., Wilensky, R., Matthai, W., Hakonarson, H. H., Rader, D. J., Franke, A., Wittig, M., Schäfer, A., Uda, M., Terracciano, A., Xiao, X., Busonero, F., Scheet, P., Schlessinger, D., St Clair, D., Rujescu, D., Abecasis, G. R., Grabe, H. J., Teumer, A., Völzke, H., Petersmann, A., John,

U., Rudan, I., Hayward, C., Wright, A. F., Kolcic, I., Wright, B. J., Thompson, J. R., Balmforth, A. J., Hall, A. S., Samani, N. J., Anderson, C. A., Ahmad, T., Mathew, C. G., Parkes, M., Satsangi, J., Caulfield, M., Munroe, P. B., Farrall, M., Dominiczak, A., Worthington, J., Thomson, W., Eyre, S., Barton, A., Wellcome Trust Case Control Consortium, Mooser, V., Francks, C., and Marchini, J. (2010). Meta-analysis and imputation refines the association of 15q25 with smoking quantity. *Nat Genet*, 42(5):436–40.

[Low et al., 2012] Low, S.-K., Takahashi, A., Cha, P.-C., Zembutsu, H., Kamatani, N., Kubo, M., and Nakamura, Y. (2012). Genome-wide association study for intracranial aneurysm in the japanese population identifies three candidate susceptible loci and a functional genetic variant at ednra. *Hum Mol Genet*, 21(9):2102–10.

[Lowe et al., 2007] Lowe, C., Cooper, J., Brusko, T., Walker, N., Smyth, D., Bailey, R., Bourget, K., Plagnol, V., Field, S., Atkinson, M., et al. (2007). Large-scale genetic fine mapping and genotype-phenotype associations implicate polymorphism in the IL2RA region in type 1 diabetes. *Nature genetics*, 39(9):1074–1082.

[Lynch M, 1998] Lynch M, W. B. (1998). *Genetics and Analysis of Quantitative Traits*. Sinauer, Sunderland, MA.

[Maier et al., 2009] Maier, L., Lowe, C., Cooper, J., Downes, K., Anderson, D., Severson, C., Clark, P., Healy, B., Walker, N., Aubin, C., et al. (2009). IL2RA genetic heterogeneity in multiple sclerosis and type 1 diabetes susceptibility and soluble interleukin-2 receptor production. *PLoS Genetics*, 5(1).

[Maller et al., 2006] Maller, J., George, S., Purcell, S., Fagerness, J., Altshuler, D., Daly, M. J., and Seddon, J. M. (2006). Common variation in three genes, including a noncoding variant in cfh, strongly influences risk of age-related macular degeneration. *Nat Genet*, 38(9):1055–9.

[Maller et al., 2007] Maller, J. B., Fagerness, J. A., Reynolds, R. C., Neale, B. M., Daly, M. J., and Seddon, J. M. (2007). Variation in complement factor 3 is associated with risk of age-related macular degeneration. *Nat Genet*, 39(10):1200–1.

[Manolio et al., 2008] Manolio, T., Brooks, L., and Collins, F. (2008). A HapMap harvest of insights into the genetics of common disease. *The Journal of clinical investigation*, 118(5):1590.

[Manolio et al., 2009] 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. C., McCarroll, S. A., and Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–53.

[Maouche and Schunkert, 2012] Maouche, S. and Schunkert, H. (2012). Strategies beyond genome-wide association studies for atherosclerosis. *Arterioscler Thromb Vasc Biol*, 32(2):170–81.

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

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

[Martone et al., 2003] Martone, R., Euskirchen, G., Bertone, P., Hartman, S., Royce, T., Luscombe, N., Rinn, J., Nelson, F., Miller, P., Gerstein, M., et al. (2003). Distribution of NF-κB-binding sites across human chromosome 22. *Proceedings of the National Academy of Sciences of the United States of America*, 100(21):12247.

[McCarthy et al., 2008] McCarthy, M., Abecasis, G., Cardon, L., Goldstein, D., Little, J., Ioannidis, J., and Hirschhorn, J. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics*, 9(5):356–369.

[McVean et al., 2004] McVean, G. A. T., Myers, S. R., Hunt, S., Deloukas, P., Bentley, D. R., and Donnelly, P. (2004). The fine-scale structure of recombination rate variation in the human genome. *Science*, 304(5670):581–4.

[Mendel, 1865] Mendel, G. (1865). Experiments in plant hybridization (1865). In *Read at the meetings of February 8th, and March 8th*.

[MIGC, 2009] MIGC (2009). Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. *Nat Genet*, 41(3):334–341.

[Mikkelsen et al., 2007] Mikkelsen, T., Ku, M., Jaffe, D., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T., Koche, R., et al.

(2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, 448(7153):553–560.

[Montgomery et al., 2006] Montgomery, S., Griffith, O., Sleumer, M., Bergman, C., Bilenky, M., Pleasance, E., Prychyna, Y., Zhang, X., and Jones, S. (2006). ORegAnno: an open access database and curation system for literature-derived promoters, transcription factor binding sites and regulatory variation. *Bioinformatics*, 22(5):637.

[Mourelatos et al., 2002] Mourelatos, Z., Dostie, J., Paushkin, S., Sharma, A., Charroux, B., Abel, L., Rappsilber, J., Mann, M., and Dreyfuss, G. (2002). mirnps: a novel class of ribonucleoproteins containing numerous micrornas. *Genes and Development*, 16(6):720–728.

[Murabito et al., 2012] Murabito, J. M., White, C. C., Kavousi, M., Sun, Y. V., Feitosa, M. F., Nambi, V., Lamina, C., Schillert, A., Coassin, S., Bis, J. C., Broer, L., Crawford, D. C., Franceschini, N., Frikke-Schmidt, R., Haun, M., Holewijn, S., Huffman, J. E., Hwang, S.-J., Kiechl, S., Kollerits, B., Montasser, M. E., Nolte, I. M., Rudock, M. E., Senft, A., Teumer, A., van der Harst, P., Vitart, V., Waite, L. L., Wood, A. R., Wassel, C. L., Absher, D. M., Allison, M. A., Amin, N., Arnold, A., Asselbergs, F. W., Aulchenko, Y., Bandinelli, S., Barbalic, M., Boban, M., Brown-Gentry, K., Couper, D. J., Criqui, M. H., Dehghan, A., den Heijer, M., Dieplinger, B., Ding, J., Dörr, M., Espinola-Klein, C., Felix, S. B., Ferrucci, L., Folsom, A. R., Fraedrich, G., Gibson, Q., Goodloe, R., Gunjaca, G., Haltmayer, M., Heiss, G., Hofman, A., Kieback, A., Kiemeney, L. A., Kolcic, I., Kullo, I. J., Kritchevsky, S. B., Lackner, K. J., Li, X., Lieb, W., Lohman, K., Meisinger, C., Melzer, D., Mohler, 3rd, E. R.,

Mudnic, I., Mueller, T., Navis, G., Oberhollenzer, F., Olin, J. W., O'Connell, J., O'Donnell, C. J., Palmas, W., Penninx, B. W., Petersmann, A., Polasek, O., Psaty, B. M., Rantner, B., Rice, K., Rivadeneira, F., Rotter, J. I., Seldenrijk, A., Stadler, M., Summerer, M., Tanaka, T., Tybjaerg-Hansen, A., Uitterlinden, A. G., van Gilst, W. H., Vermeulen, S. H., Wild, S. H., Wild, P. S., Willeit, J., Zeller, T., Zemunik, T., Zgaga, L., Assimes, T. L., Blankenberg, S., Boerwinkle, E., Campbell, H., Cooke, J. P., de Graaf, J., Herrington, D., Kardia, S. L. R., Mitchell, B. D., Murray, A., Münzel, T., Newman, A. B., Oostra, B. A., Rudan, I., Shuldiner, A. R., Snieder, H., van Duijn, C. M., Völker, U., Wright, A. F., Wichmann, H.-E., Wilson, J. F., Witteman, J. C. M., Liu, Y., Hayward, C., Borecki, I. B., Ziegler, A., North, K. E., Cupples, L. A., and Kronenberg, F. (2012). Association between chromosome 9p21 variants and the ankle-brachial index identified by a meta-analysis of 21 genome-wide association studies. *Circ Cardiovasc Genet*, 5(1):100–12.

[Musunuru et al., 2010] Musunuru, K., Post, W. S., Herzog, W., Shen, H., O'Connell, J. R., McArdle, P. F., Ryan, K. A., Gibson, Q., Cheng, Y. C., Clearfield, E., Johnson, A. D., Tofler, G., Yang, Q., O'Donnell, C. J., Becker, D. M., Yanek, L. R., Becker, L. C., Faraday, N., Bielak, L. F., Peyser, P. A., Shuldiner, A. R., and Mitchell, B. D. (2010). Association of single nucleotide polymorphisms on chromosome 9p21.3 with platelet reactivity: a potential mechanism for increased vascular disease. *Circ Cardiovasc Genet*, 3(5):445–53.

[Ng et al., 2005] Ng, P., Wei, C., Sung, W., Chiu, K., Lipovich, L., Ang, C., Gupta, S., Shahab, A., Ridwan, A., Wong, C., et al. (2005). Gene identifi-

cation signature (GIS) analysis for transcriptome characterization and genome annotation. *Nature methods*, 2(2):105–111.

[Ng et al., 2010] Ng, S. B., Buckingham, K. J., Lee, C., Bigham, A. W., Tabor, H. K., Dent, K. M., Huff, C. D., Shannon, P. T., Jabs, E. W., Nickerson, D. A., Shendure, J., and Bamshad, M. J. (2010). Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet*, 42(1):30–5.

[NHGRI, 2011] NHGRI (2011). http://www.genome.gov/gwastudies/index.cfm.

[Nica and Dermitzakis, 2008] Nica, A. C. and Dermitzakis, E. T. (2008). Using gene expression to investigate the genetic basis of complex disorders. *Hum Mol Genet*, 17(R2):R129–34.

[Nickerson et al., 1998] Nickerson, D. A., Taylor, S. L., Weiss, K. M., Clark, A. G., Hutchinson, R. G., Stengård, J., Salomaa, V., Vartiainen, E., Boerwinkle, E., and Sing, C. F. (1998). Dna sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. *Nature Genetics*, 19(3):233–240.

[Nothnagel et al., 2009] Nothnagel, M., Ellinghaus, D., Schreiber, S., Krawczak, M., and Franke, A. (2009). A comprehensive evaluation of snp genotype imputation. *Hum Genet*, 125(2):163–71.

[Omori et al., 2009] Omori, S., Tanaka, Y., Horikoshi, M., Takahashi, A., Hara, K., Hirose, H., Kashiwagi, A., Kaku, K., Kawamori, R., Kadowaki, T., Nakamura, Y., and Maeda, S. (2009). Replication study for the association of new meta-analysis-derived risk loci with susceptibility to type 2 diabetes in 6,244 japanese individuals. *Diabetologia*, 52(8):1554–60.

[Ozsolak et al., 2007] Ozsolak, F., Song, J., Liu, X., and Fisher, D. (2007). High-throughput mapping of the chromatin structure of human promoters. *Nature Biotechnology*, 25(2):244–248.

[Pasmant et al., 2007] Pasmant, E., Laurendeau, I., Heron, D., Vidaud, M., Vidaud, D., and Bieche, I. (2007). Characterization of a germ-line deletion, including the entire ink4/arf locus, in a melanoma-neural system tumor family: identification of anril, an antisense noncoding rna whose expression coclusters with arf. *Cancer Research*, 67(8):3963.

[Pasquinelli et al., 2000] Pasquinelli, A., Reinhart, B., Slack, F., Martindale, M., Kuroda, M., Maller, B., Hayward, D., Ball, E., Degnan, B., and Mller, P. (2000). Conservation of the sequence and temporal expression of let-7 heterochronic regulatory rna. *Nature*, 408(6808):86–89.

[Patil et al., 2001] Patil, N., Berno, A. J., Hinds, D. A., Barrett, W. A., Doshi, J. M., Hacker, C. R., Kautzer, C. R., Lee, D. H., Marjoribanks, C., McDonough, D. P., Nguyen, B. T., Norris, M. C., Sheehan, J. B., Shen, N., Stern, D., Stokowski, R. P., Thomas, D. J., Trulson, M. O., Vyas, K. R., Frazer, K. A., Fodor, S. P., and Cox, D. R. (2001). Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science*, 294(5547):1719–23.

[Peden and Farrall, 2011] Peden, J. F. and Farrall, M. (2011). Thirty-five common variants for coronary artery disease: the fruits of much collaborative labour. *Hum Mol Genet*, 20(R2):R198–205.

[Pedersen et al., 2006] Pedersen, J., Bejerano, G., Siepel, A., Rosenbloom, K., Lindblad-Toh, K., Lander, E., Kent, J., Miller, W., and Haussler, D. (2006). Identification and classification of conserved rna secondary structures in the human genome. *PLoS Comput Biol*, 2(4):e33.

[Pe'er et al., 2006] Pe'er, I., de Bakker, P., Maller, J., Yelensky, R., Altshuler, D., and Daly, M. (2006). Evaluating and improving power in whole-genome association studies using fixed marker sets. *Nature genetics*, 38(6):663–667.

[Pennacchio et al., 2006] Pennacchio, L., Ahituv, N., Moses, A., Prabhakar, S., Nobrega, M., Shoukry, M., Minovitsky, S., Dubchak, I., Holt, A., and Lewis, K. (2006). In vivo enhancer analysis of human conserved non-coding sequences. *Nature*, 444(7118):499–502.

[Petrykowska et al., 2008] Petrykowska, H., Vockley, C., and Elnitski, L. (2008). Detection and characterization of silencers and enhancer-blockers in the greater CFTR locus. *Genome research*, 18(8):1238.

[Pickrell et al., 2009] Pickrell, J., Coop, G., Novembre, J., Kudaravalli, S., Li, J., Absher, D., Srinivasan, B., Barsh, G., Myers, R., Feldman, M., et al. (2009). Signals of recent positive selection in a worldwide sample of human populations. *Genome Research*, 19(5):826.

[Piontkivska et al., 2009] Piontkivska, H., Yang, M., Larkin, D., Lewin, H., Reecy, J., and Elnitski, L. (2009). Cross-species mapping of bidirectional promoters enables prediction of unannotated 5' UTRs and identification of species-specific transcripts. *BMC genomics*, 10(1):189.

[Pruitt et al., 2006] Pruitt, K., Tatusova, T., and Maglott, D. (2006). Ncbi reference sequences (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research*.

[Purcell et al., 2007a] Purcell, S., Daly, M. J., and Sham, P. C. (2007a). Whap: haplotype-based association analysis. *Bioinformatics*, 23(2):255–6.

[Purcell et al., 2007b] Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., and Sham, P. C. (2007b). Plink: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*, 81(3):559–75.

[Qi et al., 2010] Qi, L., Cornelis, M. C., Kraft, P., Stanya, K. J., Linda Kao, W. H., Pankow, J. S., Dupuis, J., Florez, J. C., Fox, C. S., Paré, G., Sun, Q., Girman, C. J., Laurie, C. C., Mirel, D. B., Manolio, T. A., Chasman, D. I., Boerwinkle, E., Ridker, P. M., Hunter, D. J., Meigs, J. B., Lee, C.-H., Hu, F. B., van Dam, R. M., Meta-Analysis of Glucose and Insulin-related traits Consortium (MAGIC), and Diabetes Genetics Replication and Meta-analysis (DIAGRAM) Consortium (2010). Genetic variants at 2q24 are associated with susceptibility to type 2 diabetes. *Hum Mol Genet*, 19(13):2706–15.

[Quaranta et al., 2009] Quaranta, M., Burden, A., Griffiths, C., Worthington, J., Barker, J., Trembath, R., and Capon, F. (2009). Differential contribution of cdkal1 variants to psoriasis, crohn's disease and type ii diabetes. *Genes and Immunity*.

[Rada-Iglesias et al., 2008] Rada-Iglesias, A., Ameur, A., Kapranov, P., Enroth, S., Komorowski, J., Gingeras, T., and Wadelius, C. (2008). Whole-genome

maps of USF1 and USF2 binding and histone H3 acetylation reveal new aspects of promoter structure and candidate genes for common human disorders. *Genome Research*, 18(3):380.

[Rada-Iglesias et al., 2005] Rada-Iglesias, A., Wallerman, O., Koch, C., Ameur, A., Enroth, S., Clelland, G., Wester, K., Wilcox, S., Dovey, O., Ellis, P., et al. (2005). Binding sites for metabolic disease related transcription factors inferred at base pair resolution by chromatin immunoprecipitation and genomic microarrays. *Human molecular genetics*, 14(22):3435.

[Reich et al., 2003] Reich, D. E., Gabriel, S. B., and Altshuler, D. (2003). Quality and completeness of snp databases. *Nat Genet*, 33(4):457–8.

[Rioux et al., 2007] Rioux, J. D., Xavier, R. J., Taylor, K. D., Silverberg, M. S., Goyette, P., Huett, A., Green, T., Kuballa, P., Barmada, M. M., Datta, L. W., Shugart, Y. Y., Griffiths, A. M., Targan, S. R., Ippoliti, A. F., Bernard, E.-J., Mei, L., Nicolae, D. L., Regueiro, M., Schumm, L. P., Steinhart, A. H., Rotter, J. I., Duerr, R. H., Cho, J. H., Daly, M. J., and Brant, S. R. (2007). Genome-wide association study identifies new susceptibility loci for crohn disease and implicates autophagy in disease pathogenesis. *Nat Genet*, 39(5):596–604.

[Risch, 1990] Risch, N. (1990). Linkage strategies for genetically complex traits. II. The power of affected relative pairs. *Am J Hum Genet*, 46(2):229–241.

[Risch, 2000] Risch, N. (2000). Searching for genetic determinants in the new millennium. *Nature*, 405(6788):847–856.

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

[Rivera et al., 2005] Rivera, A., Fisher, S. A., Fritsche, L. G., Keilhauer, C. N., Lichtner, P., Meitinger, T., and Weber, B. H. F. (2005). Hypothetical loc387715 is a second major susceptibility gene for age-related macular degeneration, contributing independently of complement factor h to disease risk. *Hum Mol Genet*, 14(21):3227–36.

[Robertson et al., 2007] Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A., et al. (2007). Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nature methods*, 4(8):651–657.

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

[Rozowsky et al., 2009] Rozowsky, J., Euskirchen, G., Auerbach, R., Zhang, Z., Gibson, T., Bjornson, R., Carriero, N., Snyder, M., and Gerstein, M. (2009). PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nature biotechnology*, 27(1):66–75.

[Sabeti et al., 2007] Sabeti, P., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E., McCarroll, S., Gaudet, R., et al. (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature*, 449(7164):913–918.

[Sabo et al., 2004] Sabo, P., Hawrylycz, M., Wallace, J., Humbert, R., Yu, M., Shafer, A., Kawamoto, J., Hall, R., Mack, J., Dorschner, M., et al. (2004).

Discovery of functional noncoding elements by digital analysis of chromatin structure. *Proceedings of the National Academy of Sciences of the United States of America*, 101(48):16837.

[Sabo et al., 2006] Sabo, P., Kuehn, M., Thurman, R., Johnson, B., Johnson, E., Cao, H., Yu, M., Rosenzweig, E., Goldy, J., Haydock, A., et al. (2006). Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays. *Nature methods*, 3(7):511–518.

[Samani et al., 2007] Samani, N., Erdmann, J., Hall, A., Hengstenberg, C., Mangino, M., Mayer, B., Dixon, R., Meitinger, T., Braund, P., and Wichmann, H. (2007). Genomewide association analysis of coronary artery disease. *New England Journal of Medicine*.

[Sasieni, 1997] Sasieni, P. D. (1997). From genotypes to genes: doubling the sample size. *Biometrics*, 53(4):1253–61.

[Saxena et al., 2007] Saxena, R., Voight, B. F., Lyssenko, V., Burtt, N. P., de Bakker, P. I., Chen, H., Roix, J. J., Kathiresan, S., Hirschhorn, J. N., Daly, M. J., Hughes, T. E., Groop, L., Altshuler, D., Almgren, P., Florez, J. C., Meyer, J., Ardlie, K., Bengtsson Bostrom, K., Isomaa, B., Lettre, G., Lindblad, U., Lyon, H. N., Melander, O., Newton-Cheh, C., Nilsson, P., Orho-Melander, M., Rastam, L., Speliotes, E. K., Taskinen, M. R., Tuomi, T., Guiducci, C., Berglund, A., Carlson, J., Gianniny, L., Hackett, R., Hall, L., Holmkvist, J., Laurila, E., Sjogren, M., Sterner, M., Surti, A., Svensson, M., Tewhey, R., Blumenstiel, B., Parkin, M., Defelice, M., Barry, R., Brodeur, W., Camarata, J., Chia, N., Fava, M., Gibbons, J., Handsaker, B., Healy, C., Nguyen, K., Gates,

C., Sougnez, C., Gage, D., Nizzari, M., Gabriel, S. B., Chirn, G. W., Ma, Q., Parikh, H., Richardson, D., Ricke, D., and Purcell, S. (2007). Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science*, 316(5829):1331–6.

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

[Schmid et al., 2000] Schmid, M., Sen, M., Rosenbach, M. D., Carrera, C. J., Friedman, H., and Carson, D. A. (2000). A methylthioadenosine phosphorylase (mtap) fusion transcript identifies a new gene on chromosome 9p21 that is frequently deleted in cancer. *Oncogene*, 19(50):5747–54.

[Scott et al., 2007] Scott, L. J., Mohlke, K. L., Bonnycastle, L. L., Willer, C. J., Li, Y., Duren, W. L., Erdos, M. R., Stringham, H. M., Chines, P. S., Jackson, A. U., Prokunina-Olsson, L., Ding, C. J., Swift, A. J., Narisu, N., Hu, T., Pruim, R., Xiao, R., Li, X. Y., Conneely, K. N., Riebow, N. L., Sprau, A. G., Tong, M., White, P. P., Hetrick, K. N., Barnhart, M. W., Bark, C. W., Goldstein, J. L., Watkins, L., Xiang, F., Saramies, J., Buchanan, T. A., Watanabe, R. M., Valle, T. T., Kinnunen, L., Abecasis, G. R., Pugh, E. W., Doheny, K. F., Bergman, R. N., Tuomilehto, J., Collins, F. S., and Boehnke, M. (2007). A genome-wide association study of type 2 diabetes in finns detects multiple susceptibility variants. *Science*, 316(5829):1341–5.

[Scuteri et al., 2007] Scuteri, A., Sanna, S., Chen, W. M., Uda, M., Albai, G., Strait, J., Najjar, S., Nagaraja, R., Orru, M., Usala, G., Dei, M., Lai, S., Maschio, A., Busonero, F., Mulas, A., Ehret, G. B., Fink, A. A., Weder, A. B., Cooper, R. S., Galan, P., Chakravarti, A., Schlessinger, D., Cao, A., Lakatta, E., and Abecasis, G. R. (2007). Genome-wide association scan shows genetic variants in the fto gene are associated with obesity-related traits. *PLoS Genet*, 3(7):e115.

[Sherry et al., 2001a] Sherry, S., Ward, M., Kholodov, M., Baker, J., Phan, L., Smigielski, E., and Sirotkin, K. (2001a). dbSNP: the NCBI database of genetic variation. *Nucleic acids research*, 29(1):308.

[Sherry et al., 2001b] Sherry, S., Ward, M., Kholodov, M., Baker, J., Phan, L., Smigielski, E., and Sirotkin, K. (2001b). dbSNP: the NCBI database of genetic variation. *Nucleic acids research*, 29(1):308.

[Shu et al., 2010] Shu, X. O., Long, J., Cai, Q., Qi, L., Xiang, Y.-B., Cho, Y. S., Tai, E. S., Li, X., Lin, X., Chow, W.-H., Go, M. J., Seielstad, M., Bao, W., Li, H., Cornelis, M. C., Yu, K., Wen, W., Shi, J., Han, B.-G., Sim, X. L., Liu, L., Qi, Q., Kim, H.-L., Ng, D. P. K., Lee, J.-Y., Kim, Y. J., Li, C., Gao, Y.-T., Zheng, W., and Hu, F. B. (2010). Identification of new genetic risk variants for type 2 diabetes. *PLoS Genet*, 6(9).

[Siepel et al., 2005] Siepel, A., Bejerano, G., Pedersen, J., Hinrichs, A., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L., and Richards, S. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research*, 15(8):1034–1050.

[Simmonds et al., 2006] Simmonds, M., Heward, J., Carr-Smith, J., Foxall, H., Franklyn, J., and Gough, S. (2006). Contribution of single nucleotide polymorphisms within FCRL3 and MAP3K7IP2 to the pathogenesis of Graves' disease.

[Sladek et al., 2007] Sladek, R., Rocheleau, G., Rung, J., Dina, C., Shen, L., Serre, D., Boutin, P., Vincent, D., Belisle, A., Hadjadj, S., Balkau, B., Heude, B., Charpentier, G., Hudson, T. J., Montpetit, A., Pshezhetsky, A. V., Prentki, M., Posner, B. I., Balding, D. J., Meyre, D., Polychronakos, C., and Froguel, P. (2007). A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*, 445(7130):881–5.

[Soranzo et al., 2009] Soranzo, N., Rivadeneira, F., Chinappen-Horsley, U., Malkina, I., Richards, J., Hammond, N., Stolk, L., Nica, A., Inouye, M., and Hofman, A. (2009). Meta-analysis of genome-wide scans for human adult stature identifies novel loci and associations with measures of skeletal frame size. *PLoS Genetics*, 5(4).

[Speliotes et al., 2010] Speliotes, E. K., Willer, C. J., Berndt, S. I., Monda, K. L., Thorleifsson, G., Jackson, A. U., Lango Allen, H., Lindgren, C. M., Luan, J., Mägi, 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., Segrè, A. V., Estrada, K., Liang, L., Nemesh, J., Park, J.-H., Gustafsson, S., Kilpeläinen, 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. B., Peters, M. J., Preuss, M., Ripatti, S., Rivadeneira, F., Sandholt, C., Scott, L. J., Timpson, N. J., Tyrer, J. P., van Wingerden, 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. S., 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. F., Boes, T., Bonnycastle, L. L., Bornstein, S. R., Brown, M. J., Buchanan, T. A., Busonero, F., Campbell, H., Cappuccio, F. P., Cavalcanti-Proença, C., Chen, Y.-D. I., Chen, C.-M., Chines, P. S., Clarke, R., Coin, L., Connell, J., Day, I. N. M., den 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. C., Gieger, C., Gjesing, A. P., Goel, A., Goyette, P., Grallert, H., Grässler, 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., Jørgensen, T., Jousilahti, P., Jula, A., Kaakinen, M., Kajantie, E., Kaplan, L. M., Kathiresan, S., Kettunen, J., Kinnunen, L., Knowles, J. W., Kolcic, I., König, I. R., Koskinen, S., Kovacs, P., Kuusisto, J., Kraft, P., Kvaløy, K., Laitinen, J., Lantieri, O., Lanzani, C., Launer, L. J., Lecoeur, C., Lehtimäki, T., Lettre, G., Liu, J., Lokki, M.-L., Lorentzon, M., Luben,

R. N., Ludwig, B., MAGIC, 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., Paré, G., Parker, A. N., Perola, M., Pichler, I., Pietiläinen, K. H., Platou, C. G. P., Polasek, O., Pouta, A., Rafelt, S., Raitakari, O., Rayner, N. W., Ridderstråle, 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., Tönjes, A., Tuomi, T., van Meurs, J. B. J., van Ommen, G.-J., Vatin, V., Viikari, J., Visvikis-Siest, S., Vitart, V., Vogel, C. I. G., Voight, B. F., Waite, L. L., Wallaschofski, H., Walters, G. B., Widen, E., Wiegand, S., Wild, S. H., Willemsen, G., Witte, D. R., Witteman, J. C., Xu, J., Zhang, Q., Zgaga, L., Ziegler, A., Zitting, P., Beilby, J. P., Farooqi, I. S., Hebebrand, J., Huikuri, H. V., James, A. L., Kähönen, M., Levinson, D. F., Macciardi, F., Nieminen, M. S., Ohlsson, C., Palmer, L. J., Ridker, P. M., Stumvoll, M., Beckmann, J. S., Boeing, H., Boerwinkle, E., Boomsma, D. I., Caulfield, M. J., Chanock, S. J., Collins, F. S., Cupples, L. A., Smith, G. D., Erdmann, J., Froguel, P., Grönberg, H., Gyllensten, U., Hall, P., Hansen, T., Harris, T. B., Hattersley, A. T., Hayes, R. B., Heinrich, J., Hu, F. B., Hveem, K., Illig, T., Jarvelin, M.-R., Kaprio, J., Karpe, F., Khaw, K.-T., Kiemeney, L. A., Krude, H., Laakso, M., Lawlor, D. A., Metspalu, A., Munroe, P. B., Ouwehand, W. H., Pedersen,

O., Penninx, B. W., Peters, A., Pramstaller, P. P., Quertermous, T., Reinehr, T., Rissanen, A., Rudan, I., Samani, N. J., Schwarz, P. E. H., Shuldiner, A. R., Spector, T. D., Tuomilehto, J., Uda, M., Uitterlinden, A., Valle, T. T., Wabitsch, M., Waeber, G., Wareham, N. J., Watkins, H., Procardis Consortium, Wilson, J. F., Wright, A. F., Zillikens, M. C., Chatterjee, N., McCarroll, S. A., Purcell, S., Schadt, E. E., Visscher, P. M., Assimes, T. L., Borecki, I. B., Deloukas, P., Fox, C. S., Groop, L. C., Haritunians, T., Hunter, D. J., Kaplan, R. C., Mohlke, K. L., O'Connell, J. R., Peltonen, L., Schlessinger, D., Strachan, D. P., van Duijn, C. M., Wichmann, H.-E., Frayling, T. M., Thorsteinsdottir, U., Abecasis, G. R., Barroso, I., Boehnke, M., Stefansson, K., North, K. E., McCarthy, M. I., Hirschhorn, J. N., Ingelsson, E., and Loos, R. J. F. (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat Genet*, 42(11):937–48.

[Steinthorsdottir et al., 2007] Steinthorsdottir, V., Thorleifsson, G., Reynisdottir, I., Benediktsson, R., Jonsdottir, T., Walters, G. B., Styrkarsdottir, U., Gretarsdottir, S., Emilsson, V., Ghosh, S., Baker, A., Snorradottir, S., Bjarnason, H., Ng, M. C., Hansen, T., Bagger, Y., Wilensky, R. L., Reilly, M. P., Adeyemo, A., Chen, Y., Zhou, J., Gudnason, V., Chen, G., Huang, H., Lashley, K., Doumatey, A., So, W. Y., Ma, R. C., Andersen, G., Borch-Johnsen, K., Jorgensen, T., van Vliet-Ostaptchouk, J. V., Hofker, M. H., Wijmenga, C., Christiansen, C., Rader, D. J., Rotimi, C., Gurney, M., Chan, J. C., Pedersen, O., Sigurdsson, G., Gulcher, J. R., Thorsteinsdottir, U., Kong, A., and Stefansson, K. (2007). A variant in cdkal1 influences insulin response and risk of type 2 diabetes. *Nat Genet*, 39(6):770–5.

[Stephens et al., 2001a] Stephens, J. C., Schneider, J. A., Tanguay, D. A., Choi, J., Acharya, T., Stanley, S. E., Jiang, R., Messer, C. J., Chew, A., Han, J. H., Duan, J., Carr, J. L., Lee, M. S., Koshy, B., Kumar, A. M., Zhang, G., Newell, W. R., Windemuth, A., Xu, C., Kalbfleisch, T. S., Shaner, S. L., Arnold, K., Schulz, V., Drysdale, C. M., Nandabalan, K., Judson, R. S., Ruano, G., and Vovis, G. F. (2001a). Haplotype variation and linkage disequilibrium in 313 human genes. *Science*, 293(5529):489–93.

[Stephens et al., 2001b] Stephens, M., Smith, N. J., and Donnelly, P. (2001b). A new statistical method for haplotype reconstruction from population data. *American journal of human genetics*, 68(4):978–989.

[Stranger et al., 2007] Stranger, B., Nica, A., Forrest, M., Dimas, A., Bird, C., Beazley, C., Ingle, C., Dunning, M., Flicek, P., Koller, D., et al. (2007). Population genomics of human gene expression. *Nature genetics*, 39(10):1217–1224.

[Stumpf and McVean, 2003] Stumpf, M. P. H. and McVean, G. A. T. (2003). Estimating recombination rates from population-genetic data. *Nat Rev Genet*, 4(12):959–68.

[Syvänen, 1999] Syvänen, A. C. (1999). From gels to chips: "minisequencing" primer extension for analysis of point mutations and single nucleotide polymorphisms. *Hum Mutat*, 13(1):1–10.

[Teo et al., 2007] Teo, Y. Y., Inouye, M., Small, K. S., Gwilliam, R., Deloukas, P., Kwiatkowski, D. P., and Clark, T. G. (2007). A genotype calling algorithm for the Illumina BeadArray platform. *Bioinformatics*, 23(20):2741–2746.

[Teo et al., 2010] Teo, Y.-Y., Small, K. S., and Kwiatkowski, D. P. (2010). Methodological challenges of genome-wide association analysis in africa. *Nat Rev Genet*, 11(2):149–60.

[Teslovich et al., 2010] Teslovich, T. M., Musunuru, K., Smith, A. V., Edmondson, A. C., Stylianou, I. M., Koseki, M., Pirruccello, J. P., Ripatti, S., Chasman, D. I., Willer, C. J., Johansen, C. T., Fouchier, S. W., Isaacs, A., Peloso, G. M., Barbalic, M., Ricketts, S. L., Bis, J. C., Aulchenko, Y. S., Thorleifsson, G., Feitosa, M. F., Chambers, J., Orho-Melander, M., Melander, O., Johnson, T., Li, X., Guo, X., Li, M., Shin Cho, Y., Jin Go, M., Jin Kim, Y., Lee, J.-Y., Park, T., Kim, K., Sim, X., Twee-Hee Ong, R., Croteau-Chonka, D. C., Lange, L. A., Smith, J. D., Song, K., Hua Zhao, J., Yuan, X., Luan, J., Lamina, C., Ziegler, A., Zhang, W., Zee, R. Y. L., Wright, A. F., Witteman, J. C. M., Wilson, J. F., Willemsen, G., Wichmann, H.-E., Whitfield, J. B., Waterworth, D. M., Wareham, N. J., Waeber, G., Vollenweider, P., Voight, B. F., Vitart, V., Uitterlinden, A. G., Uda, M., Tuomilehto, J., Thompson, J. R., Tanaka, T., Surakka, I., Stringham, H. M., Spector, T. D., Soranzo, N., Smit, J. H., Sinisalo, J., Silander, K., Sijbrands, E. J. G., Scuteri, A., Scott, J., Schlessinger, D., Sanna, S., Salomaa, V., Saharinen, J., Sabatti, C., Ruokonen, A., Rudan, I., Rose, L. M., Roberts, R., Rieder, M., Psaty, B. M., Pramstaller, P. P., Pichler, I., Perola, M., Penninx, B. W. J. H., Pedersen, N. L., Pattaro, C., Parker, A. N., Pare, G., Oostra, B. A., O'Donnell, C. J., Nieminen, M. S., Nickerson, D. A., Montgomery, G. W., Meitinger, T., McPherson, R., McCarthy, M. I., McArdle, W., Masson, D., Martin, N. G., Marroni, F., Mangino, M., Magnusson, P. K. E., Lucas, G., Luben, R., Loos, R. J. F., Lokki, M.-L., Lettre, G., Langen-

berg, C., Launer, L. J., Lakatta, E. G., Laaksonen, R., Kyvik, K. O., Kronenberg, F., König, I. R., Khaw, K.-T., Kaprio, J., Kaplan, L. M., Johansson, A., Jarvelin, M.-R., Janssens, A. C. J. W., Ingelsson, E., Igl, W., Kees Hovingh, G., Hottenga, J.-J., Hofman, A., Hicks, A. A., Hengstenberg, C., Heid, I. M., Hayward, C., Havulinna, A. S., Hastie, N. D., Harris, T. B., Haritunians, T., Hall, A. S., Gyllensten, U., Guiducci, C., Groop, L. C., Gonzalez, E., Gieger, C., Freimer, N. B., Ferrucci, L., Erdmann, J., Elliott, P., Ejebe, K. G., Döring, A., Dominiczak, A. F., Demissie, S., Deloukas, P., de Geus, E. J. C., de Faire, U., Crawford, G., Collins, F. S., Chen, Y.-d. I., Caulfield, M. J., Campbell, H., Burtt, N. P., Bonnycastle, L. L., Boomsma, D. I., Boekholdt, S. M., Bergman, R. N., Barroso, I., Bandinelli, S., Ballantyne, C. M., Assimes, T. L., Quertermous, T., Altshuler, D., Seielstad, M., Wong, T. Y., Tai, E.-S., Feranil, A. B., Kuzawa, C. W., Adair, L. S., Taylor, Jr, H. A., Borecki, I. B., Gabriel, S. B., Wilson, J. G., Holm, H., Thorsteinsdottir, U., Gudnason, V., Krauss, R. M., Mohlke, K. L., Ordovas, J. M., Munroe, P. B., Kooner, J. S., Tall, A. R., Hegele, R. A., Kastelein, J. J. P., Schadt, E. E., Rotter, J. I., Boerwinkle, E., Strachan, D. P., Mooser, V., Stefansson, K., Reilly, M. P., Samani, N. J., Schunkert, H., Cupples, L. A., Sandhu, M. S., Ridker, P. M., Rader, D. J., van Duijn, C. M., Peltonen, L., Abecasis, G. R., Boehnke, M., and Kathiresan, S. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, 466(7307):707–13.

[the DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium et al., 2012] the DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, Morris, A. P., Voight, B. F., Teslovich, T. M., Ferreira, T., Segrè, A. V.,

Steinthorsdottir, V., Strawbridge, R. J., Khan, H., Grallert, H., Mahajan, A., Prokopenko, I., Kang, H. M., Dina, C., Esko, T., Fraser, R. M., Kanoni, S., Kumar, A., Lagou, V., Langenberg, C., Luan, J., Lindgren, C. M., Müller-Nurasyid, M., Pechlivanis, S., Rayner, N. W., Scott, L. J., Wiltshire, S., Yengo, L., Kinnunen, L., Rossin, E. J., Raychaudhuri, S., Johnson, A. D., Dimas, A. S., Loos, R. J. F., Vedantam, S., Chen, H., Florez, J. C., Fox, C., Liu, C.-T., Rybin, D., Couper, D. J., Kao, W. H. L., Li, M., Cornelis, M. C., Kraft, P., Sun, Q., van Dam, R. M., Stringham, H. M., Chines, P. S., Fischer, K., Fontanillas, P., Holmen, O. L., Hunt, S. E., Jackson, A. U., Kong, A., Lawrence, R., Meyer, J., Perry, J. R. B., Platou, C. G. P., Potter, S., Rehnberg, E., Robertson, N., Sivapalaratnam, S., Stančáková, A., Stirrups, K., Thorleifsson, G., Tikkanen, E., Wood, A. R., Almgren, P., Atalay, M., Benediktsson, R., Bonnycastle, L. L., Burtt, N., Carey, J., Charpentier, G., Crenshaw, A. T., Doney, A. S. F., Dorkhan, M., Edkins, S., Emilsson, V., Eury, E., Forsen, T., Gertow, K., Gigante, B., Grant, G. B., Groves, C. J., Guiducci, C., Herder, C., Hreidarsson, A. B., Hui, J., James, A., Jonsson, A., Rathmann, W., Klopp, N., Kravic, J., Krjutškov, K., Langford, C., Leander, K., Lindholm, E., Lobbens, S., Männistö, S., Mirza, G., Mühleisen, T. W., Musk, B., Parkin, M., Rallidis, L., Saramies, J., Sennblad, B., Shah, S., Sigursson, G., Silveira, A., Steinbach, G., Thorand, B., Trakalo, J., Veglia, F., Wennauer, R., Winckler, W., Zabaneh, D., Campbell, H., van Duijn, C., Uitterlinden, A. G., Hofman, A., Sijbrands, E., Abecasis, G. R., Owen, K. R., Zeggini, E., Trip, M. D., Forouhi, N. G., Syvänen, A.-C., Eriksson, J. G., Peltonen, L., Nöthen, M. M., Balkau, B., Palmer, C. N. A., Lyssenko, V., Tuomi, T., Isomaa, B., Hunter, D. J., Qi, L., Wellcome Trust Case Control Consortium, Meta-Analyses of Glucose and Insulin-related traits

Consortium (MAGIC) Investigators, Genetic Investigation of ANthropometric Traits (GIANT) Consortium, Asian Genetic Epidemiology Network–Type 2 Diabetes (AGEN-T2D) Consortium, South Asian Type 2 Diabetes (SAT2D) Consortium, Shuldiner, A. R., Roden, M., Barroso, I., Wilsgaard, T., Beilby, J., Hovingh, K., Price, J. F., Wilson, J. F., Rauramaa, R., Lakka, T. A., Lind, L., Dedoussis, G., Njølstad, I., Pedersen, N. L., Khaw, K.-T., Wareham, N. J., Keinanen-Kiukaanniemi, S. M., Saaristo, T. E., Korpi-Hyövälti, E., Saltevo, J., Laakso, M., Kuusisto, J., Metspalu, A., Collins, F. S., Mohlke, K. L., Bergman, R. N., Tuomilehto, J., Boehm, B. O., Gieger, C., Hveem, K., Cauchi, S., Froguel, P., Baldassarre, D., Tremoli, E., Humphries, S. E., Saleheen, D., Danesh, J., Ingelsson, E., Ripatti, S., Salomaa, V., Erbel, R., Jöckel, K.-H., Moebus, S., Peters, A., Illig, T., de Faire, U., Hamsten, A., Morris, A. D., Donnelly, P. J., Frayling, T. M., Hattersley, A. T., Boerwinkle, E., Melander, O., Kathiresan, S., Nilsson, P. M., Deloukas, P., Thorsteinsdottir, U., Groop, L. C., Stefansson, K., Hu, F., Pankow, J. S., Dupuis, J., Meigs, J. B., Altshuler, D., Boehnke, M., and McCarthy, M. I. (2012). Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat Genet*, 44(9):981–990.

[The ENCODE Project Consortium, 2007] The ENCODE Project Consortium (2007). Identification and analysis of functional elements in 1genome by the encode pilot project. *Nature*, 447(7146):799–816.

[Thomas et al., 2008] Thomas, G., Jacobs, K., Yeager, M., Kraft, P., Wacholder, S., Orr, N., Yu, K., Chatterjee, N., Welch, R., and Hutchinson, A. (2008). Multiple loci identified in a genome-wide association study of prostate cancer.

Nature genetics, 40(3):310–315.

[Tsai et al., 2010] Tsai, F.-J., Yang, C.-F., Chen, C.-C., Chuang, L.-M., Lu, C.-H., Chang, C.-T., Wang, T.-Y., Chen, R.-H., Shiu, C.-F., Liu, Y.-M., Chang, C.-C., Chen, P., Chen, C.-H., Fann, C. S. J., Chen, Y.-T., and Wu, J.-Y. (2010). A genome-wide association study identifies susceptibility variants for type 2 diabetes in han chinese. *PLoS Genet*, 6(2):e1000847.

[Ubeda et al., 2006] Ubeda, M., Rukstalis, J., and Habener, J. (2006). Inhibition of cyclin-dependent kinase 5 activity protects pancreatic beta cells from glucotoxicity. *Journal of Biological Chemistry*, 281(39):28858.

[Ueda et al., 2003] Ueda, H., Howson, J., Esposito, L., Heward, J., et al. (2003). Association of the T-cell regulatory gene CTLA4 with susceptibility to autoimmune disease. *Nature*, 423(6939):506–511.

[Vaidya et al., 2003] Vaidya, B., Oakes, E., Imrie, H., Dickinson, A., Perros, P., Kendall-Taylor, P., and Pearce, S. (2003). CTLA4 gene and Graves' disease: association of Graves' disease with the CTLA4 exon 1 and intron 1 polymorphisms, but not with the promoter polymorphism. *Clinical Endocrinology*, 58(6):732.

[Valouev et al., 2008] Valouev, A., Johnson, D., Sundquist, A., Medina, C., Anton, E., Batzoglou, S., Myers, R., and Sidow, A. (2008). Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nature methods*, 5(9):829–834.

[Vella et al., 2005] Vella, A., Cooper, J., Lowe, C., Walker, N., Nutland, S., Widmer, B., Jones, R., Ring, S., McArdle, W., Pembrey, M., et al. (2005). Lo-

calization of a type 1 diabetes locus in the IL2RA/CD25 region by use of tag single-nucleotide polymorphisms. *The American Journal of Human Genetics*, 76(5):773–779.

[Voight et al., 2006] Voight, B., Kudaravalli, S., Wen, X., and Pritchard, J. (2006). A map of recent positive selection in the human genome. *PLoS biology*, 4(3):446.

[Voight et al., 2012] Voight, B. F., Kang, H. M., Ding, J., Palmer, C. D., Sidore, C., Chines, P. S., Burtt, N. P., Fuchsberger, C., Li, Y., Erdmann, J., Frayling, T. M., Heid, I. M., Jackson, A. U., Johnson, T., Kilpeläinen, T. O., Lindgren, C. M., Morris, A. P., Prokopenko, I., Randall, J. C., Saxena, R., Soranzo, N., Speliotes, E. K., Teslovich, T. M., Wheeler, E., Maguire, J., Parkin, M., Potter, S., Rayner, N. W., Robertson, N., Stirrups, K., Winckler, W., Sanna, S., Mulas, A., Nagaraja, R., Cucca, F., Barroso, I., Deloukas, P., Loos, R. J. F., Kathiresan, S., Munroe, P. B., Newton-Cheh, C., Pfeufer, A., Samani, N. J., Schunkert, H., Hirschhorn, J. N., Altshuler, D., McCarthy, M. I., Abecasis, G. R., and Boehnke, M. (2012). The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS Genet*, 8(8):e1002793.

[Voight et al., 2010] 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., Segrè, A. V., van Hoek, M., Navarro, P., Ardlie, K., Balkau, B., Benediktsson, R., Bennett, A. J., Blagieva, R., Boerwinkle, E., Bonny-

castle, L. L., Bengtsson Boström, K., Bravenboer, B., Bumpstead, S., Burtt, N. P., Charpentier, G., Chines, P. S., Cornelis, M., Couper, D. J., Crawford, G., Doney, A. S. F., 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. V., Jørgensen, T., Kao, W. H. L., 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. B., Petersen, A.-K., Platou, C., Proença, 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., Sparsø, 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 Herpt, 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. A., 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., MAGIC investigators, and GIANT Consortium (2010). Twelve type 2 diabetes sus-

ceptibility loci identified through large-scale association analysis. *Nat Genet*, 42(7):579–89.

[Vukcevik, 2009] Vukcevik, D. (2009). *Bayesian and frequentist methods and analyses of genome-wide association studies*. PhD thesis, Oxford University.

[Wall and Pritchard, 2003] Wall, J. D. and Pritchard, J. K. (2003). Haplotype blocks and linkage disequilibrium in the human genome. *Nature Reviews Genetics*, 4(8):587–597.

[Wang et al., 1998] Wang, D. G., Fan, J. B., Siao, C. J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J., Kruglyak, L., Stein, L., Hsie, L., Topaloglou, T., Hubbell, E., Robinson, E., Mittmann, M., Morris, M. S., Shen, N., Kilburn, D., Rioux, J., Nusbaum, C., Rozen, S., Hudson, T. J., Lipshutz, R., Chee, M., and Lander, E. S. (1998). Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science*, 280(5366):1077–82.

[Wang et al., 2005] Wang, W., Barratt, B., Clayton, D., and Todd, J. (2005). Genome-wide association studies: theoretical and practical concerns. *Nature Reviews Genetics*, 6(2):109–118.

[Watson and Crick, 1953] Watson, J. D. and Crick, F. H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–8.

[Weber, 1990] Weber, J. L. (1990). Human dna polymorphisms and methods of analysis. *Curr Opin Biotechnol*, 1(2):166–171.

[Weber, 2005] Weber, M. (2005). New human and mouse microrna genes found by homology search. *FEBS*, 272(1):59–73.

[Wei et al., 2006] Wei, C., Wu, Q., Vega, V., Chiu, K., Ng, P., Zhang, T., Shahab, A., Yong, H., Fu, Y., Weng, Z., et al. (2006). A global map of p53 transcription-factor binding sites in the human genome. *Cell*, 124(1):207–219.

[Wei et al., 2005] Wei, F., Nagashima, K., Ohshima, T., Saheki, Y., Lu, Y., Matsushita, M., Yamada, Y., Mikoshiba, K., Seino, Y., and Matsui, H. (2005). Cdk5-dependent regulation of glucose-stimulated insulin secretion. *Nature medicine*, 11(10):1104–1108.

[Wiggs et al., 2012] Wiggs, J. L., Yaspan, B. L., Hauser, M. A., Kang, J. H., Allingham, R. R., Olson, L. M., Abdrabou, W., Fan, B. J., Wang, D. Y., Brodeur, W., Budenz, D. L., Caprioli, J., Crenshaw, A., Crooks, K., Delbono, E., Doheny, K. F., Friedman, D. S., Gaasterland, D., Gaasterland, T., Laurie, C., Lee, R. K., Lichter, P. R., Loomis, S., Liu, Y., Medeiros, F. A., McCarty, C., Mirel, D., Moroi, S. E., Musch, D. C., Realini, A., Rozsa, F. W., Schuman, J. S., Scott, K., Singh, K., Stein, J. D., Trager, E. H., Vanveldhuisen, P., Vollrath, D., Wollstein, G., Yoneyama, S., Zhang, K., Weinreb, R. N., Ernst, J., Kellis, M., Masuda, T., Zack, D., Richards, J. E., Pericak-Vance, M., Pasquale, L. R., and Haines, J. L. (2012). Common variants at 9p21 and 8q22 are associated with increased susceptibility to optic nerve degeneration in glaucoma. *PLoS Genet*, 8(4):e1002654.

[Willer et al., 2008] Willer, C. J., Sanna, S., Jackson, A. U., Scuteri, A., Bonnycastle, L. L., Clarke, R., Heath, S. C., Timpson, N. J., Najjar, S. S., Stringham,

H. M., Strait, J., Duren, W. L., Maschio, A., Busonero, F., Mulas, A., Albai, G., Swift, A. J., Morken, M. A., Narisu, N., Bennett, D., Parish, S., Shen, H., Galan, P., Meneton, P., Hercberg, S., Zelenika, D., Chen, W.-M., Li, Y., Scott, L. J., Scheet, P. A., Sundvall, J., Watanabe, R. M., Nagaraja, R., Ebrahim, S., Lawlor, D. A., Ben-Shlomo, Y., Davey-Smith, G., Shuldiner, A. R., Collins, R., Bergman, R. N., Uda, M., Tuomilehto, J., Cao, A., Collins, F. S., Lakatta, E., Lathrop, G. M., Boehnke, M., Schlessinger, D., Mohlke, K. L., and Abecasis, G. R. (2008). Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nat Genet*, 40(2):161–9.

[Wingender, 2008] Wingender, E. (2008). The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. *Briefings in bioinformatics*.

[Wray NR, 2008] Wray NR, V. P. (2008). *Statistical genetics : gene mapping through linkage and association*, chapter 6, pages 100–104. Taylor & Francis.

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

[WTCCC, 2012] WTCCC (2012). Bayesian refinement of association signals for 14 loci in three common diseases. *Nature Genetics - in press*.

[Yamauchi et al., 2010] Yamauchi, T., Hara, K., Maeda, S., Yasuda, K., Takahashi, A., Horikoshi, M., Nakamura, M., Fujita, H., Grarup, N., Cauchi, S., Ng, D. P. K., Ma, R. C. W., Tsunoda, T., Kubo, M., Watada, H., Maegawa, H., Okada-Iwabu, M., Iwabu, M., Shojima, N., Shin, H. D., Andersen, G.,

Witte, D. R., Jørgensen, T., Lauritzen, T., Sandbæk, A., Hansen, T., Ohshige, T., Omori, S., Saito, I., Kaku, K., Hirose, H., So, W.-Y., Beury, D., Chan, J. C. N., Park, K. S., Tai, E. S., Ito, C., Tanaka, Y., Kashiwagi, A., Kawamori, R., Kasuga, M., Froguel, P., Pedersen, O., Kamatani, N., Nakamura, Y., and Kadowaki, T. (2010). A genome-wide association study in the japanese population identifies susceptibility loci for type 2 diabetes at ube2e2 and c2cd4a-c2cd4b. *Nat Genet*, 42(10):864–8.

[Yanagawa et al., 1995] Yanagawa, T., Hidaka, Y., Guimaraes, V., Soliman, M., and DeGroot, L. (1995). CTLA-4 gene polymorphism associated with Graves' disease in a Caucasian population. *Journal of Clinical Endocrinology & Metabolism*, 80(1):41–45.

[Yang et al., 2012] Yang, J., Ferreira, T., Morris, A. P., Medland, S. E., Genetic Investigation of ANthropometric Traits (GIANT) Consortium, DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, Madden, P. A. F., Heath, A. C., Martin, N. G., Montgomery, G. W., Weedon, M. N., Loos, R. J., Frayling, T. M., McCarthy, M. I., Hirschhorn, J. N., Goddard, M. E., and Visscher, P. M. (2012). Conditional and joint multiple-snp analysis of gwas summary statistics identifies additional variants influencing complex traits. *Nat Genet*, 44(4):369–75, S1–3.

[Yang et al., 2008] Yang, M., Taylor, J., and Elnitski, L. (2008). Comparative analyses of bidirectional promoters in vertebrates. *BMC bioinformatics*, 9(Suppl 6):S9.

[Yang and Nielsen, 2002] Yang, Z. and Nielsen, R. (2002). Codon-Substitution Models for Detecting Molecular Adaptation at Individual Sites Along Specific Lineages. *Mol. Biol. Evol*, 19(6):908–917.

[Yeo et al., 2009a] Yeo, G., Coufal, N., Liang, T., Peng, G., Fu, X., and Gage, F. (2009a). An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells. *Nature Structural & Molecular Biology*, 16(2):130–137.

[Yeo et al., 2009b] Yeo, G., Coufal, N., Liang, T., Peng, G., Fu, X., and Gage, F. (2009b). An rna code for the fox2 splicing regulator revealed by mapping rna-protein interactions in stem cells. *Nature Structural and Molecular Biology*, 16(2):130–137.

[Yeo et al., 2005] Yeo, G., Van Nostrand, E., Holste, D., Poggio, T., and Burge, C. (2005). Identification and analysis of alternative splicing events conserved in human and mouse. *Proceedings of the National Academy of Sciences of the United States of America*, 102(8):2850.

[Zeggini and Morris, 2010] Zeggini, E. and Morris, A. (2010). *Analysis of Complex Disease Association Studies*. Elsevier Science & Technology.

[Zeggini et al., 2008] Zeggini, E., Scott, L. J., Saxena, R., Voight, B. F., Marchini, J. L., Hu, T., de Bakker, P. I., Abecasis, G. R., Almgren, P., Andersen, G., Ardlie, K., Bostrom, K. B., Bergman, R. N., Bonnycastle, L. L., Borch-Johnsen, K., Burtt, N. P., Chen, H., Chines, P. S., Daly, M. J., Deodhar, P., Ding, C. J., Doney, A. S., Duren, W. L., Elliott, K. S., Erdos, M. R., Frayling, T. M., Freathy, R. M., Gianniny, L., Grallert, H., Grarup, N., Groves, C. J.,

Guiducci, C., Hansen, T., Herder, C., Hitman, G. A., Hughes, T. E., Isomaa, B., Jackson, A. U., Jorgensen, T., Kong, A., Kubalanza, K., Kuruvilla, F. G., Kuusisto, J., Langenberg, C., Lango, H., Lauritzen, T., Li, Y., Lindgren, C. M., Lyssenko, V., Marvelle, A. F., Meisinger, C., Midthjell, K., Mohlke, K. L., Morken, M. A., Morris, A. D., Narisu, N., Nilsson, P., Owen, K. R., Palmer, C. N., Payne, F., Perry, J. R., Pettersen, E., Platou, C., Prokopenko, I., Qi, L., Qin, L., Rayner, N. W., Rees, M., Roix, J. J., Sandbaek, A., Shields, B., Sjogren, M., Steinthorsdottir, V., Stringham, H. M., Swift, A. J., Thorleifsson, G., Thorsteinsdottir, U., Timpson, N. J., Tuomi, T., Tuomilehto, J., Walker, M., Watanabe, R. M., Weedon, M. N., Willer, C. J., Illig, T., Hveem, K., Hu, F. B., Laakso, M., Stefansson, K., Pedersen, O., Wareham, N. J., Barroso, I., Hattersley, A. T., Collins, F. S., Groop, L., McCarthy, M. I., Boehnke, M., and Altshuler, D. (2008). Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat Genet*, 40(5):638–45.

[Zeggini et al., 2007] Zeggini, E., Weedon, M. N., Lindgren, C. M., Frayling, T. M., Elliott, K. S., Lango, H., Timpson, N. J., Perry, J. R., Rayner, N. W., Freathy, R. M., Barrett, J. C., Shields, B., Morris, A. P., Ellard, S., Groves, C. J., Harries, L. W., Marchini, J. L., Owen, K. R., Knight, B., Cardon, L. R., Walker, M., Hitman, G. A., Morris, A. D., Doney, A. S., McCarthy, M. I., and Hattersley, A. T. (2007). Replication of genome-wide association signals in uk samples reveals risk loci for type 2 diabetes. *Science*, 316(5829):1336–41.

[Zeller et al., 2006] Zeller, K., Zhao, X., Lee, C., Chiu, K., Yao, F., Yustein, J., Ooi, H., Orlov, Y., Shahab, A., Yong, H., et al. (2006). Global mapping of

c-Myc binding sites and target gene networks in human B cells. *Proceedings of the National Academy of Sciences*, 103(47):17834.

[Zheng et al., 2007] Zheng, H., Fu, G., Dai, T., and Huang, H. (2007). Migration of endothelial progenitor cells mediated by stromal cell-derived factor-1 [alpha]/cxcr4 via pi3k/akt/enos signal transduction pathway. *Journal of Cardiovascular Pharmacology*, 50(3):274.