BIOINFORMATICS SOFTWARE AND METHODS FOR GENOME-WIDE ASSOCIATION AND CHIP-SEQ STUDIES

by

Ryan Patrick Welch

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in the University of Michigan
2013

Doctoral Committee:

Professor Michael Boehnke, Co-Chair
Associate Research Professor Laura J. Scott, Co-Chair
Professor Goncalo R. Abecasis
Professor David T. Burke
Professor Margit Burmeister
Assistant Professor Maureen Sartor
Peter J. Woolf, FoodWiki LLC

Dedication

To my parents, Richard and Diane, for their love, support, and providing every opportunity a son could hope for.

Acknowledgments

I am very fortunate in my graduate career to have been advised by some truly wonderful mentors. My first lab rotation was with Dr. Peter Woolf, whom I credit with giving a former engineer an interest in statistics, where none previously existed (not an easy task!) I can recall many long meetings, gone completely off topic to the research at hand, that were very inspiring and shaped the way I would think for the rest of my time in graduate school. While our work together did not become a part of my thesis, I can safely say I would not have arrived at this destination without his guidance. I cannot thank him enough. I would like to thank Prof. Michael Boehnke and Prof. Laura Scott for their mentoring and support, and for taking a chance on a Bioinformatics student. They provided me an amazing opportunity to learn in a world class environment, to present and learn at many conferences, and to make a meaningful contribution to type 2 diabetes genetics. Both were infinitely patient with my frequent “it’s almost done!” software projects. And I certainly would not be the public speaker I am today without their advice about giving scientific presentations, and their willingness to listen to my rambling practice talks. I am also grateful to have studied under the guidance of Prof. Maureen Sartor. While we began working together only a year or so before the end of my studies, she taught me many things both about science, and about life, in that short time that I will not forget. I would like to thank my committee members: Prof. David Burke, Prof. Margit Burmeister, Prof. Goncalo Abecasis for their advice, and taking the time to attend my frequent (and long) committee meetings.

One of the greatest benefits to working in our group was the opportunity to work with outstanding postdoctoral fellows and staff members, without whom I could not have finished my dissertation. I would like to thank Dr. Cristen Willer and Dr. Tanya Teslovich for their mentoring, and giving me the chance to work on two very successful thesis projects. Dr. Anne Jackson and Dr. Heather Stringham helped me on countless projects and I would not have made the progress I did without them. Terry Gliedt and Peter

Chines taught me how to write good software, how to manage databases, and provided me many tools and tricks that I was fortunate to learn. My dissertation would not have been possible without support from many staff members and friends within the Bioinformatics Program, and the Center for Statistical Genetics (CSG). Susan Burke, Dawn Keene, Peggy White, and Julia Eussen were amazingly patient with all of problems and questions. I am sure I always had a perfectly good reason to show up in their offices at the last minute.

I think it goes without saying that graduate school is not just about science, but about the relationships and friends that are made throughout. I am fortunate to have made many friends through the Bioinformatics Program, my lab, and my housemates, without whom I would not have had nearly as much fun as I did. From the early days, I remember many great outings with friends Patrick Harrington, Yongsheng Huang, Abhik Shah, Angel Alvarez, Greg Boggy, Huy Vuong, Jayson Falkner, Arun Manoharan, and Datta (whose last name is too long for me to even attempt to spell.) I thank my many housemates from over the years Steffen Froehlich, Chris Iacovella, Andreas Sophocleous, Gang Su, Michael Workman, Aaron Reina, Cameron Nelson, and Mark Reppell, for giving me a great place to come home to, and a moderately clean one at that. I am grateful to my friends at CSG and SPH, who made many long days in the office enjoyable: Clement Ma, Adrian Tan, Pranav Yajnik, Robert Weyant, Peng Zhang, and Matthew Zawistowski. And finally, but most certainly not least, I cannot express my gratitude and thanks to longtime friends Ted Arneault and Connie Chen for listening to my ramblings and frustrations, and being there when I needed it most.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

2hGlu: 2-hr glucose

ChIP-seq: Chromatin immunoprecipitation followed by deep sequencing

DIAGRAM: Diabetes Genetics Replication And Meta-analysis Consortium

FG: Fasting glucose

FI: Fasting insulin

GWAS: Genome-wide association study

LD: Linkage disequilibrium

MAF: Minor allele frequency

MAGIC: Meta-Analysis of Glucose and Insulin-related traits Consortium

OGTT: Oral glucose tolerance test

SNP: Single nucleotide polymorphism

T2D: Type 2 diabetes

TSS: Transcription start site

Abstract

Much of the work in modern human genetics and bioinformatics is focused on identifying the connection between our genomes and disease. Genome-wide association studies (GWAS) aim to identify common genetic variants across the genome associated with a disease or trait. Chromatin immunoprecipitation followed by sequencing (ChIP-seq) studies aim to identify genes potentially regulated by disease-related DNA-binding proteins. I describe two bioinformatic tools for researching associated variants from GWAS and visualizing the properties of their genomic regions. The first tool, Snipper, automatically reports the biological functions of genes near associated variants. The second tool, LocusZoom, creates visualizations of the genomic regions near associated variants. I identify 41 genetic variants associated with glycemic traits (fasting levels of glucose, insulin, and glucose measured 2 hours after ingestion), bringing the total number of variants associated with these traits to 53. Of these variants, 33 are also associated with increased type 2 diabetes (T2D) risk. I apply Snipper and LocusZoom to investigate the variants for connections with their associated traits. Future functional follow-up work to investigate these variants will yield additional insights into the mechanisms behind glucose control, and potentially T2D. For ChIP-seq studies, I describe ChIP-Enrich, a method that identifies likely biological function(s) of DNA-binding proteins given knowledge of the functions of the genes surrounding the proteins' binding sites. ChIP-Enrich is compared with existing methods. The results show that earlier methods do not properly control for the confounding effect of gene length and intergenic distance, and as a result exhibit an inflated type 1 error rate. ChIP-Enrich uses all ChIP-seq peaks for analysis, rather than only those near transcription start sites (TSSs) as in earlier methods, and can therefore potentially identify functions of proteins binding distally to TSSs. ChIP-Enrich may prove useful in future studies for identifying the function of DNA-binding proteins involved in T2D and other multi-genic diseases, and could also be applied to other whole-genome experiments, such as DNA

methylation experiments (MeDIP-seq) and open chromatin sequencing (DNAse-seq or FAIRE.)

Chapter 1 Introduction

Much of the work in my dissertation has been in response to research questions raised by genome-wide association studies (GWAS). GWAS have become an important approach for identifying common genetic variants associated with various complex traits. As of this writing, over 1600 variants for 249 traits have been identified either through individual GWAS, or GWAS meta-analysis (http://www.genome.gov/gwastudies/). It is clear that these studies have been quite successful in identifying genetic variants underlying complex traits, and it is likely that such discoveries will lead to an improved understanding of many diseases and phenotypes present in the general population.

To perform a GWAS, one typically starts by collecting a large sample of many hundreds or thousands of individuals (Hirschhorn and Daly 2005; McCarthy et al. 2008). The number of individuals may vary considerably, depending on the availability of samples, the cost of phenotyping/genotyping, the penetrance of the disease, and the rarity of genetic variants one wishes to consider. These subjects may be sampled by a number of potential strategies depending on the trait or disease, for example: case/control sampling, population cohorts, sib-pairs or family pedigrees. The samples are then genotyped at hundreds of thousands or millions of single nucleotide polymorphisms (SNPs) using a commercial genotyping array. Once the data are collected, tests of association between the trait or disease phenotype and the genotypes of each SNP are conducted. Significant SNPs (either those reaching a genome-wide significance threshold, or at the very least the top ranked SNPs) are then usually taken forward for follow-up genotyping in a stage 2 study. Joint analysis of the significant SNPs in both the original GWAS (stage 1) and the stage 2 study is then performed, and those SNPs exceeding a genome-wide significance threshold ($P \leq 5x10^{-8}$) are then considered to be identified associations (Skol et al. 2006).

It is often the case, however, that individual GWAS are under-powered to detect association owing to the small effect size of the common SNPs being assayed and relatively small sample sizes. To increase sample sizes and therefore power, GWAS on the same trait are meta-analyzed together (Begum et al. 2012). This has proven to be an effective strategy by a number of large-scale meta-analysis consortia identifying low effect variants with sample sizes numbering in the tens (Dupuis et al. 2010; Voight et al. 2010; Strawbridge et al. 2011) and hundreds (Speliotes et al. 2010) (Teslovich et al. 2010) of thousands.

Once associated SNPs have been identified, either by GWAS or GWAS meta-analysis, it is usually unclear how they may be functionally related to the trait of interest (Wang et al. 2010). The associated SNP is often in linkage disequilibrium (LD) with other SNPs or variants in the genomic region, making identification of the true causal variant(s) difficult. Even if the truly causal variant(s) were known, it still may not be clear how that variant(s) functions biologically. One possible mechanism is that the variant causes an amino acid substitution, potentially altering the function of a protein. Another possible mechanism could be that SNPs influence the expression of one or more genes, perhaps by altering DNA regulatory domains such as promoters, enhancers, and silencers, or by changing a genes rate of transcription via synonymous variants within a gene. Therefore, we examine the genes underlying association signals to shed light on the potential mechanism through which these variants are acting.

In chapter 1, I introduce software called Snipper that is designed to aid in the research of genes near associated SNPs. This software eliminates many of the challenges present in systematically researching the numerous genes near association signals. For example, in a recent meta-analysis of GWAS for lipid traits, there were 866 genes within 300KB of 101 independent associated SNPs. Clearly, researching this number of genes manually would prove difficult for a number of reasons: 1) retrieving each gene's information is itself a time-consuming task, 2) reviewing each gene for relevance to the trait, without a method of searching the available information, is quite difficult, and 3) organizing the information in a fashion suitable for review would be exhaustively time-consuming. Snipper provides solutions for these problems by automatically identifying genes near associated SNPs, downloading their latest information from public databases, searching the information for relevant terms, and formatting the information for display to the user.

When considering a genome-wide association signal, we are not only interested in the genes, but other genetic and genomic features in the region. Visualizing these features simultaneously can aid in our understanding of the nature of the association. Such features include: the extent of the association signal, either in terms of p-values, or linkage disequilibrium with nearby variants, position of the signal relative to nearby genes, recombination hotspots, and variant annotations. Seeing these features plotted together can help identify interesting patterns in the region. For example, one might identify potential secondary signals (SNPs) contained within recombination hotspots that are not in LD with the most strongly associated SNP.

In chapter 2, I introduce a tool called LocusZoom that allows visualization of results from a genome-wide association scan or meta-analysis. LocusZoom creates plots that can display all of the features mentioned above. LocusZoom also provides the ability to create plots from published meta-analysis results for multiple traits. This allows researchers to look up genomic regions of interest or genes from their current study in one of our available GWAS or meta-analysis scans and to look for association signals with their trait of interest. LocusZoom improves on previous tools by providing support for multiple sources of LD information (1000 Genomes, HapMap), functional annotation of SNPs, the ability to plot large (> 500 kb) regions, greatly improved automated gene spacing, and a batch mode that can create plots for many traits and loci at once. LocusZoom exists as both a web program and a downloadable software package that can be run locally, without the need to upload results to a server thus providing confidentiality. Our paper introducing LocusZoom was published in Bioinformatics in 2010 (Pruim et al. 2010), and became the 8th most cited paper in the journal for that year. To date, LocusZoom has been used to create over 110,000 plots on our website, the standalone software has been downloaded over 500 times, and our plots are often featured in many prominent publications and journals, for example: (Willer et al. 2009; Teslovich et al. 2010; Voight et al. 2010; Burdon et al. 2011; Scott et al. 2012). My role as part of the LocusZoom team was in collaborating on the design of the plots, creating the software and backend databases, and contributing to the manuscript.

Meta-analysis of GWAS allows one to achieve greater power to detect association by increasing sample sizes (Chapman et al. 2011), but as meta-analysis sample sizes have increased the feasibility of testing small set of SNPs in a second stage of 10's or 100's of thousands of samples had become more limiting. Genotyping costs and small panel sizes (of typically 20-30 SNPs) have prevented follow-up of many potentially associated variants. To genotype a greater number of potentially associated variants in a cost-effective manner, a number of meta-analysis consortia for various heart and metabolic traits partnered with Illumina to create the CardioMetabochip (hereafter referred to simply as the Metabochip) (Voight et al. 2012), a custom Illumina array of ~ 200,000 SNPs designed primarily for two purposes: to facilitate large-scale investigation of loci from previous meta-analyses (~66,000 SNPs), and to provide SNPs for fine-mapping of >250 known associated regions (~120,000 SNPs.)

In chapter 3, I present my work with the Meta-Analysis of Glycaemic and Insulin-related Traits (MAGIC) consortium. Previously, MAGIC conducted meta-analyses of GWAS for glycaemic traits in non-diabetic individuals, identifying 16 novel loci for fasting glucose, 2 for fasting insulin, and 5 for 2-hr glucose (Prokopenko et al. 2009; Dupuis et al. 2010; Saxena et al. 2010). Second stage genotyping was limited to a relatively small number of SNPs for each of these traits, and therefore only a handful of loci were able to be identified as genome-wide significant. In our present work, multiple studies within the consortium genotyped additional samples using the Metabochip, increasing our total discovery and replication sample size to 133,010 / 108,557 / 42,854 for studies of fasting glucose, fasting insulin, and 2-hour glucose, respectively. Our efforts lead to the discovery of 41 additional glycemic associations. My contributions as one of three lead analysts in this effort were to perform the statistical analyses and quality control for GWAS of 2-hr glucose, to research the discovered loci (all traits) with Snipper and visualize them with LocusZoom, to determine the effects of the glucose or insulin-raising alleles on the risk of diabetes in a meta-analysis of T2D GWAS and metabochip-based studies (Morris et al. 2012), and to contribute to our manuscript published in Nature Genetics (Scott et al. 2012).

While GWAS and meta-analysis studies, such as those we describe here, have been successful in identifying SNPs associated with common traits and diseases, they are dependent on having a deep catalog of SNPs and other sequence variants available for inclusion on a genotyping array. Within the last few years, whole genome sequencing of 100's or 1000's of samples by individual research groups had been made possible by the introduction of efficient and large-scale genome sequencers by Illumina, Agilent, and other companies (Shendure and Ji 2008; Altshuler et al. 2010). Genome sequencing allows an unbiased view of the genome without the need for *a priori* knowledge of variants potentially present in a sample. This technology also allows for the discovery and genotyping of much rarer genetic variants (MAF < 1%) than were previously available in GWAS (Cirulli and Goldstein 2010).

Sequencing technology has been quickly adapted for use in other areas beyond detection and analysis of sequence variants. One such application is to identify the binding regions of DNA-binding proteins, using chromatin immunoprecipitation with massively parallel DNA sequencing (ChIP-Seq) (Park 2009). This method allows for sequencing millions of short reads from regions in the genome where a protein of interest is bound. These reads are then fed into a statistical algorithm to call “peaks”, i.e. regions where reads have significantly piled up relative to a background distribution of reads (Zhang et al. 2008; Spyrou et al. 2009). Peaks are then considered to be the locations of the genome where the protein binds, under the biological conditions for the experiment (such as tissue type or cell line, or drug treatment.)

In many cases, the biological function of the protein of interest is not well understood. A first step in understanding the function of the protein in a ChIP-seq experiment is to consider the functions of genes near the binding locations (peaks) of the protein. To do this, one could use gene set enrichment testing (Subramanian et al. 2005), which tests for an enrichment of genes containing a peak within or near biologically related sets of genes. Databases such as Gene Ontology (GO) (Ashburner et al. 2000) maintain sets of genes that are grouped together according to their known molecular functions, biological processes, and cellular localizations. Using GO (or other databases) and this type of analysis, we can attempt to infer the biological function of the protein.

Gene set enrichment testing applied to ChIP-seq peaks requires careful consideration of the potential biases present in the data. In the classic application of gene set enrichment analysis with differential gene expression data, each gene is *a priori* equally likely to be selected as differentially expressed. The same is often not true for ChIP-seq peaks, where genes of longer locus length are more likely to contain a peak (Taher and Ovcharenko 2009). Genes with longer (or shorter) locus lengths also tend to belong to specific GO terms (Ovcharenko et al. 2005; Taher and Ovcharenko 2009). Therefore, gene locus length can act as a confounder and can lead to spurious detection of enriched GO terms. Enrichment testing requires a way to account for gene locus length, and other potential confounders as they are discovered.

In chapter 4, I introduce a method for performing gene set enrichment testing on ChIP-seq peaks, called "ChIP-Enrich." This method is capable of empirically adjusting for gene locus length and other confounding variables that could lead to biased enrichment results. I also compare ChIP-Enrich to two existing methods: Fisher's exact test and a binomial test on the count of peaks within GO terms (Taher and Ovcharenko 2009), Through simulation and permutation, I show examples under which these methods have strongly anti-conservative type 1 error rates, while ChIP-Enrich has the expected (or slightly conservative) type 1 error rate, and therefore produces the best calibrated results (as defined by not exceeding the nominal thresholds for type 1 error) for performing gene set enrichment on ChIP-seq data. I also identify another potential confounder not currently considered by other methods – the mappability of a gene locus – and show how ChIP-Enrich is able to adjust for it. Through the application of ChIP-Enrich to experimental data, I show that it is able to identify the function of two well-studied transcription factors, E2F transcription factor 4 (E2F4) and glucocorticoid receptor (GR), and also discover a potentially novel function for GR in angiogenesis. With each dataset, I investigate whether the same GO terms are detectable using only those peaks near the transcription start site (TSS), instead of using all peaks in the data, and show that much of the information regarding the function of GR would be lost without considering all peaks in the data. I have implemented ChIP-Enrich as an R package soon to be available through Bioconductor (Gentleman et al. 2004), providing researchers with an easy-to-use implementation that can test for gene set enrichment on over 15 databases spanning 20,000 sets of genes.

Chapter 2 Snipper: a tool for extracting and searching biological annotations of genes implicated by trait-associated SNPs

2.1 Introduction

During the last few years, genome-wide association studies (GWAS) have been conducted for a multitude of traits. A GWAS may identify a handful to hundreds of trait-associated SNPs, located both within and outside of genes. It is usually not known how these genetic variants, or variants being tagged, function to influence trait variability. Biological follow-up could focus on SNPs with interesting annotation or on genes with potentially relevant biological function. To prioritize genes within an associated region, typical first steps are to search 1) the literature for known associations with the trait or related phenotypes and for biological functions, 2) biological databases for relevant gene annotation, and 3) OMIM entries for candidate genes. GWAS SNPs are more likely than randomly chosen SNPs to have eQTL associations (Gamazon et al. 2010). Searching the annotations of eQTL-associated genes can provide additional biological evidence. The rapid expansion in the number of associated regions identified from GWAS (Manolio et al. 2009) makes this type of search increasingly challenging to perform manually.

To assist in this task, we developed Snipper, a command-line bioinformatics utility. Snipper performs four primary functions to investigate genes potentially related to associated SNPs (either by proximity or by association with gene expression) and identify genes biologically-related to the trait of interest. For each user-provided SNP, Snipper 1) finds genes located nearby in the genome or whose expression is associated with the SNP; 2) downloads relevant information on each of these genes from public databases; 3) searches information on these genes with user-supplied search terms; and 4) collects and organizes the resulting information into a single report.

There are a number of tools available for candidate gene prioritization, but there are few that can research a large number of genes and compile the information in a single, organized report. Existing tools that partially address this task include GeneSniffer (Thornblad et al. 2007) and VarioWatch (formerly GenoWatch) (Chen et al. 2008), however, these tools are focused more on gene prioritization than exhaustive yet easy-to-read compilation of information. Snipper also places a greater emphasis on text matches: searching all gene information, searching PubMed and OMIM for relevant matches to user-supplied terms, and highlighting relevant terms throughout the report, making reading and identifying interesting biological evidence easier. Snipper is a downloadable open-source tool, which allows users to integrate it into their own pipelines, and avoids upload of sensitive information such as SNP names to an external site. Finally, Snipper includes information on genes whose expression is associated with the user-provided SNPs, in addition to focusing on genes in the nearby genomic intervals.

2.2 Methods

2.2.1 Implementation

Snipper is a command line-driven software utility written entirely in Python, with XML parsing capabilities provided by BeautifulSoup (http://www.crummy.com/software/BeautifulSoup/). Positions of SNPs and genes are downloaded from the UCSC FTP site and stored locally. Annotation information on genes is downloaded from NCBI using the E-Utilities API (Maglott et al. 2011) at each invocation of the program to ensure that returned data are up to date. The same API is used to download data from OMIM and PubMed. All information is stored locally in an object-oriented database in memory before being formatted for display. The HTML report is created using Sphinx (http://sphinx.pocoo.org), a module for Python used for creating formatted documentation. Typical runtime is ~6 seconds per gene on an Intel Xeon X5660 (2.8 Ghz) CPU, although runtime varies depending on the amount of information available for each gene. Snipper requires Python version 2.6 or later.

2.2.2 Databases and gene annotations

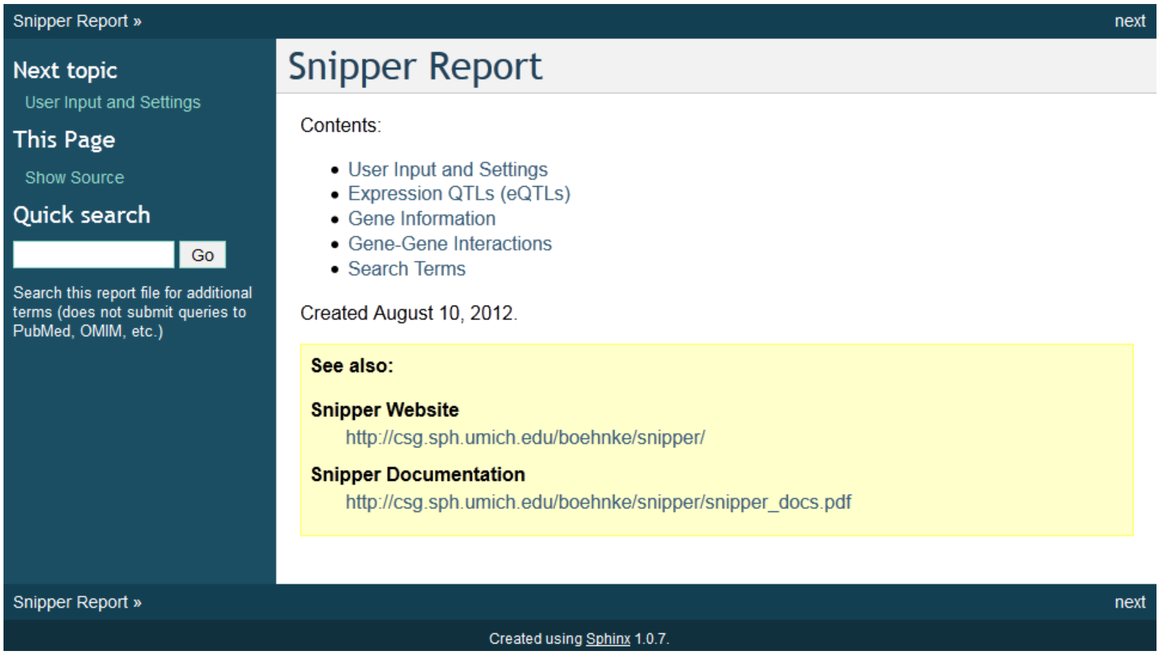
We currently use NCBI databases as our primary resource for gene-related information. Gene annotations are retrieved from Entrez Gene, and include: full gene description, summary text regarding the gene's function and cellular localization, phenotypes, KEGG pathways, Gene ontology (GO) terms, and GeneRIFs. We query the OMIM database for phenotypes related to each gene, and return the OMIM description and OMIM ID. We also query the PubMed data-base for articles linked to a gene, and display a user-defined number of the most recent article citations. If a user provides search terms, we search for PubMed articles matching any one of the relevant terms and the gene, and display these articles as well. The user can optionally choose to perform a search for each term and gene, rather than performing a single search that matches any term. Genes with mRNA expression levels associated with user-provided SNPs are found by querying the SCAN database (Gamazon et al. 2010). We search the Michigan Molecular Interactions (MiMI) (Jayapandian et al. 2007) database for direct interactions between pairs of genes. For each interaction, we display the GO cellular component/process/function terms, the directionality of the interaction, the public databases in which the interaction was reported, whether the interaction was identified in vivo or in vitro, and the PubMed IDs for articles that cite this interaction.

2.2.3 Usage

The user provides a SNP or list of SNPs, for example the most significant trait-associated SNP from each GWAS region. Snipper searches for genes located nearby or whose mRNA expression levels are associated with each SNP, and downloads information for each gene. For genes selected based on proximity to the SNP, the user can control how far from each SNP to search and/or the maximum number of genes to return. For genes selected based on association of expression with a given SNP, the user can control the p-value threshold used to select genes. The user may supply a list of search terms, in which case Snipper will search the information on each gene for matches with these terms. Snipper also searches the PubMed database for each combination of gene and search term, and provides a list of articles that match both. All this information is collected together into an HTML formatted report. A full list of parameters can be found in the documentation online.

2.2.4 HTML report

Snipper creates a formatted HTML report containing the information pulled from the various databases listed previously. The report consists of 5 main sections (Figure 2.1). The first section lists user input and settings given to run Snipper (Figure 2.2). An expression QTL section lists each gene whose expression was associated with a SNP given by the user, and the relevant information on each association from the SCAN database (Figure 2.3) (Gamazon et al. 2010). The gene information section contains each gene and its data from Entrez Gene, as well as PubMed articles, and search terms that matched each gene (Figure 2.5, Figure 2.6.) The gene-gene interaction section lists direct interactions between genes in the MiMI database (Figure 2.4) (Jayapandian et al. 2007). Finally, a search terms section lists each search term given by the user, and where they matched within each gene's information (Figure 2.7). This allows the user to easily navigate to genes that were relevant to a particular search query, for example all of those genes which match "diabetes" or "insulin."


Snipper Report »
next
Next topic
User Input and Settings
This Page
Show Source
Quick search
Go
Search this report file for additional terms (does not submit queries to PubMed, OMIM, etc.)
Snipper Report »
next
Created using Sphinx 1.0.7.

Snipper Report

Contents:

- User Input and Settings
- Expression QTLs (eQTLs)
- Gene Information
- Gene-Gene Interactions
- Search Terms

Created August 10, 2012.

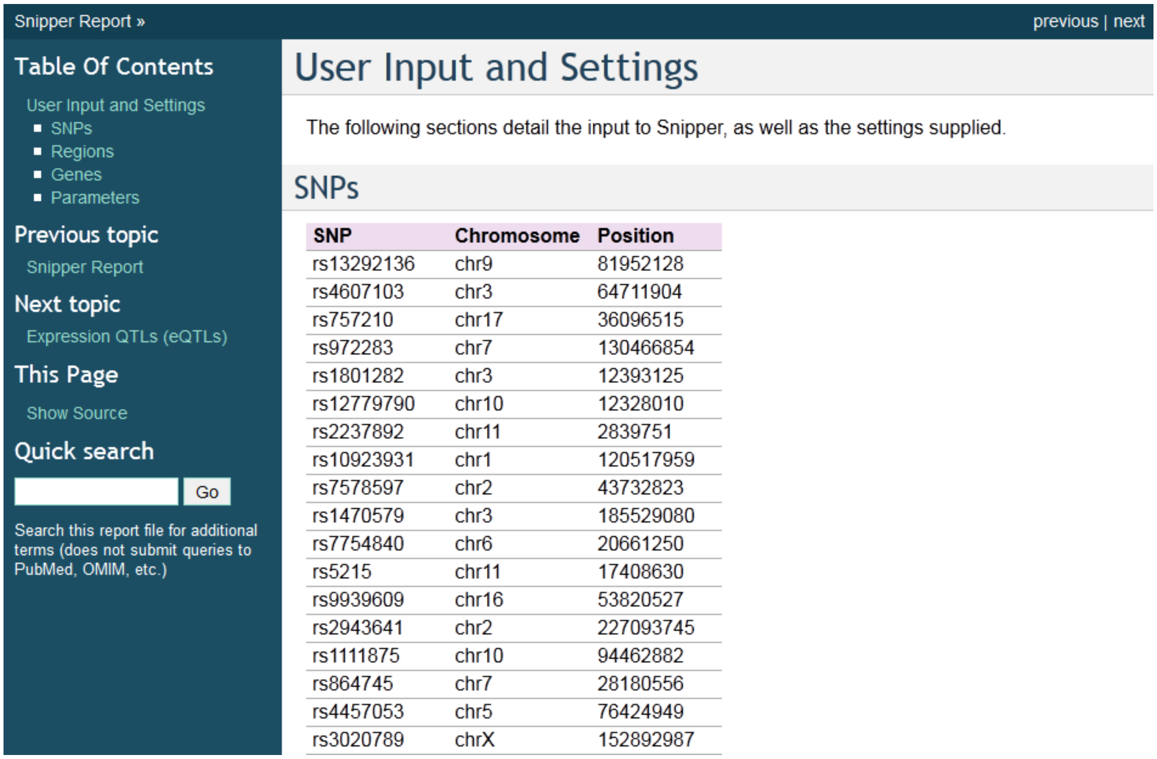
See also:

Snipper Website
http://csg.sph.umich.edu/boehnke/snipper/

Snipper Documentation
http://csg.sph.umich.edu/boehnke/snipper/snipper_docs.pdf

Figure 2.1. Snipper HTML report table of contents. The initial table of contents page lists each section of information, the date when the report was created, and references to the Snipper website and documentation. A

search bar on the left is also available for searching through the report after-the-fact (queries will not be sent to NCBI, OMIM, PubMed, etc.).



Search this report file for additional terms (does not submit queries to PubMed, OMIM, etc.)

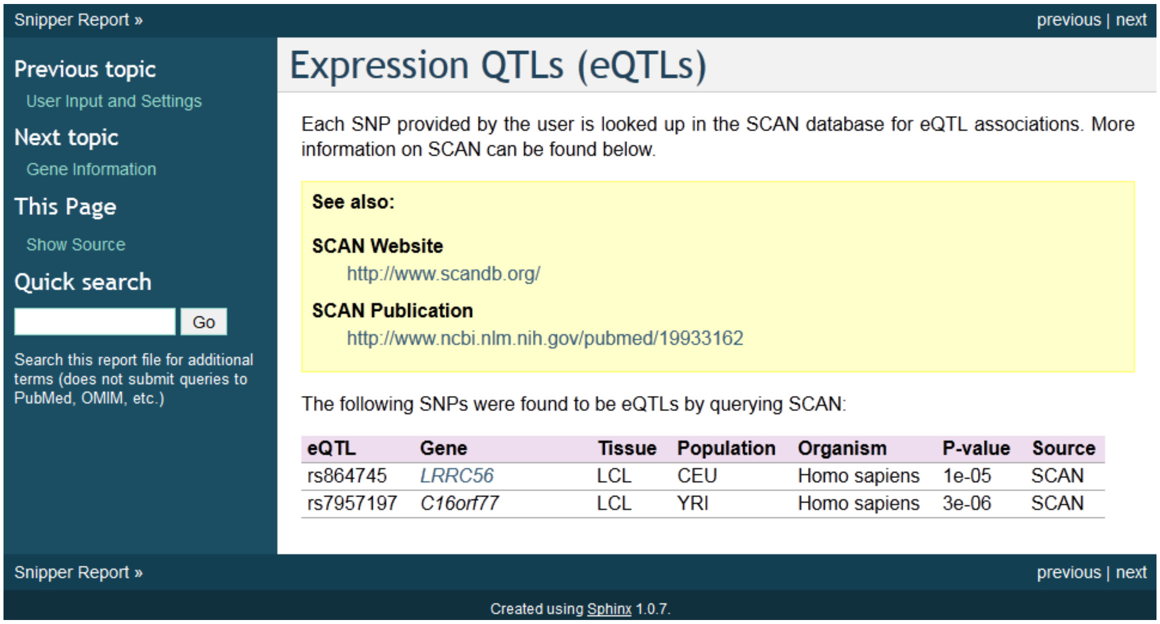
User Input and Settings

The following sections detail the input to Snipper, as well as the settings supplied.

SNPs

SNP	Chromosome	Position
rs13292136	chr9	81952128
rs4607103	chr3	64711904
rs757210	chr17	36096515
rs972283	chr7	130466854
rs1801282	chr3	12393125
rs12779790	chr10	12328010
rs2237892	chr11	2839751
rs10923931	chr1	120517959
rs7578597	chr2	43732823
rs1470579	chr3	185529080
rs7754840	chr6	20661250
rs5215	chr11	17408630
rs9939609	chr16	53820527
rs2943641	chr2	227093745
rs1111875	chr10	94462882
rs864745	chr7	28180556
rs4457053	chr5	76424949
rs3020789	chrX	152892987

Figure 2.2. User input and settings section. All of the settings used to run Snipper are listed on this page (truncated to fit – additional parameters are listed further down the page.)



Expression QTLs (eQTLs)

Each SNP provided by the user is looked up in the SCAN database for eQTL associations. More information on SCAN can be found below.

See also:

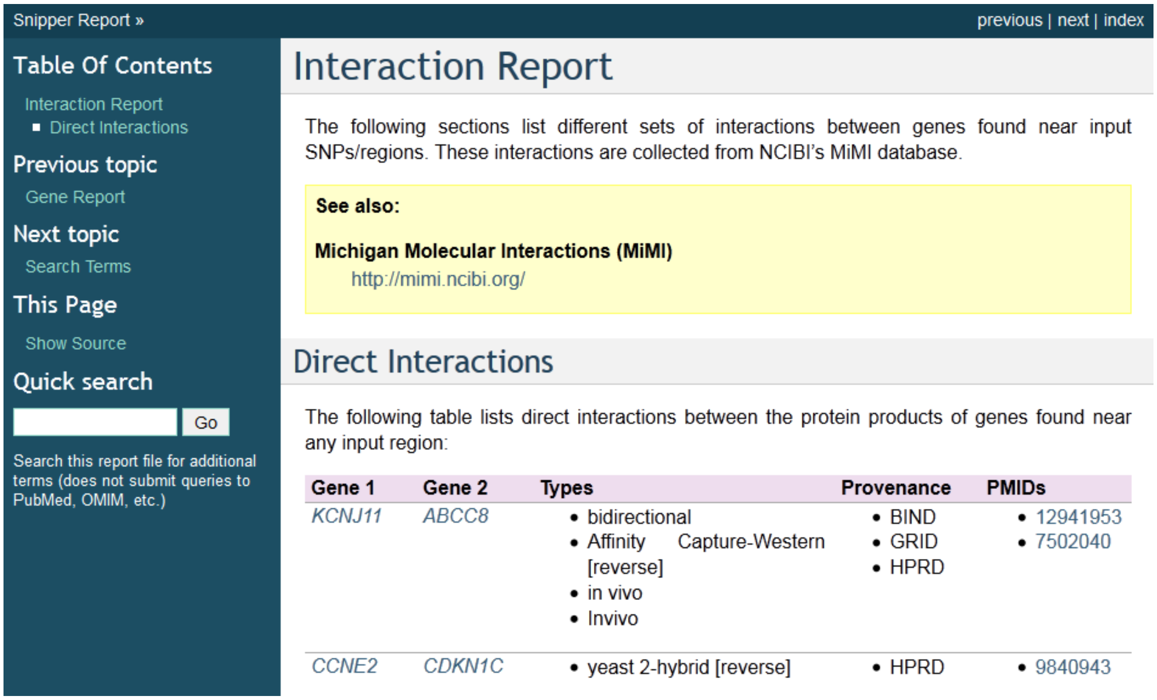
SCAN Website
http://www.scandb.org/

SCAN Publication
http://www.ncbi.nlm.nih.gov/pubmed/19933162

The following SNPs were found to be eQTLs by querying SCAN:

eQTL	Gene	Tissue	Population	Organism	P-value	Source
rs864745	*LRRC56*	LCL	CEU	Homo sapiens	1e-05	SCAN
rs7957197	*C16orf77*	LCL	YRI	Homo sapiens	3e-06	SCAN

Figure 2.3. Expression QTLs from SCAN. Each gene whose expression is associated with a user-given SNP is shown here, along with the relevant data from SCAN: the tissue type/cell line, population, organism, and p-value for the association test.



Interaction Report

The following sections list different sets of interactions between genes found near input SNPs/regions. These interactions are collected from NCIBI's MiMI database.

See also:

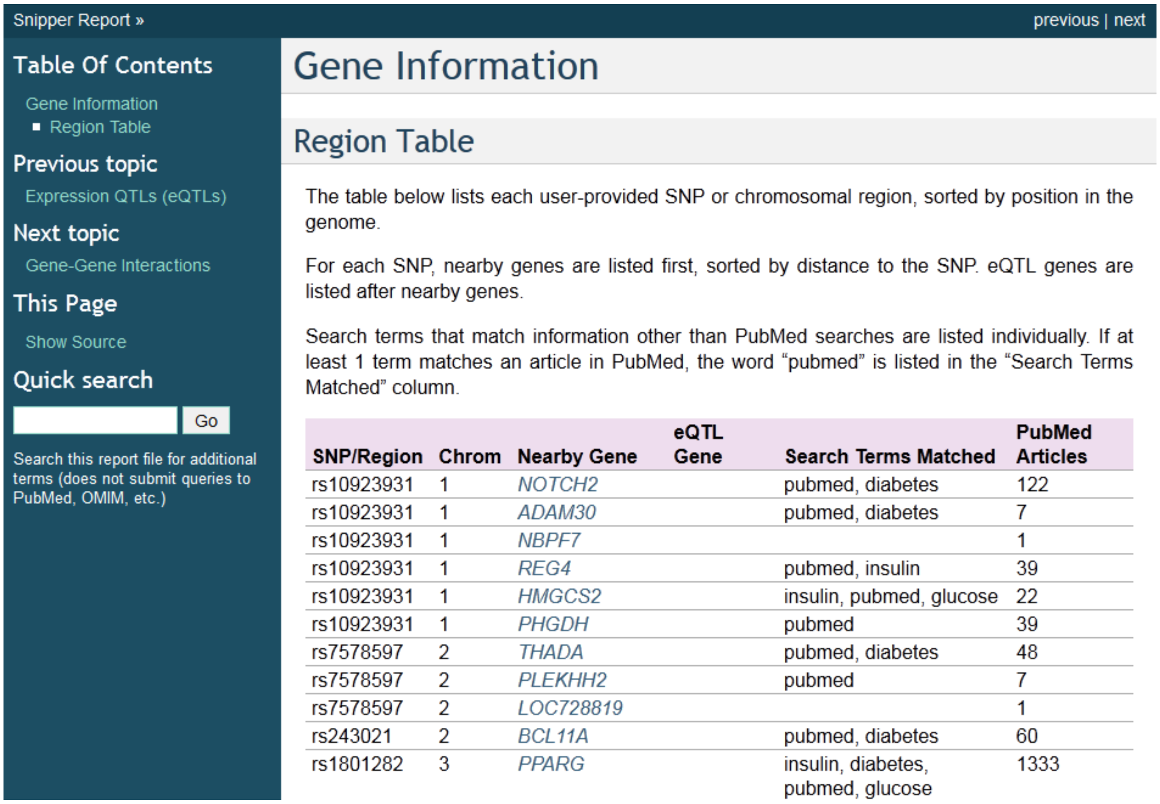
Michigan Molecular Interactions (MiMI)
http://mimi.ncibi.org/

Direct Interactions

The following table lists direct interactions between the protein products of genes found near any input region:

Gene 1	Gene 2	Types	Provenance	PMIDs
KCNJ11	*ABCC8*	• bidirectional • Affinity Capture-Western [reverse] • in vivo • Invivo	• BIND • GRID • HPRD	• 12941953 • 7502040
CCNE2	*CDKN1C*	• yeast 2-hybrid [reverse]	• HPRD	• 9840943

Figure 2.4. Gene-gene interactions section. This section lists directed interactions between all genes (those both near SNPs and within regions), along with any accompanying metadata downloaded from MiMI, including the type of interaction, the database from which the interaction was obtained, and PubMed IDs for articles that either cite the interaction, or claim to have discovered it.



Gene Information

Region Table

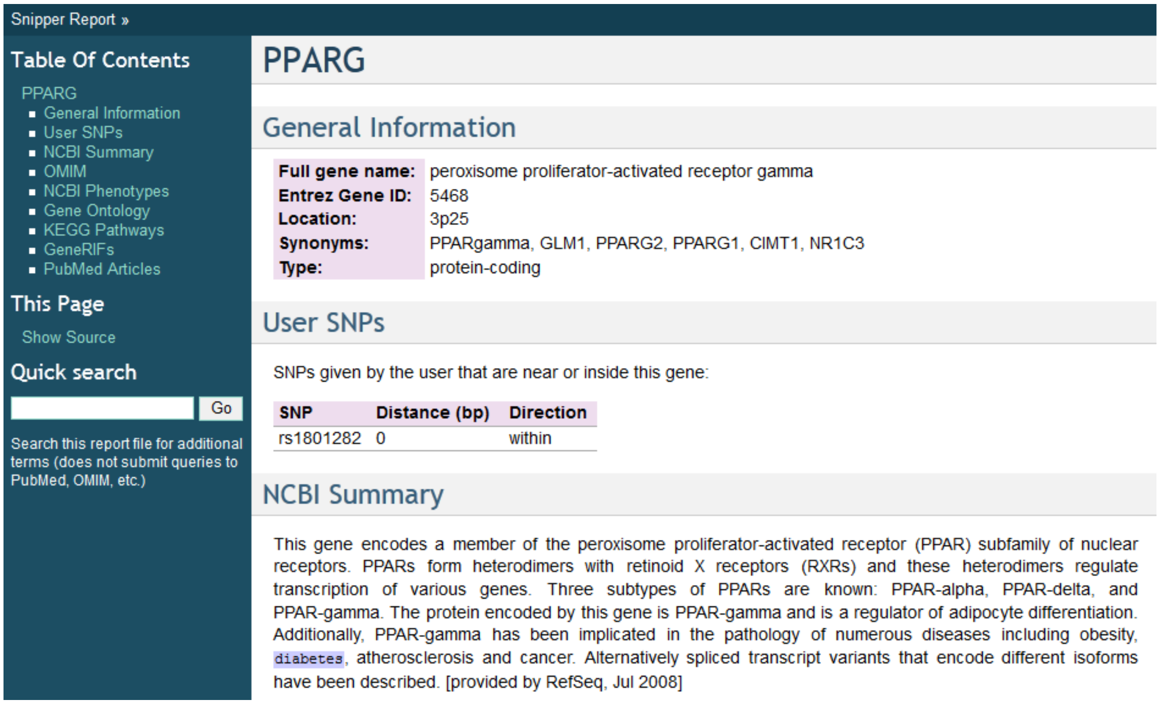
The table below lists each user-provided SNP or chromosomal region, sorted by position in the genome.

For each SNP, nearby genes are listed first, sorted by distance to the SNP. eQTL genes are listed after nearby genes.

Search terms that match information other than PubMed searches are listed individually. If at least 1 term matches an article in PubMed, the word “pubmed” is listed in the “Search Terms Matched” column.

SNP/Region	Chrom	Nearby Gene	eQTL Gene	Search Terms Matched	PubMed Articles
rs10923931	1	*NOTCH2*		pubmed, diabetes	122
rs10923931	1	*ADAM30*		pubmed, diabetes	7
rs10923931	1	*NBPF7*			1
rs10923931	1	*REG4*		pubmed, insulin	39
rs10923931	1	*HMGCS2*		insulin, pubmed, glucose	22
rs10923931	1	*PHGDH*		pubmed	39
rs7578597	2	*THADA*		pubmed, diabetes	48
rs7578597	2	*PLEKHH2*		pubmed	7
rs7578597	2	*LOC728819*			1
rs243021	2	*BCL11A*		pubmed, diabetes	60
rs1801282	3	*PPARG*		insulin, diabetes, pubmed, glucose	1333

Figure 2.5. Gene information section. This section contains a table listing each gene found near the user-provided SNPs, or those genes and/or regions given by the user. Search terms that match the information for that gene are listed in the table. The number of PubMed articles is also listed for each gene. An example of an individual gene is given in Figure 2.6.



PPARG

General Information

Full gene name:	peroxisome proliferator-activated receptor gamma
Entrez Gene ID:	5468
Location:	3p25
Synonyms:	PPARgamma, GLM1, PPARG2, PPARG1, CIMT1, NR1C3
Type:	protein-coding

User SNPs

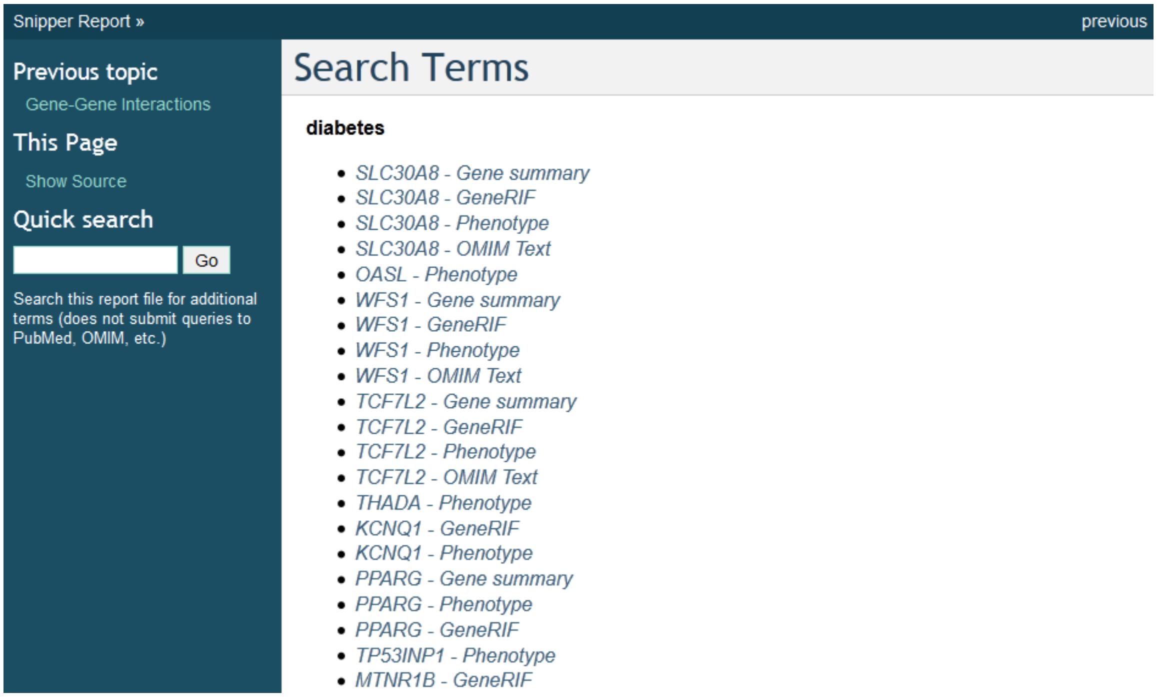
SNPs given by the user that are near or inside this gene:

SNP	Distance (bp)	Direction
rs1801282	0	within

NCBI Summary

This gene encodes a member of the peroxisome proliferator-activated receptor (PPAR) subfamily of nuclear receptors. PPARs form heterodimers with retinoid X receptors (RXRs) and these heterodimers regulate transcription of various genes. Three subtypes of PPARs are known: PPAR-alpha, PPAR-delta, and PPAR-gamma. The protein encoded by this gene is PPAR-gamma and is a regulator of adipocyte differentiation. Additionally, PPAR-gamma has been implicated in the pathology of numerous diseases including obesity, diabetes, atherosclerosis and cancer. Alternatively spliced transcript variants that encode different isoforms have been described. [provided by RefSeq, Jul 2008]

Figure 2.6. Example of a gene's information section. Each gene is listed along with the user-provided SNPs that were near it, information retrieved from NCBI, OMIM, and PubMed such as the gene summary, OMIM report, Gene Ontology and KEGG pathways, Gene References Into Function (GeneRIFs), and recent PubMed citations. Search terms are highlighted in purple (e.g. "diabetes" within the NCBI summary shown above.)



Search Terms

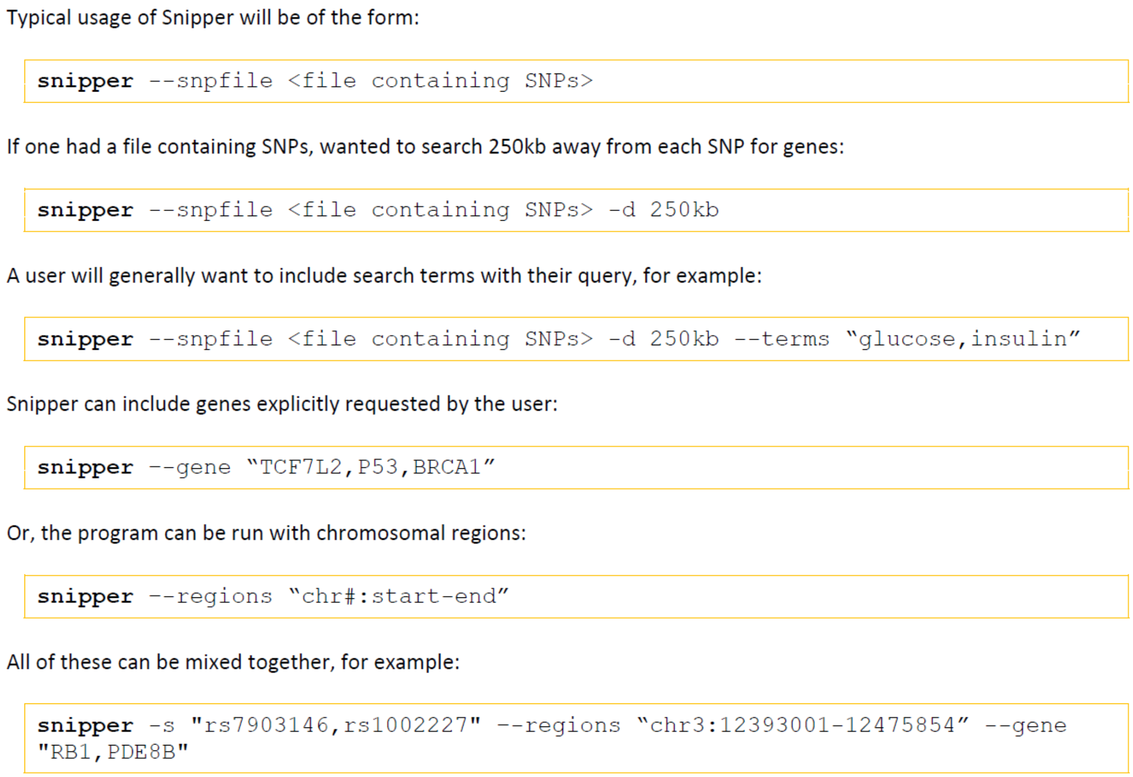
diabetes

- *SLC30A8 - Gene summary*
- *SLC30A8 - GeneRIF*
- *SLC30A8 - Phenotype*
- *SLC30A8 - OMIM Text*
- *OASL - Phenotype*
- *WFS1 - Gene summary*
- *WFS1 - GeneRIF*
- *WFS1 - Phenotype*
- *WFS1 - OMIM Text*
- *TCF7L2 - Gene summary*
- *TCF7L2 - GeneRIF*
- *TCF7L2 - Phenotype*
- *TCF7L2 - OMIM Text*
- *THADA - Phenotype*
- *KCNQ1 - GeneRIF*
- *KCNQ1 - Phenotype*
- *PPARG - Gene summary*
- *PPARG - Phenotype*
- *PPARG - GeneRIF*
- *TP53INP1 - Phenotype*
- *MTNR1B - GeneRIF*

Figure 2.7. Search terms section. Search terms provided by the user before Snipper is executed are listed here, along with the location of where it matched within each gene's information. The user can click one of the links and jump directly to where the term matched.

2.2.5 Snipper interface

Snipper can be run either as a command-line program and using a graphical user interface. Figure 2.8 gives a synopsis of the multiple ways of running Snipper from the command-line. A full list of options is available in the program documentation (Supplementary Data 2.4.) Snipper also provides a GUI interface, which simplifies the operation of the program for users unfamiliar with command-line applications Figure 2.9. The interface is implemented using Python/Tk (http://wiki.python.org/moin/TkInter.) All of the options available to the command-line are also available from within the GUI.



Typical usage of Snipper will be of the form:

```
snipper --snpfile <file containing SNPs>
```

If one had a file containing SNPs, wanted to search 250kb away from each SNP for genes:

```
snipper --snpfile <file containing SNPs> -d 250kb
```

A user will generally want to include search terms with their query, for example:

```
snipper --snpfile <file containing SNPs> -d 250kb --terms “glucose,insulin”
```

Snipper can include genes explicitly requested by the user:

```
snipper --gene “TCF7L2,P53,BRCA1”
```

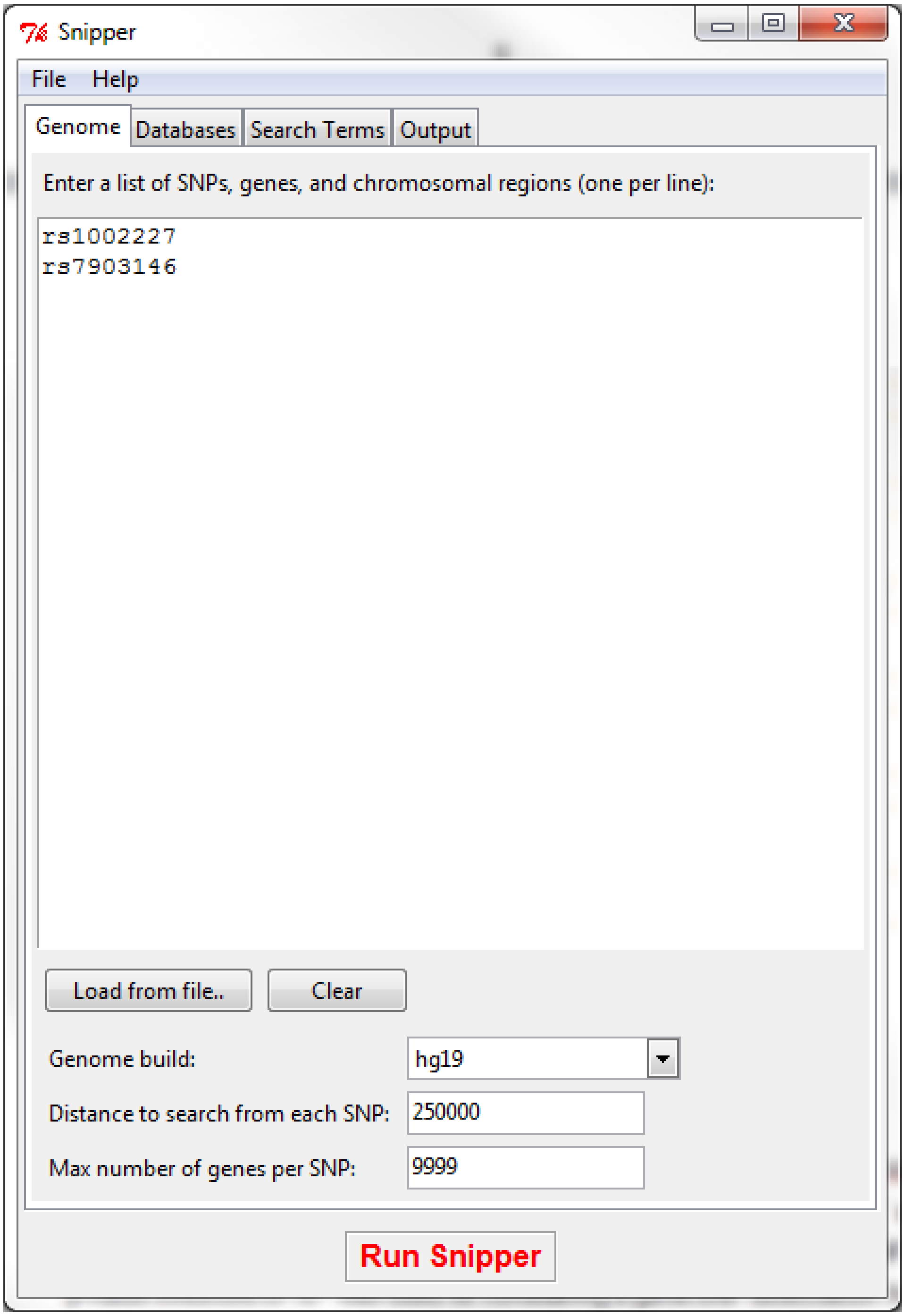
Or, the program can be run with chromosomal regions:

```
snipper --regions “chr#:start-end”
```

All of these can be mixed together, for example:

```
snipper -s "rs7903146,rs1002227" --regions “chr3:12393001-12475854” --gene
"RB1,PDE8B"
```

Figure 2.8. Synopsis of operating Snipper from the command-line.



Enter a list of SNPs, genes, and chromosomal regions (one per line):

```
rs1002227
rs7903146
```

Genome build: hg19

Distance to search from each SNP: 250000

Max number of genes per SNP: 9999

Figure 2.9. Snipper's graphical user interface.

2.2.6 Example

We ran Snipper on a list of 32 SNPs associated with type 2 diabetes (see Voight et al. 2010, Table 3). We included "diabetes," "insulin," and "glucose" as search terms, and extracted all genes within 250 kb of each SNP. A p-value threshold of 10^{-5} was used for considering a gene/SNP association from the eQTL database SCAN as significant. Figure 2.6 shows a truncated version of the “gene report,” a section of the report which contains a table listing all genes located near the SNPs, the search terms that matched information about this gene, and the number of PubMed articles found linked to this gene. The gene report also provides a list of each gene and its annotations and PubMed articles, with search terms highlighted throughout. A link to the full program output can be found in Supplementary Data (section 2.4.)

2.3 Conclusion

We have developed an easy to use command-line tool Snipper for efficient investigation of genes located near, or whose expression levels are associated with, SNPs identified by GWAS. Snipper quickly identifies these genes, downloads information from Entrez Gene, OMIM, PubMed, SCAN, and MiMI, and searches this information with user-specified search terms. A comprehensive report on each gene is provided, as are summary tables to quickly identify genes with entries that match search terms. Through our use of online APIs for each database, the data are obtained at run-time and so guaranteed to be up-to-date each time a user runs Snipper. Finally, because it is command-line driven, Snipper can be easily incorporated into analysis pipelines or run in parallel for multiple projects. We also provide a graphical user interface, which simplifies execution for users unfamiliar with command-line applications. Snipper is open-source and available under the GNU General Public License v3.0.

2.4 Supplementary Data

Snipper program output can be found online at the following link:

http://www.umich.edu/~welchr/snipper_supplemental/

Full program documentation (including a list of all available options) is available at the following link:

http://csg.sph.umich.edu/boehnke/snipper/snipper_docs.pdf

Chapter 3 LocusZoom: regional visualization of genome-wide association scans

3.1 Introduction

Genome-wide association studies (GWAS) have identified hundreds of loci associated with complex human diseases and traits (Manolio et al. 2009). GWAS test for association with dichotomous or quantitative traits at millions of SNPs across the genome and can identify variants many hundreds of kilobases away from any known gene. The next challenge in human genetics will be to identify the causal variants and genes responsible for disease association at the many disease-associated loci identified from GWAS. An associated region may contain only a single strongly associated SNP, or more commonly, a set of SNPs with varying degrees of association due to local linkage disequilibrium (LD) patterns. When examining results from a GWAS, it is important to visually inspect regions showing association to determine the extent of the association signal and the position relative to nearby genes. Genes several hundred kb or more from an associated SNP may be functionally relevant (Loos et al. 2008). We have developed a web-based tool that provides graphical display of locus-specific association results and gives an overview of the extent of LD and the position relative to nearby genes and local recombination hotspots.

3.2 Implementation

3.2.1 Features and functionality

The main panel of a LocusZoom plot shows association P-values on the -log10 scale on the vertical axis, and the chromosomal position along the horizontal axis (Fig. 1). The user can specify the region to display in one of three ways: (i) an index SNP and a window size, (ii) the chromosome together with start and stop positions or (iii) gene name and size of flanking region. We allow for the display of a ‘rug’ above the main panel which gives a tick for any SNP in the results file, or for all SNPs from HapMap Phase II. The plots were designed to display ~1Mb windows of the genome, although for regions with several association signals or long-range LD patterns, plots extending further can be drawn. To identify SNPs that may be potentially causative, LocusZoom plots show not only the magnitude of association for each SNP, but also the pairwise LD pattern with the most strongly associated SNP or another user-specified SNP. Quick inspection can reveal the extent of the associated region and the location and number of SNPs in strong LD with the index SNP. In addition, a locus may show strongly associated variants that are weakly correlated, suggesting the presence of multiple independent association signals. Users may choose to display LD (r^2 or D') estimates from HapMap Phase II (CEU, YRI or JPT + CHB) or from the 1000 Genomes Project. Users of the standalone version of the software may also supply their own pre-calculated LD, or provide genotype files in either MAP/PED or binary PLINK formats from which to calculate LD. LocusZoom caches the r^2/D' estimates to speed-up plot creation for future runs. LocusZoom is compatible with 1000 Genomes SNP naming format (chr:position) and will plot association results for novel SNPs identified by sequencing studies. We provide an option for the data point symbol to reflect genomic annotation (nonsense, non-synonymous, coding, UTR, splice variants, transcription factor binding sites and multi-species conservation), which is available for all SNPs in dbSNP or the 1000 Genomes Project (August 2009 release). The size of the data points can optionally reflect the square root of the sample size. The bottom panel of a LocusZoom plot shows the name and location of genes in the UCSC Genome Browser (Kent et al. 2002). Positions of exons are displayed, and the transcribed strand is indicated with an arrow. This allows the visual comparison of association results relative to coding regions. Gene names are automatically spaced relative to one another to avoid overlap. Currently used plotting tools include regional association plotter SNAP (Johnson et al. 2008) and LD-based viewers such as LD-Plus (Bush et al. 2010), CandiSNPer (Schmitt et al. 2010) and VALID (Jorgenson et al. 2009). LocusZoom provides additional features not currently available in any other single tool, such as: (i) the display of 1000 Genomes or novel SNPs from sequencing studies, (ii) functional annotation of SNPs, (iii) exon/intron distinction and automated gene spacing, (iv) ability to plot regions larger than 500 kb, (v) no pre-selection of input files and (vi) web-based batch mode and availability of source code and databases for download and local installation of LocusZoom.

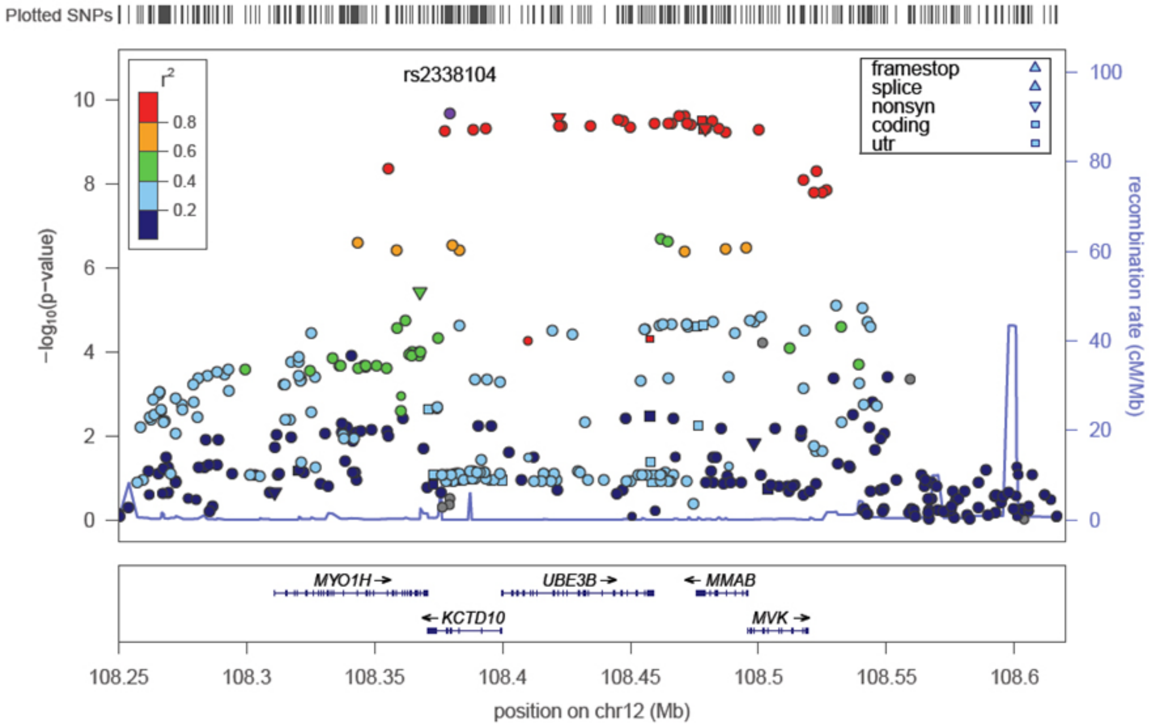


Figure 3.1. Example LocusZoom plot. An example LocusZoom plot showing the HDL cholesterol-associated region near the MMAB gene (Kathiresan et al. 2009).

3.2.2 Usage

LocusZoom was written in R using the grid and lattice graphics packages and runs within a Python wrapper. SQLite tables with relevant data for recombination rate, SNP position, annotation and gene information can be accessed using Python's built-in SQLite tools. A simple plot can be generated from the web form by uploading a file with SNP names and P-values, and specifying the region to be plotted and optional features using drop-down buttons. Typical run time for a single plot returned to the browser window is ~20 s, not including time required to upload a meta-analysis file, which varies according to the user's internet upload speed and file size. To reduce upload time, users may choose to restrict data files to the region being plotted, or to compress the meta-analysis file using gzip. To generate a series of locus plots from the web form, users can submit a specification file where custom specifications for each plot can be listed. When a specification file is used to draw many plots, a single PDF containing all generated plots is returned to the user by e-mail. Finally, users can download our scripts, which require R and Python, and associated databases in SQLite format to enable plot generation on their local Unix machine. Our databases are simple to create, and can be easily adapted to other organisms by following the instructions on our website. Full documentation of all features is available on the LocusZoom website (http://csg.sph.umich.edu/locuszoom/.) The LocusZoom webpage comes pre-loaded with genome-wide association results for HDL cholesterol, LDL cholesterol and triglycerides in ~20,000 individuals of European ancestry (Kathiresan et al. 2009) and a number of other large-scale meta-analyses (Table 3.1).

Trait Consortium	Trait	Reference
Lipids	Total cholesterol	(Teslovich et al. 2010)
	HDL-C	(Teslovich et al. 2010)
	LDL-C	(Teslovich et al. 2010)
	Triglycerides	(Teslovich et al. 2010)
GIANT	BMI	(Speliotes et al. 2010)
	Height	(Lango Allen et al. 2010)
	Waist-hip ratio	(Heid et al. 2010)
MAGIC	2-hr Glucose (adjusted for BMI)	(Saxena et al. 2010)
	Fasting glucose	(Dupuis et al. 2010)
	Fasting insulin	(Dupuis et al. 2010)
	HOMA-B	(Dupuis et al. 2010)
	HOMA-IR	(Dupuis et al. 2010)
	Hemoglobin A1(C)	(Soranzo et al. 2010)
	Fasting proinsulin	(Strawbridge et al. 2011)
ICBP-GWAS	Diastolic blood pressure	(Ehret et al. 2011)
	Systolic blood pressure	(Ehret et al. 2011)

Table 3.1. Pre-loaded datasets. Each dataset is a meta-analysis file provided by a consortium investigating a particular trait. Users can create plots using these datasets from the web.

3.3 Conclusion

We have created a user-friendly tool to generate regional plots of association results in their genomic context. LocusZoom allows for quick visual inspection of the strength of association evidence, the extent of the association signal and LD, and the position of the associated SNPs relative to genes in the region. LocusZoom plots provide an option to size the data points relative to sample size and can display functional annotation. LocusZoom can be accessed from a simple web-based form with drop-down menus or by uploading a specification file to generate many plots at once. LocusZoom Python application, source code in R, and associated databases are available for download and we provide instruction for users to create custom database tables. It is anticipated that, in the future, additional publicly available result sets will be available for convenient viewing.

Chapter 4 Large-scale association study using the Metabochip array reveals new loci influencing glycemic traits

4.1 Introduction

The Meta-Analyses of Glucose and Insulin-related traits Consortium (MAGIC) previously undertook meta-analyses of genome-wide association studies (GWAS) of glycemic traits in non-diabetic individuals, leading to the discovery of multiple robustly associated loci; 16 for fasting glucose concentrations (FG), two for fasting insulin concentrations (FI), and five for post-challenge glucose concentrations (2hGlu) (Prokopenko et al. 2009; Dupuis et al. 2010; Saxena et al. 2010). These and subsequent studies highlighted important biological pathways implicated in glucose and insulin regulation in non-diabetic children and adults (Ingelsson et al. 2010; Barker et al. 2011). They also demonstrated that some, but not all, loci associated with differences in glycemic traits in non-diabetic individuals also affect the risk of type 2 diabetes (T2D) (Dupuis et al. 2010; Voight et al. 2010), and provided insights into the genetic architecture of each trait. Despite the success of these efforts, the identification of new loci was limited by de novo genotyping capacity and cost, such that only a limited number of promising loci from discovery analyses were taken forward to follow-up analyses (often those reaching a threshold of ~$P<10^{-5}$ in discovery). Therefore, it is likely that many additional associations with common, low penetrance variants remain to be found among SNPs not previously selected for replication (Park et al. 2010; Yang et al. 2010a).

The Illumina CardioMetabochip (Metabochip) is a custom Illumina iSELECT array of 196,725 SNPs, developed to support cost-effective large-scale follow-up studies of putative association signals for a range cardiovascular and metabolic traits (~66,000 SNPs) and to fine-map established trait-associated loci (~120,000 SNPs) (Figure 4.6) ((Voight et al. 2012)). The ~66,000 follow-up SNPs were selected to enable genotyping of the top 5,000 or 1,000 most significant association signals for each of 23 metabolic traits contributed by a range of consortia. As such, MAGIC contributed ~5,000 top ranking SNPs for FG, and ~1,000 each for FI and 2hGlu that had shown nominal association in discovery analyses ($P_{discovery}<0.02$) (Dupuis et al. 2010; Saxena et al. 2010) as well as 16,202 SNPs to fine-map previously established loci.

In the present study, we combined newly available samples with genotype data for these 66,000 follow-up SNPs with previous discovery meta-analyses to discover novel association signals with glycemic traits. This approach identified 38 glycemic associations not described in the previous discovery approaches in Europeans (Saxena et al. 2010): 20 for FG, 17 for FI and four for 2hGlu (in two loci, we observed associations with more than one trait). This takes the total number of loci associated with glycemic traits to 36 for FG, 19 for FI and 9 for 2hGlu explaining 4.8%, 1.2% and 1.7% of the variance in these traits, respectively. Of these 53 non-overlapping loci, 33 were also associated with T2D (P<0.05), which whilst supporting the previous assertion of an imperfect correlation between these traits, also implicates new T2D loci and increases the overlap between glycemic and T2D loci.

4.2 Methods

4.2.1 Study design

The Illumina CardioMetabochip (Metabochip) is a custom Illumina iSELECT array of 196,725 SNPs. It has been designed to support efficient large-scale follow-up of putative associations for glycemic (including fasting glucose (FG), fasting insulin (FI) and post-challenge glucose concentrations (2hGlu)) and other metabolic and cardiovascular traits (Figure 4.6) ((Voight et al. 2012)), and to enable the fine-mapping of established loci. Overall, there were 65,435 SNPs genotyped on the Metabochip for follow-up of previous associations including a total of 23 cardio-metabolic traits. Traits contributing SNPs to the Metabochip were prioritized into “primary” (including FG) and “secondary” (including FI and 2hGlu) contributing ~5k and ~1k SNPs, respectively, from the most significantly associated variants for each phenotype in the discovery meta-analyses from each contributing consortium. This included 5,055 SNPs for follow-up of

FG, 1,046 for FI, and 1,038 for follow up of 2hGlu associations. In the present analysis we focused our analysis on this set of “follow-up” SNPs available on the Metabochip to establish variants amongst these SNP associated with glycemic traits. While we also included newly available studies genotyped on genome-wide platforms, we limited our primary analyses to only these ~66,000 SNPs.

4.2.2 Studies

In the present effort, collaborating studies within the Meta-Analysis of Glucose and Insulin related traits Consortium (MAGIC) provided results for the 66,000 “follow-up” SNPs genotyped on Metabochip on a maximum total of 133,010 (FG)/ 108,557(FI)/ 42,854(2hGlu) individuals. As well as those newly genotyped on the Metabochip platform, in our overall meta-analysis we were able to include further studies which had genotyped or imputed the same SNPs on other platforms. The largest proportion of our entire sample was directly genotyped on the Metabochip and comprised 53,622 (FG)/ 42,384 (FI)/ 27,602 (2hGlu) individuals from 26/21/12 studies, respectively. We were also able to recruit 11,690 (FG)/8,813 (FI) individuals from up to 4 additional GWA studies (Prevend, Ascot (FG-only), Prosper, and TRAILS) (Table ST1) not included in the original meta-analysis (Dupuis et al. 2010). From another MAGIC study of sex-specific associations with glycemic traits (Prokopenko on behalf of the MAGIC authors, personal communication), we were able to recruit another 15/13 independent studies comprising up to 25,618/23,130 individuals for FG and FI, respectively. The above studies were combined in a single fixed-effects meta-analysis with those studies included in the 2 original GWAS (Dupuis et al. 2010; Saxena et al. 2010): 20 (FG)/ 19 (FI)/ 9 (2hGlu) studies and 42,080 (FG)/34,230 (FI)/15,252 (2hGlu) individuals, as described previously (Dupuis et al. 2010; Saxena et al. 2010). The study and individual counts from the original GWAS excluded the family-based SardiNIA study where, initially, a large number of the individuals had imputed genotype data only. The entire sample was directly genotyped on Metabochip, so those data were included in place of the original GWAS. Some studies had genotyping data available from both Metabochip and genome-wide arrays but from entirely independent samples within the studies (Table ST1). Full study characteristics of all Metabochip studies are shown in Table

ST1, while data from discovery genome-wide studies and those from the sex-specific analyses are reported elsewhere (Dupuis et al. 2010; Saxena et al. 2010); Prokopenko on behalf of the MAGIC authors, personal communication). All participants of the main analysis were of European descent and mostly adults, although data from a total of 7,872/7,164 adolescents were also included in the FG/FI meta-analyses (NFBC86, Leipzig-childhood_IFB, TRAILS and ALSPAC studies). All studies were approved by local research ethic committees and all participants gave informed consent. Results from the CLHNS study of Filipino women (N = 1,682/1,635 for FG/FI, respectively) genotyped on Metabochip were also available and were included in supplementary analyses to compare effect directions with European-descent studies alone.

4.2.3 Phenotypes

Analyses were undertaken for FG and FI measured in mmol/l and pmol/l, respectively. 2hGlu was measured in mmol/l. Similar to the previous MAGIC discovery analysis (Dupuis et al. 2010; Saxena et al. 2010), individuals were excluded from the analysis if they had a physician diagnosis of diabetes, were on diabetes treatment (oral or insulin), or had a fasting plasma glucose equal to or greater than 7 mmol/l. Individual studies applied further sample exclusions, including pregnancy, non-fasting individuals and type 1 diabetes, as detailed in Table ST1. Individuals from case control studies (Table ST1) were excluded if they had hospitalization or blood transfusion in the 2-3 months before phenotyping took place. 2hGlu measures were done 120 min after a glucose challenge during an oral glucose tolerance test (OGTT). Measures of FG and 2Glu made in whole blood were corrected to plasma level using the correction factor of 1.133. FI was measured in serum. Detailed descriptions of study-specific glycemic measurements are given in the Table ST1.

4.2.4 Trait transformations and adjustment

Analyses were performed for untransformed levels of FG, natural logarithm transformed FI and untransformed 2hGlu using a linear regression model. All analyses were adjusted for age (if applicable), study site (if applicable) and geographical covariates (if applicable) to evaluate the association using an additive genetic model at each genetic SNP variant.

4.2.4.1 BMI-adjusted analysis

In the Fenland study (Table ST1), we investigated the correlation between BMI and natural logarithm-transformed FI, FG, and 2hGlu to establish the variation in each trait explained by BMI. Meta-analyses for each trait were also adjusted for body mass index (BMI). Metabochip and new GWA studies performed study-level analyses adjusted for BMI. Most studies in the original GWAS (except deCode, GEMs, KORAF4, TwinsUK studies) as well as from the studies analyzed in a sex-specific manner were included in BMI-adjusted meta-analysis. The original discovery 2hGlu meta-analysis adjusted for BMI (Saxena et al. 2010) was also included in these analyses. We also performed an analysis for 2hGlu adjusted for FG to investigate if additional variants would be identified with an effect on 2hGlu independent of FG and also to establish whether identified 2hGlu associations were driven by FG.

4.2.5 Genotyping and quality control

The Metabochip or other commercial genome-wide arrays were used by individual studies for genotyping. Details are presented in Table ST1, or are reported elsewhere (Dupuis et al. 2010; Saxena et al. 2010). The quality control criteria for both Metabochip and genome-wide arrays for filtering of poorly genotyped individuals or low quality SNPs prior to imputation included: (1) call rate<0.95; (2) sex-discrepancies; (3) ethnic outliers; (4) heterozygosity (Table ST1); (5) SNP minor allele frequency<0.01; (6) SNP Hardy-Weinberg equilibrium P<10-4; (7) SNP effect estimate standard error (SE) =10; (8) SNPs minor allele count (MAC) < 10 (calculated as total number of observed alleles at each SNP multiplied by MAF).

Studies with genome-wide arrays undertook imputation using the HapMap CEU reference panel using MACH and IMPUTE software (Table ST1). Parameters used in imputation and filters applied to imputed genotypes are described in Table ST1 or reported previously (Saxena et al. 2010). From a total of ~2.5M genome-wide directly genotyped or imputed autosomal SNPs, study-specific results for the ~66,000

Metabochip follow-up SNPs were considered for the present meta-analyses. SNPs with a meta-analysis result for more than a total 10,000 individuals were included in the analysis.

4.2.6 Statistical analysis

Analyses of previous discovery studies are reported elsewhere (Dupuis et al. 2010; Saxena et al. 2010), while those studies genotyped on the Metabochip are described in Table ST1. SNP effect estimates and their standard errors (for additive genetic model) were combined by inverse-variance weighted fixed effects meta-analysis using METAL (Willer et al. 2010) and GWAMA (Magi and Morris 2010). Two parallel meta-analyses for each trait by different analysts were compared for consistency. Individual cohort results were corrected for residual inflation of the test statistics using lambda of genomic control (GC) estimates. The GC values were estimated for each study using either test statistics from all SNPs for the GWA studies, while for those studies genotyped on the Metabochip, GC lambda estimates were derived from test statistics for 5,041 SNPs selected for follow-up of QT-interval associations, as we perceived these to have the lowest likelihood of common architecture of associations with glycemic traits. Individual study-level lambda GC estimates are shown in Table ST1. Overall QQ plots for the QT follow-up SNPs are shown in Figure 4.27 - Figure 4.30.

4.2.7 Fasting glucose, fasting insulin, and 2-hr glucose associated signals selection strategy

Meta-analysis results for each trait were considered as genome-wide significant if they achieved P<=5x10-8 threshold and were not in LD (r^2<0.05) or within 500Kb of an established signal. The most significantly associated SNP (lowest P-value) in each region (500Kb) was selected as the lead SNP. Associated loci are referred to by the name of the nearest gene, unless a more biologically plausible gene was nearby, or a nearby gene was previously associated with another trait. In such cases, we maintain consistency with the previous naming, but list the nearest genes in Tables ST2 (b-e.) To establish the variance in each trait explained by these SNPS, in the Framingham Heart Study, we included all SNPs in a model adjusted for age, sex, BMI and cohort.

4.2.8 Fine-mapping of known glycemic trait loci

To undertake preliminary fine-mapping analyses, we investigated the patterns of association at 17 known FG/FI loci (Dupuis et al. 2010) and 5 known 2hGlu loci (Saxena et al. 2010) using meta-analysis results from 13,644/1,309/1,249 SNPs genotyped on the Metabochip in 53,622, 42,384, and 27,602 individuals for FG, FI and 2hGlu, respectively. Only studies genotyped directly on the Metabochip were used for fine-mapping purposes in order to have equal sample size and availability of all SNPs. Regional plots for each locus were created using the previous lead SNP (Dupuis et al. 2010) or a suitable proxy (r^2>0.8) as the index SNP if that marker was not present on Metabochip. The plots were generated on the LocusZoom web-based plotting software (Pruim et al. 2010) utilizing LD information from 1000Genomes (hg19/Nov2010/EUR data). Prior to generating the plots, all SNP names and positions from the Metabochip-only meta-analysis files were aligned to build37 using the Lift Genome annotation tool on the UCSC website (http://genome.ucsc.edu/cgibin/ hgLiftOver) in order to be compatible with the 1000 Genomes SNP naming format (chr:position) and allow more thorough assessment of the pairwise LD patterns around the established SNPs.

4.2.9 Associations of glycemic trait variants with related traits

For those SNPs which we identified to be genome-wide significant, we also investigated their association with other metabolic and disease traits. We exchanged reciprocal data for such SNPs with the latest DIAGRAM Metabochip analyses ((Morris et al. 2012)), and checked associations of these SNPs in publicly available data from previous studies of lipid traits from the GLGC7 (Triglycerides, HDL- and LDL-cholesterol - http://www.sph.umich.edu/csg/abecasis/public/lipids2010/) as well as BMI and waist-hip ratio (WHR) from the GIANT consortium (Heid et al. 2010; Speliotes et al. 2010) (http://www.broadinstitute.org/collaboration/giant/index.php/Main_Page). From these data, we were able to establish the presence of any association and the direction of effect for these other traits aligned to our trait-raising alleles. We highlighted associations with other traits at P<0.05, and also performed FDR analyses. We performed FDR analyses for each trait separately (removing duplicate loci that were associated with more than one glycaemic trait) and identified those where q<0.05.

4.2.10 SNP/gene biology and functional annotation

4.2.10.1 eQTL (ASAP Liver)

Liver gene expression data from the Advanced Study of Aortic Pathology (ASAP) study has been described previously (Folkersen et al. 2010). In brief, liver biopsies were collected from patients at the Karolinska University Hospital, Stockholm, Sweden undergoing aortic valve surgery alone or combined with surgery for aortic aneurysm, starting from February 13, 2007. All subjects gave their informed consent and the study was approved by the ethics committee of Karolinska Institute, Stockholm, Sweden. After hybridisation of extracted RNA to Affymetrix ST 1.0 Exon arrays, data was RMA normalized and log-transformed. DNA was extracted from whole blood and genotyping was carried out using the Illumina 610w-Quad bead array platform. Imputation was carried out on SNPs with a call rate exceeding 95%, using the MACH algorithm. Imputation quality scores of RSQ < 0.3 were excluded from analysis. An additive genetic model was used to test for association between SNPs and gene expression.

4.2.10.2 VEGAS

To identify genes with multiple associated SNPs we performed gene-based analysis using VEGAS, described in detail previously (Liu et al. 2010). Briefly, on all available samples and among the ~66,000 follow-up SNPs, VEGAS pooled the information for all SNPs within each gene (± 50kb) to identify genes with higher evidence of association than expected by chance, while adjusting for gene size and the linkage disequilibrium structure of the SNPs, by simulation (maximum number of simulations used was 106). We identified genomic regions (separated by >1Mb) showing evidence of association and described the genes contained within those regions. While we often identified multiple genes within an associated region, it is probable that some of these are significant via linkage disequilibrium. Bonferroni correction was used to adjust for multiple testing, based on the number of independent tests (number of genes tested) (~9,300) and P-values < 5.0x10-6 were considered significant. While the number of genes represented was constrained by those SNPs submitted to the Metabochip, our analyses asked the question: of the genes represented on the Metabochip, all with a slightly raised prior likelihood of association, which genes show the most evidence for association with glycemic traits?

4.2.10.3 GRAIL (Gene Relationships Among Implicated Loci)

We used the GRAIL (Raychaudhuri et al. 2009a) to evaluate whether genome-wide loci associated with glycemic traits were enriched for connectivity between genes representing particular pathways or molecular processes. As described in detail previously (Raychaudhuri et al. 2009a), to define the genes near each SNP, GRAIL finds the furthest neighboring SNPs in the 3' and 5' direction in LD (Hapmap CEU: $r^2 > 0.5$) and proceeds outwards in each direction to the nearest recombination hotspot14. All genes that overlap that interval are considered implicated by the SNP. If there are no genes in that region, the interval is extended by 250 kb in either direction. The method performs a text-based analysis looking at abstracts in PubMed prior to Dec 2006 (to avoid confounding from GWAS results arising after that date). We performed two analyses for each trait: first, we took all genome-wide signals for each trait as seed and query loci to investigate biological connectivity amongst those loci (FG=35, FI=16, 2hGlu=9). For FI, we did not include FTO as the association with FI was entirely mediated by BMI. Secondly, we also investigated connectivity between these established signals (as seed regions) and those which did not reach genome-wide significance but were suggestively associated with each trait (P<0.0005) (as query regions) as described previously (Raychaudhuri et al. 2009b). For FI, we used BMI-adjusted results to define the query regions. Query regions were defined by taking all SNPs more significant than P<0.0005, removing those associated at genome-wide levels of significance and pruning SNPs of $r^2>0.05$ in each region using PLINK (Purcell et al. 2007). As GRAIL tests connectivity of regions, we also removed any duplicates where a region was represented by more than one SNP. For those SNPs not found by the software, we submitted the region as a 500Kb window centered at the location of the SNP. This approach identified 218, 155, and 100 query regions (representing 715, 639 and 298 genes) for FG, FI (adjusted for BMI), and 2hGlu, respectively. The number of loci reaching $P_{grail}<0.01$ was identified from these analyses and to establish the level of enrichment, we randomly sampled 1,000 random sets of matched numbers of SNPs and calculated the proportion with as many or more reaching $P_{grail}<0.01$ to derive a permutation based P-value ($P_{permutation}$).

4.2.10.4 Pathway analyses

Pathway analysis was carried out for FG, FI and 2hGlu (uniform or adjusted for fasting glucose or adjusted for BMI) using data from previous discovery GWAS only (Dupuis et al. 2010) to avoid bias towards pathways represented on the Metabochip (Build 36, N>10,000 and MAF≥1% cutoff used). The software used for this analysis was MAGENTA 2.4 (July 2011, http://www.broadinstitute.org/mpg/magenta/). SNPs from the meta-analysis file were assigned to a gene if they mapped within 110kb upstream and 40kb downstream of transcript boundaries. The smallest P-value for the set of SNPs assigned to the gene was adjusted for confounders, such as gene length, marker density, LD in a linear regression, creating a gene association score. If a top SNP was assigned to multiple genes, only the gene with the lowest score was kept to avoid positional clustering. The HLA region was removed due to high LD and gene density. Pathway terms from multiple databases (GO, PANTHER, Ingenuity, KEGG) was attached to each gene. The genes were ranked on their association score and a GSEA test was performed testing all pathway terms using a 5% and 75% cutoff. Initially, 10,000 gene set permutations were performed for GSEA P-value estimation. This number was then increased with GSEA P-value<$1x10^{-4}$, and up to 1,000,000 permutations were performed. Results were sorted on FDR (5% cutoff) and FDR<0.05 was considered to be significant.

4.2.10.5 Analyses of directional consistency of associations between discovery and follow-up studies among SNPs not reaching genome-wide significance

We investigated whether the Metabochip follow-up SNPs were likely to contain further “real” associations in addition to those SNPs which reached genome-wide significance. To do so, we meta-analyzed those studies involved in the original discovery analyses (Dupuis et al. 2010; Saxena et al. 2010) comprising 42,078 individuals for FG, 34,230 for FI and 15,252 for 2hGlu, and also then separately meta-analyzed all studies newly available to follow up, comprising 85,710 individuals for FG, 69,240 for FI and 27,602 for 2hGlu. For each trait (FG, FI, FI-BMIadj, and 2hGlu), we then identified all SNPs which had a nominally significant association ($P<0.05$) in the follow-up studies alone and, for these SNPs, performed a two-sided binomial test of whether more SNPs than expected by chance (50%) had a consistent direction of effect with that observed in the discovery analyses. Before performing these analyses, SNPs were filtered by LD ($r^2<0.01$) to identify independent variants, and all SNPs (and those in LD, $r^2=0.01$) associated with glycemic traits (FG, FI, 2hGlu, HbA1c and proinsulin) at genome-wide levels of significance (including those SNPs identified in the present study) were excluded. These analyses were initially performed for all 66,000 SNPs, but we were then also able to compare across SNPs submitted to the Metabochip by different consortia and for SNPs submitted to follow up on particular traits amongst these consortia. The results of each of these tests were plotted overall, within SNPs from each consortium, and within SNPs submitted for follow-up of each trait in Figure 4.22- Figure 4.25. The numbers of SNPs meeting these criteria are shown are Table ST7. We supplemented these results with FDR analyses and noted the q-value at $P=0.05$ in the follow-up 8 studies to identify the likelihood of true positives amongst these nominally significant SNPs (Table ST7).

4.3 Results

4.3.1 Approaches to identify loci associated with glycemic traits

To follow-up loci showing evidence of association ($P_{\text{discovery}}$ one<0.02) with glycemic traits in discovery GWAS, we investigated the 66,000 Metabochip follow-up SNPs for association with FG, FI and 2hGlu. We combined in meta-analyses data from up to 133,010 (FG) /108,557 (FI) /42,854 (2hGlu) non-diabetic individuals of European ancestry, including individuals from the previous GWAS meta-analyses (Dupuis et al. 2010; Saxena et al. 2010), individuals from new GWAS and individuals newly genotyped on the Metabochip array (Figure 4.1). Genome-wide association data for Filipino women were available to us (Table ST1), for which we report the effect directions relative to the overall effects in Tables ST2 (b-c.) All study characteristics are shown in Table ST1. We considered SNPs to represent novel association signals if they were genome-wide significant ($P<5\times10^{-8}$) and located more than 500Kb from, and not in

LD (Hapmap CEU: $r^2 < 0.05$) with, any variant already known to be associated with the trait. Associated loci are referred to by the name of the nearest gene, unless a more biologically plausible gene was nearby, or a nearby gene was previously associated with another trait. In such cases, we maintain consistency with the previous naming, but list the nearest genes in Tables ST2 (b-e.) As BMI is a major risk factor for T2D and is correlated with glycemic traits, we also performed analyses adjusted for BMI.

Figure 1

MAGIC Metabochip Study Design

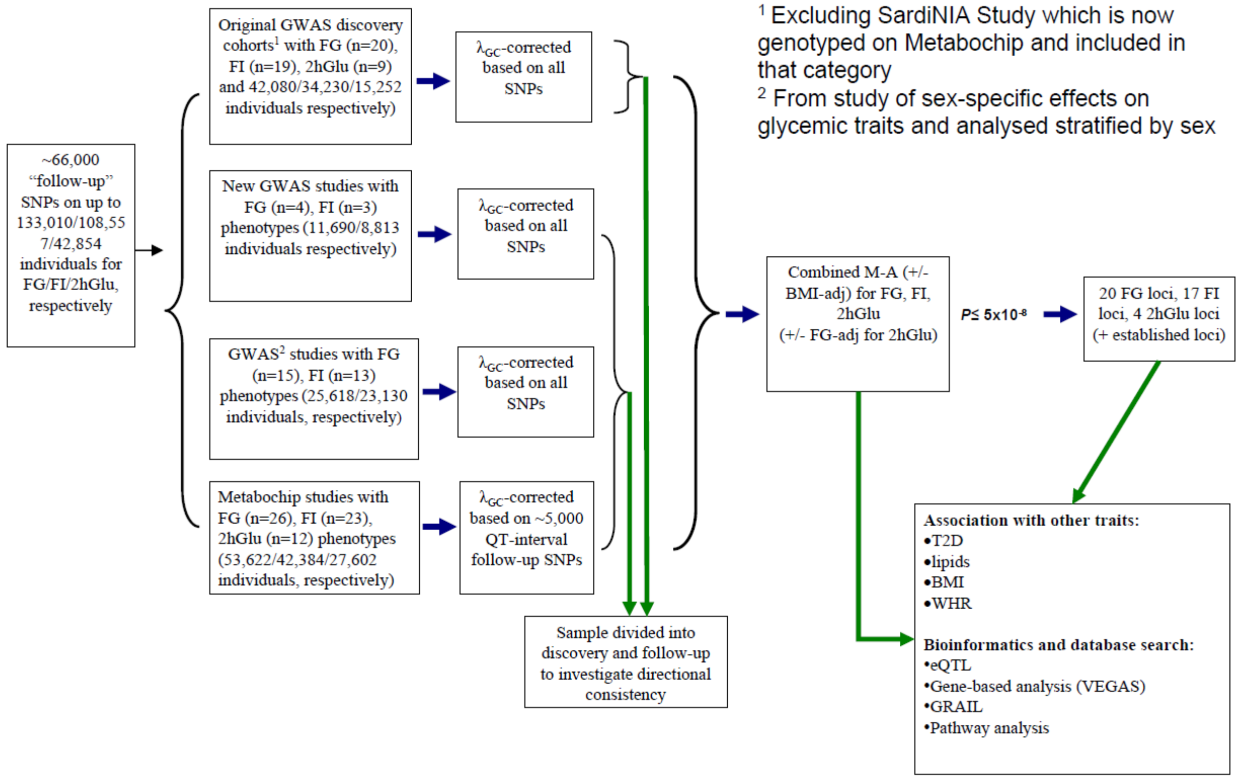

~66,000 "follow-up" SNPs on up to 133,010/108,557/42,854 individuals for FG/FI/2hGlu, respectively
Original GWAS discovery cohorts[1] with FG (n=20), FI (n=19), 2hGlu (n=9) and 42,080/34,230/15,252 individuals respectively)
New GWAS studies with FG (n=4), FI (n=3) phenotypes (11,690/8,813 individuals respectively)
GWAS[2] studies with FG (n=15), FI (n=13) phenotypes (25,618/23,130 individuals, respectively)
Metabochip studies with FG (n=26), FI (n=23), 2hGlu (n=12) phenotypes (53,622/42,384/27,602 individuals, respectively)
λ_{GC}-corrected based on all SNPs
λ_{GC}-corrected based on all SNPs
λ_{GC}-corrected based on all SNPs
λ_{GC}-corrected based on ~5,000 QT-interval follow-up SNPs
Sample divided into discovery and follow-up to investigate directional consistency
[1] Excluding SardiNIA Study which is now genotyped on Metabochip and included in that category
[2] From study of sex-specific effects on glycemic traits and analysed stratified by sex
Combined M-A (+/- BMI-adj) for FG, FI, 2hGlu (+/- FG-adj for 2hGlu)
P≤ 5x10-8
20 FG loci, 17 FI loci, 4 2hGlu loci (+ established loci)
Association with other traits:
•T2D
•lipids
•BMI
•WHR
Bioinformatics and database search:
•eQTL
•Gene-based analysis (VEGAS)
•GRAIL
•Pathway analysis

Figure 4.1. MAGIC Metabochip study design. The overall design for the follow-up of ~ 66,000 SNPs is shown above.

Though not the main focus of this effort, given the increased variant density available on the Metabochip for previously established glycemic loci, we investigated whether these data would enable fine-mapping of the functional variants potentially underlying these signals (Dupuis et al. 2010; Saxena et al. 2010). In these analyses we included data from up to 53,622 individuals for FG, 42,384 for FI, and 27,602 for 2hGlu from studies with Metabochip genotypes only. However, given the lack of samples from different ancestries and the absence of full conditional analysis, for the most part these analyses did not improve the resolution of association signals.

Beyond individual SNP investigations for each glycemic trait, we also tested the hypothesis that gene-based analyses using VEGAS (Liu et al. 2010) would identify genes that harbor multiple association signals, which individually did not reach genome-wide significance. Among the ~66,000 SNPs we used VEGAS to pool the information for all SNPs within each gene (± 50kb) to identify genes with more evidence of association than expected by chance (given the gene size and linkage disequilibrium structure) by simulation. Genes were considered to be significantly associated if they were significant after Bonferroni-correction for multiple testing ($P<5x10^{-6}$).

Below we provide details of these analyses for fasting glucose, fasting insulin and post-challenge glucose levels.

4.3.1.1 Fasting glucose

In analyses of up to 133,010 individuals we identified 20 loci with genome-wide significant associations ($P<5\times10^{-8}$) (Table 4.1, Figure 4.7, Figure 4.11) and confirmed previously established loci (Table ST2a). Of these 20 loci, nine (in or near IBKAP, LOC728489, WARS, KL, TOP1, P2RX2, AMT, RREB1 and GLS2) had not previously been associated with other metabolic traits (Table 4.2). Among these, KL (Klotho) is of particular interest. In addition to being associated with FG (but not FI), the FG-raising allele is also associated with an increased risk of T2D (OR=1.08(1.04-1.11), $P=1.1x10^{-5}$) (Figure 4.2). KL was first identified as a gene related to suppression of aging: its reduced expression was associated with reduced lifespan, as well as hypoglycemia (Kuro-o et al. 1997). Despite further animal studies supporting a role for KL in glucose metabolism (Ohnishi et al. 2011) and insulin sensitivity (Utsugi et al. 2000), human studies were generally small and inconclusive (Rhee et al. 2006; Paroni et al. 2012).

We also identified new associations with FG in regions previously associated with other metabolic traits or disease outcomes including T2D6 (ARAP1, CDKN2B, GRB1015, CDKAL1, IGF2BP2 and ZBED3, which was identified in BMI-adjusted models) and 2hGlu2 (GIPR) as well as confirming the recently identified signals for FG at FOXA216, PPP1R3B, PCSK1 and PDX1 ((Manning et al. 2012)). FOXA2 is a forkhead transcription factor that regulates PDX1 expression, while PDX1 encodes a transcription factor critical to pancreatic development (Jonsson et al. 1994). PDX1 mutations have been linked to MODY4 (Stoffers et al. 1997a), pancreatic agenesis (Stoffers et al. 1997b) and permanent neonatal diabetes (Nicolino et al. 2010); however, we did not observe significant association with T2D based on associations in DIAGRAM Metabochip analyses (Morris et al. 2012) (Figure 4.2).

							Primary trait				I^2		FI (BMI-adjusted)				2hGlu			
Primary trait	SNP	Chr.	Position	Gene	Alleles (effect/ other)	Freq. effect allele	Effect	SE	Global analysis *P* value	Global analysis *n*	estimate	*P* value	Effect	SE	Global analysis *P* value	Global analysis *n*	Effect	SE	Global analysis *P* value	Global analysis *n*
FG	**rs10811661**	9	22124094	CDKN2B	T/C	0.82	0.0238	0.003	5.6×10^{-18}	128,488	0.00	1.00	–0.0065	0.003	0.019	98,880	0.0567	0.014	8.8×10^{-5}	42,801
	rs4869272	5	95565204	PCSK1*	T/C	0.69	0.0177	0.002	1.0×10^{-15}	131,872	0.00	1.00	0.0016	0.002	0.469	103,493	–0.0322	0.012	0.006	42,848
	rs11619319	13	27385599	PDX1	G/A	0.23	0.0195	0.002	1.3×10^{-15}	132,996	0.00	1.00	0.0001	0.002	0.977	103,492	0.0185	0.013	0.156	42,848
	rs983309	8	9215142	PPP1R3B*	T/G	0.12	0.0256	0.003	6.3×10^{-15}	127,470	0.14	0.32	0.0223	0.003	1.2×10^{-12}	99,024	–0.0548	0.016	0.001	42,846
	rs6943153	7	50759073	GRB10	T/C	0.34	0.0154	0.002	1.6×10^{-12}	131,795	0.00	1.00	0.0091	0.002	2.3×10^{-5}	103,447	0.0110	0.011	0.333	42,794
	rs11603334	11	72110633	ARAP1	G/A	0.83	0.0192	0.003	1.1×10^{-11}	128,139	0.00	1.00	–0.0046	0.003	0.086	99,026	0.0294	0.014	0.037	42,839
	rs6113722	20	22505099	FOXA2	G/A	0.96	0.0353	0.005	2.5×10^{-11}	123,665	0.04	0.78	–0.0095	0.005	0.064	103,471	0.0493	0.030	0.101	41,416
	rs16913693	9	110720180	IKBKAP	T/G	0.97	0.0434	0.007	3.5×10^{-11}	125,115	0.00	1.00	–0.0018	0.007	0.785	96,357	0.0639	0.034	0.062	40,522
	rs3829109	9	138376587	DNLZ	G/A	0.71	0.0172	0.003	1.1×10^{-10}	115,310	0.25	0.07	–0.0002	0.003	0.948	94,964	0.0343	0.014	0.013	36,803
	rs3783347	14	99909014	WARS	G/T	0.79	0.0168	0.003	1.3×10^{-10}	132,544	0.02	0.89	0.0017	0.003	0.515	103,339	0.0274	0.014	0.044	42,850
	rs2302593	19	50888474	GIPR	C/G	0.50	0.0144	0.002	9.3×10^{-10}	116,141	0.27	0.05	0.0025	0.002	0.265	96,976	–0.0322	0.012	0.006	40,781
	rs9368222	6	20794975	CDKAL1	A/C	0.28	0.0143	0.002	1.0×10^{-9}	128,453	0.09	0.50	–0.0047	0.002	0.037	98,894	0.0279	0.012	0.023	42,825
	rs10747083	12	131551691	P2RX2	A/G	0.66	0.0133	0.002	7.6×10^{-9}	127,111	0.00	1.00	–0.0006	0.002	0.785	99,895	0.0269	0.012	0.026	42,790
	rs6072275	20	39177319	TOP1	A/G	0.16	0.0159	0.003	1.7×10^{-8}	128,616	0.00	1.00	0.0038	0.003	0.169	99,018	–0.0110	0.014	0.435	42,853
	rs7651090	3	186996086	IGF2BP2	G/A	0.31	0.0128	0.002	1.75×10^{-8}	128,548	0.02	0.86	0.0003	0.002	0.900	98,924	0.0583	0.012	1.05×10^{-6}	42,814
	rs576674	13	32452302	KL	G/A	0.15	0.0167	0.003	2.3×10^{-8}	131,856	0.00	1.00	–0.0001	0.003	0.983	103,472	0.0308	0.016	0.060	42,849
	rs11715915	3	49430334	AMT	C/T	0.68	0.0120	0.002	4.9×10^{-8}	131,523	0.30	0.02	0.0059	0.002	0.006	103,398	0.0273	0.012	0.018	42,851
FG (BMI-adjusted)	rs17762454	6	7158199	RREB1	T/C	0.26	0.0140	0.002	9.6×10^{-9}	123,247	0.00	1.00	–0.0002	0.002	0.919	103,470	0.0007	0.013	0.953	42,848
	rs7708285	5	76461623	ZBED3	G/A	0.27	0.0150	0.003	1.2×10^{-8}	117,931	0.00	1.00	0.0027	0.002	0.265	98,341	0.0349	0.013	0.008	42,803
	rs2657879	12	55151605	GLS2	G/A	0.18	0.0157	0.003	3.9×10^{-8}	121,596	0.39	0.03	–0.0024	0.003	0.366	102,175	0.0200	0.014	0.164	42,670
							Primary trait						FG				2hGlu			
FI	rs1421085	16	52358455	FTO	C/T	0.42	0.0200	0.003	1.9×10^{-15}	104,062	0.00	1.00	0.0074	0.002	0.001	128,597	0.0122	0.011	0.278	42,849
	rs983309	8	9215142	PPP1R3B*	T/G	0.12	0.0287	0.004	3.8×10^{-14}	103,030	0.04	0.77	0.0256	0.003	6.3×10^{-15}	127,470	–0.0548	0.016	0.001	42,846
	rs9884482	4	106301085	TET2	C/T	0.39	0.0165	0.002	1.4×10^{-11}	108,420	0.00	1.00	0.0001	0.002	0.946	132,869	0.0004	0.011	0.973	42,745
	rs7903146	10	114748339	TCF7L2	C/T	0.72	0.0181	0.003	6.1×10^{-11}	103,037	0.31	0.02	–0.0220	0.002	2.7×10^{-20}	127,477	–0.0885	0.013	5.6×10^{-12}	42,851
	rs10195252	2	165221337	GRB14*	T/C	0.59	0.0159	0.003	4.9×10^{-10}	99,126	0.00	1.00	0.0053	0.002	0.014	127,005	0.0361	0.011	0.001	42,846
	rs1167800	7	75014132	HIP1	A/G	0.54	0.0156	0.003	2.6×10^{-9}	91,416	0.00	1.00	0.0016	0.002	0.470	118,536	–0.0133	0.012	0.272	38,884
	rs2820436	1	217707303	LYPLAL1	C/A	0.67	0.0153	0.003	4.4×10^{-9}	104,044	0.01	0.97	0.0077	0.002	0.001	128,580	–0.0041	0.012	0.723	42,843
	rs2745353	6	127494628	RSPO3	T/C	0.51	0.0143	0.002	5.5×10^{-9}	104,075	0.06	0.67	–0.0009	0.002	0.677	128,615	–0.0005	0.011	0.962	42,853
	rs731839	19	38590905	PEPD	G/A	0.34	0.0145	0.003	1.7×10^{-8}	104,636	0.13	0.38	0.0046	0.002	0.038	132,528	0.0142	0.012	0.220	42,846
	rs4865796	5	53308421	ARL15	A/G	0.67	0.0146	0.003	2.1×10^{-8}	100,001	0.03	0.81	0.0043	0.002	0.052	127,784	0.0337	0.012	0.004	42,852
	rs2972143	2	226824609	IRS1	G/A	0.62	0.0142	0.003	3.2×10^{-8}	99,566	0.00	1.00	0.0035	0.002	0.107	127,473	0.0195	0.011	0.082	42,853
	rs1530559	2	135472099	YSK4	A/G	0.52	0.0145	0.003	3.4×10^{-8}	107,281	0.19	0.18	0.0037	0.002	0.100	129,880	0.0200	0.011	0.077	42,849
FI (BMI-adjusted)	rs2943645	2	226807424	IRS1	T/C	0.63	0.0193	0.002	2.3×10^{-19}	99,023	0.00	1.00	0.0034	0.002	0.112	127475	0.0210	0.011	0.061	42,846
	rs10195252	2	165221337	GRB14*	T/C	0.60	0.0174	0.002	1.3×10^{-16}	98,997	0.00	1.00	0.0053	0.002	0.014	127005	0.0361	0.011	0.001	42,846
	rs2126259	8	9222556	PPP1R3B	T/C	0.11	0.0238	0.003	3.3×10^{-13}	99,021	0.14	0.51	0.0213	0.003	5.4×10^{-10}	127480	–0.0877	0.017	1.8×10^{-7}	42,849
	rs4865796	5	53308421	ARL15	A/G	0.67	0.0154	0.002	2.2×10^{-12}	98,314	0.48	0.01	0.0043	0.002	0.052	127784	0.0337	0.012	0.004	42,852
	rs17036328	3	12365484	PPARG	T/C	0.86	0.0212	0.003	3.6×10^{-12}	98,984	0.21	0.31	0.0051	0.003	0.103	128567	0.0335	0.016	0.031	42,843
	rs731839	19	38590905	PEPD	G/A	0.34	0.0148	0.002	5.1×10^{-12}	103,252	0.13	0.55	0.0046	0.002	0.038	132528	0.0142	0.012	0.220	42,847
	rs974801	4	106290513	TET2	G/A	0.38	0.0139	0.002	3.3×10^{-11}	103,489	0.09	0.67	0.0012	0.002	0.582	131866	0.0052	0.011	0.643	42,849
	rs459193	5	55842508	ANKRD55-MAP3K1	G/A	0.73	0.0147	0.002	1.12×10^{-10}	103,378	0.27	0.17	0.0111	0.002	1.6×10^{-6}	132989	0.0276	0.012	0.023	42,849
	rs6822892	4	157954125	PDGFC	A/G	0.68	0.0138	0.002	2.6×10^{-10}	103,432	0.00	1.00	0.0010	0.002	0.636	132951	0.0256	0.012	0.031	42,836
	rs4846565	1	217788727	LYPLAL1	G/A	0.67	0.0132	0.002	1.8×10^{-9}	99,014	0.00	1.00	0.0066	0.002	0.003	127468	0.0132	0.012	0.254	42,853
	rs3822072	4	89960292	FAM13A	A/G	0.48	0.0116	0.002	1.8×10^{-8}	99,977	0.00	1.00	0.0025	0.002	0.236	129432	0.0161	0.011	0.143	42,850
	rs6912327	6	34872900	UHRF1BP1	T/C	0.80	0.0165	0.003	2.3×10^{-8}	80,010	0.04	0.91	0.0074	0.003	0.011	103826	0.0139	0.0139	0.391	34,761
							Primary trait						FG				FI (BMI-adjusted)			
2hGlu	rs6975024	7	44198411	GCK	C/T	0.15	0.1026	0.016	5.2×10^{-11}	42,842	0.00	1.00	0.0605	0.003	2.9×10^{-99}	103,517	0.0063	0.003	0.030	98,458
	rs11782386	8	9239197	PPP1R3B*	C/T	0.87	0.0985	0.017	2.2×10^{-9}	42,852	0.00	1.00	–0.0167	0.003	5.5×10^{-7}	100,595	–0.0164	0.003	6.9×10^{-7}	95,565
	rs1019503	5	96280573	ERAP2	A/G	0.48	0.0628	0.011	8.9×10^{-9}	42,851	19.6	0.42	–0.0061	0.002	0.003	108,113	0.0004	0.002	0.851	103,448
2hGlu (BMI-adjusted)	rs7651090	3	186996086	IGF2BP2	G/A	0.30	0.064	0.012	4.5×10^{-8}	42,792	63.4	0.01	0.0128	0.002	1.8×10^{-8}	104,019	0.0003	0.002	0.900	98,924

Table 4.1. SNPs associated with fasting glucose, fasting insulin and 2-hour glucose at genome-wide significance in Europeans. Genome-wide loci for fasting glucose (FG), fasting insulin (FI), FI (adjusted for BMI) and 2hGlu are shown along with results for the other traits aligned to the trait-raising allele for the primary trait. Non-MAGIC SNPs (identified in other consortia and selected for the Metabochip to follow up on other non-MAGIC traits) are indicated in bold. Freq denotes the allele frequency of the primary trait-raising allele. Per-allele effect (standard error, SE) for FI represents differences in natural log–transformed levels of FI. N represents sample size. Heterogeneity was assessed using the I^2 index (Higgins and Thompson 2002). The gene shown is the nearest gene to the lead SNP, except for those marked with an asterisk, for which the nearest gene is also listed in Table ST2b–e.

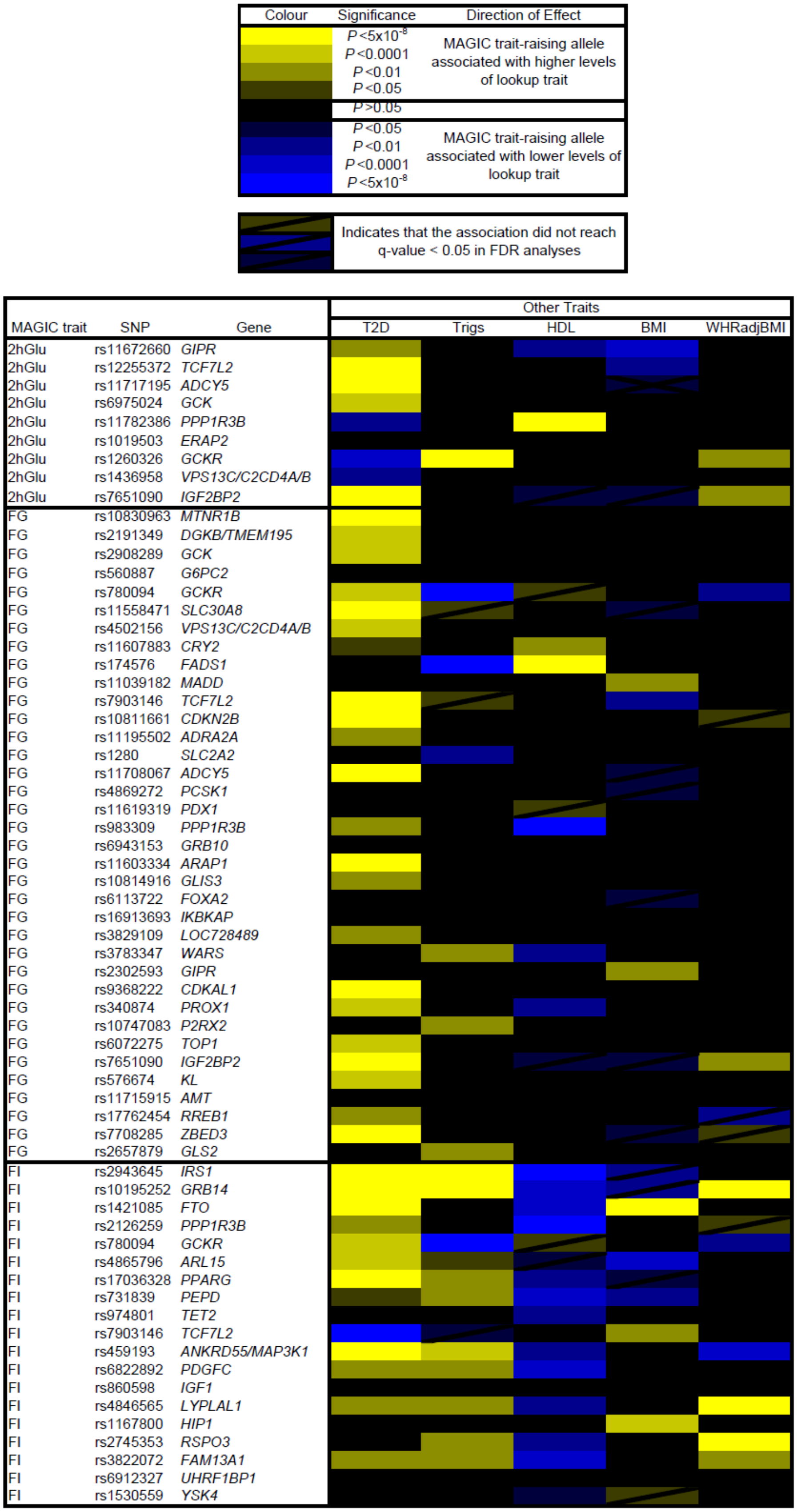

Colour
Significance
Direction of Effect
$P<5\times10^{-8}$
$P<0.0001$
$P<0.01$
$P<0.05$
MAGIC trait-raising allele associated with higher levels of lookup trait
$P>0.05$
$P<0.05$
$P<0.01$
$P<0.0001$
$P<5\times10^{-8}$
MAGIC trait-raising allele associated with lower levels of lookup trait
Indicates that the association did not reach q-value < 0.05 in FDR analyses
MAGIC trait
SNP
Gene
Other Traits
T2D
Trigs
HDL
BMI
WHRadjBMI
2hGlu rs11672660 GIPR
2hGlu rs12255372 TCF7L2
2hGlu rs11717195 ADCY5
2hGlu rs6975024 GCK
2hGlu rs11782386 PPP1R3B
2hGlu rs1019503 ERAP2
2hGlu rs1260326 GCKR
2hGlu rs1436958 VPS13C/C2CD4A/B
2hGlu rs7651090 IGF2BP2
FG rs10830963 MTNR1B
FG rs2191349 DGKB/TMEM195
FG rs2908289 GCK
FG rs560887 G6PC2
FG rs780094 GCKR
FG rs11558471 SLC30A8
FG rs4502156 VPS13C/C2CD4A/B
FG rs11607883 CRY2
FG rs174576 FADS1
FG rs11039182 MADD
FG rs7903146 TCF7L2
FG rs10811661 CDKN2B
FG rs11195502 ADRA2A
FG rs1280 SLC2A2
FG rs11708067 ADCY5
FG rs4869272 PCSK1
FG rs11619319 PDX1
FG rs983309 PPP1R3B
FG rs6943153 GRB10
FG rs11603334 ARAP1
FG rs10814916 GLIS3
FG rs6113722 FOXA2
FG rs16913693 IKBKAP
FG rs3829109 LOC728489
FG rs3783347 WARS
FG rs2302593 GIPR
FG rs9368222 CDKAL1
FG rs340874 PROX1
FG rs10747083 P2RX2
FG rs6072275 TOP1
FG rs7651090 IGF2BP2
FG rs576674 KL
FG rs11715915 AMT
FG rs17762454 RREB1
FG rs7708285 ZBED3
FG rs2657879 GLS2
FI rs2943645 IRS1
FI rs10195252 GRB14
FI rs1421085 FTO
FI rs2126259 PPP1R3B
FI rs780094 GCKR
FI rs4865796 ARL15
FI rs17036328 PPARG
FI rs731839 PEPD
FI rs974801 TET2
FI rs7903146 TCF7L2
FI rs459193 ANKRD55/MAP3K1
FI rs6822892 PDGFC
FI rs860598 IGF1
FI rs4846565 LYPLAL1
FI rs1167800 HIP1
FI rs2745353 RSPO3
FI rs3822072 FAM13A1
FI rs6912327 UHRF1BP1
FI rs1530559 YSK4

Figure 4.2. Associations between glycemic loci and T2D, HDL-cholesterol and triglycerides, BMI, and WHR. Loci associated with the above traits (P<0.05) are highlighted. Those with positively correlated effect distributions are colored yellow, and negative correlations are blue. Those which did not reach a q-value < 0.05 in FDR analyses are also marked.

Only for rs2302593, near GIPR, did we observe that BMI-adjustment materially altered the magnitude and significance of FG effect estimate (BMI-adjusted: ß=0.010(SE=0.0023) mmol/L/allele, P=2.9x10^{-5}; unadjusted: ß=0.014(0.0023) mmol/L/allele, P=9.3 x10^{-10}), suggesting that its effect on FG was mediated in part by its association with BMI (Lyssenko et al. 2011) (the FG-raising allele was also associated with higher BMI) (Figure 4.2).

Given the overlap between genetic loci for fasting glucose and other metabolic traits, we performed a systematic look-up of all glycemic loci and their associations with other metabolic traits using data available through other consortia (Heid et al. 2010; Speliotes et al. 2010; Teslovich et al. 2010). In the latest data from DIAGRAM Metabochip analyses ((Morris et al. 2012)), 22 (>60%) of the now 36 genome-wide significant FG loci showed association (P<0.05; FDR q<0.05) with T2D (Figure 4.2). In all cases, the glucose-raising allele was associated with increased risk of T2D, yet the FG effect size and T2D OR were only weakly correlated (Figure 4.3). FG loci also showed associations with other traits, although with inconsistent directionality (Figure 4.2).

Gene-based analyses confirmed many of the loci identified in individual SNP analyses (Table ST3a) and identified another nine genomic regions (containing 14 genes) with significant association signals (P<5x10^{-6}), including some with biological candidacy, such as the HKDC1 gene, encoding a putative hexokinase (Brandstatter et al. 2004; Irwin and Tan 2008).

4.3.1.2 Fasting insulin

In a combined sample size of 108,557 individuals, we identified 17 additional loci with genome-wide significant associations and confirmed both previously established associations (Dupuis et al. 2010). These newly identified loci include variants in or near HIP1, TET2, YSK4, PEPD and FAM13A1 genes (Table 4.1, Figure 4.8, Figure 4.12, Table 4.2), as well as SNPs near loci previously associated with other metabolic traits including T2D (Voight et al. 2010) (TCF7L2, PPARG), BMI (Frayling et al. 2007) (FTO), waist-hip ratio (Heid et al. 2010) (WHR) (LYPLAL1, RSPO3, GRB14), triglycerides (Teslovich et al. 2010) (ANKRD55/MAP3K1) and adiponectin (Richards et al. 2009) (ARL15). We also confirmed the recent associations with FI for GRB14, PPP1R3B, LYPLAL1, IRS1, UHRF1BP1 and PDGFC ((Manning et al. 2012)). The ANKRD55/MAP3K1 association is of interest as MAP3K1 regulates expression of IRS129 and activation of the JNK pathway (Yujiri et al. 1998) and NF-kB (Meyer et al. 1996; Lee et al. 1997), both centrally implicated in insulin resistance (Hirosumi et al. 2002; Cai et al. 2005). Furthermore, data from DIAGRAM Metabochip analyses show that the FI-raising allele at this SNP is strongly associated with increased risk of T2D ((Morris et al. 2012)), yet was also associated with lower WHR (adjusted for BMI) (Figure 4.2).

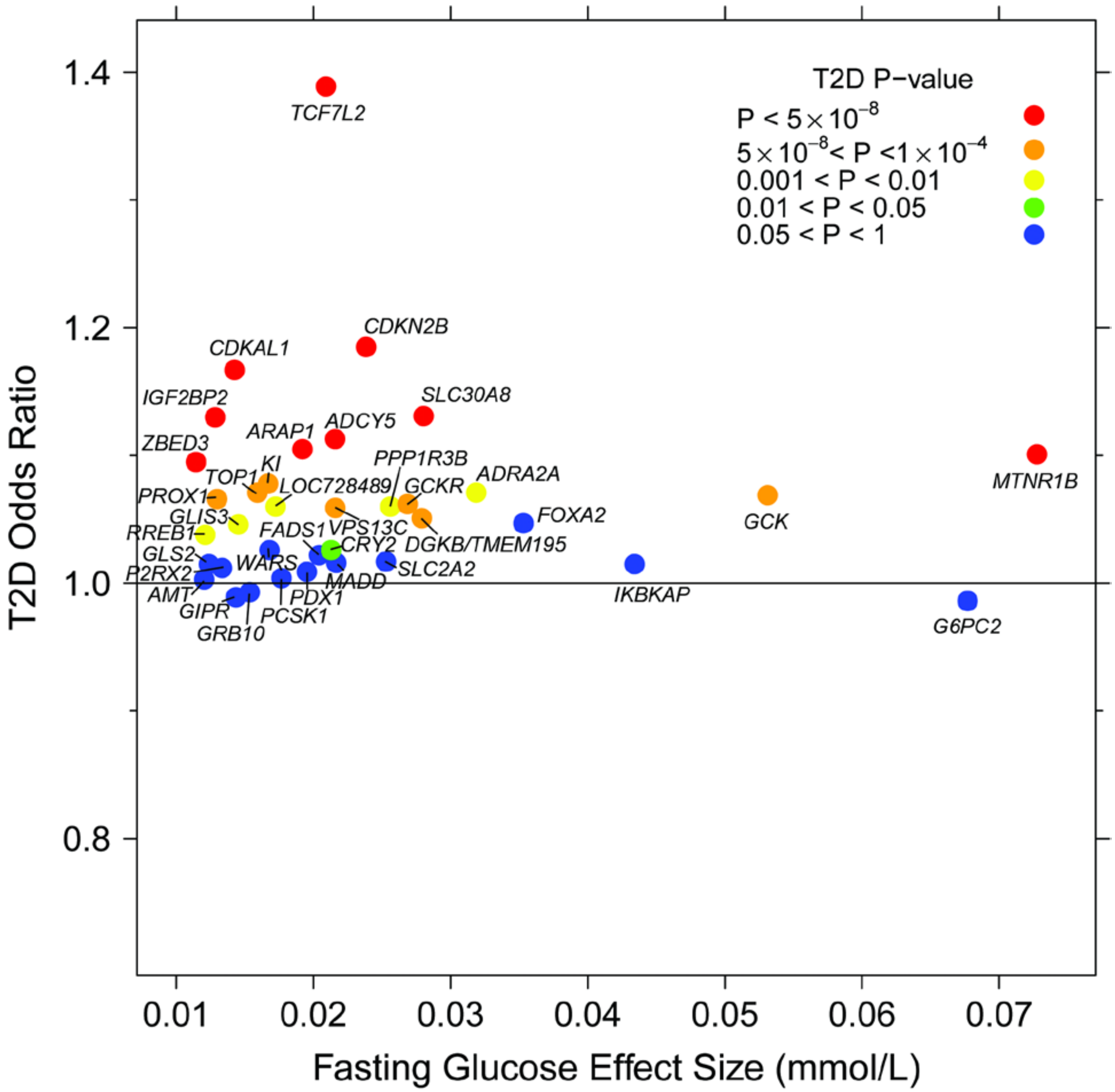


Figure 4.3. Per-allele beta-coefficients for fasting glucose concentration vs. odds ratios for T2D.

In contrast to FG (Figure 4.19), in FI analyses adjusted for BMI we observed a systematic decrease in the standard errors of the SNP effect estimates (Figure 4.20), perhaps because BMI explains more of the variance in FI (R2=32.6%) than in FG (R2=8.6%) or 2hGlu (11.0%) (Data from the Fenland study). Therefore, adjusting for BMI removes more variance in FI thereby rendering genetic associations more readily detectable. This is supported by the 11 genome-wide significant loci detected by this approach (Table 4.1, Figure 4.9, Figure 4.13), of which only four overlapped with the

BMI-unadjusted analysis (Table 4.1). As expected, BMI-adjustment abolished FI associations at FTO (ß=-0.001(SE=0.002), P=0.71) (Table ST2c) suggesting that the association with FI is mediated entirely through the association with BMI.

In total, 13 of the 19 loci associated with FI also showed associations with T2D (P<0.05; FDR q<0.05) (Figure 4.2), with the FI-raising allele associated with higher risk of T2D, except for TCF7L2 (Figure 4.4), a locus known to exert effects on insulin secretion and where the allele associated with lower FI is associated with higher FG (Table 4.1). Notably, the loci associated with fasting insulin showed a pattern of association with lipid traits consistent with insulin resistance not observed for either FG or 2hGlu (Figure 4.2). Thirteen (~68%) of the 19 loci were associated with HDL-cholesterol (q<0.05): all FI-raising alleles were associated with lower HDL levels, 9 of which were also associated with higher triglycerides (q<0.05) (Figure 4.2). Further, the FI-raising alleles of four SNPs were associated with higher WHR (adjusted for BMI) (q<0.05) (Figure 4.2), another trait linked to insulin resistance while five SNPs were also associated with BMI, although with inconsistent direction (q<0.05) (Figure 4.2).

In gene-based analyses, we focused on BMI-adjusted results to account for the variance in FI explained by BMI. Beyond those loci containing genome-wide significant SNPs we identified 7 distinct regions (containing 22 genes) to be associated with FI after Bonferroni-correction ($P< 5x10^{-6}$). Amongst these genes, we identified many for which prior biological evidence suggests their role in pathways involved in insulin secretion or action (Table ST3b). While the lead SNP in PPARD was not genome-wide significant ($P=3.9x10^{-6}$), the PPARD gene, a regulator of adipose, hepatic and skeletal muscle metabolism (Barish et al. 2006) reached the gene-based significance threshold ($P<1x10^{-6}$). PPARD agonists have also been shown to induce insulin sensitizing effects in a murine model (Tanaka et al. 2003). In addition, we identified PTEN (Table ST3b), a gene previously suggested to affect glucose metabolism through regulation of insulin signaling (Butler et al. 2002), and in which a muscle-specific deletion protected mice from insulin resistance and diabetes resulting from high fat feeding (Wijesekara et al. 2005).

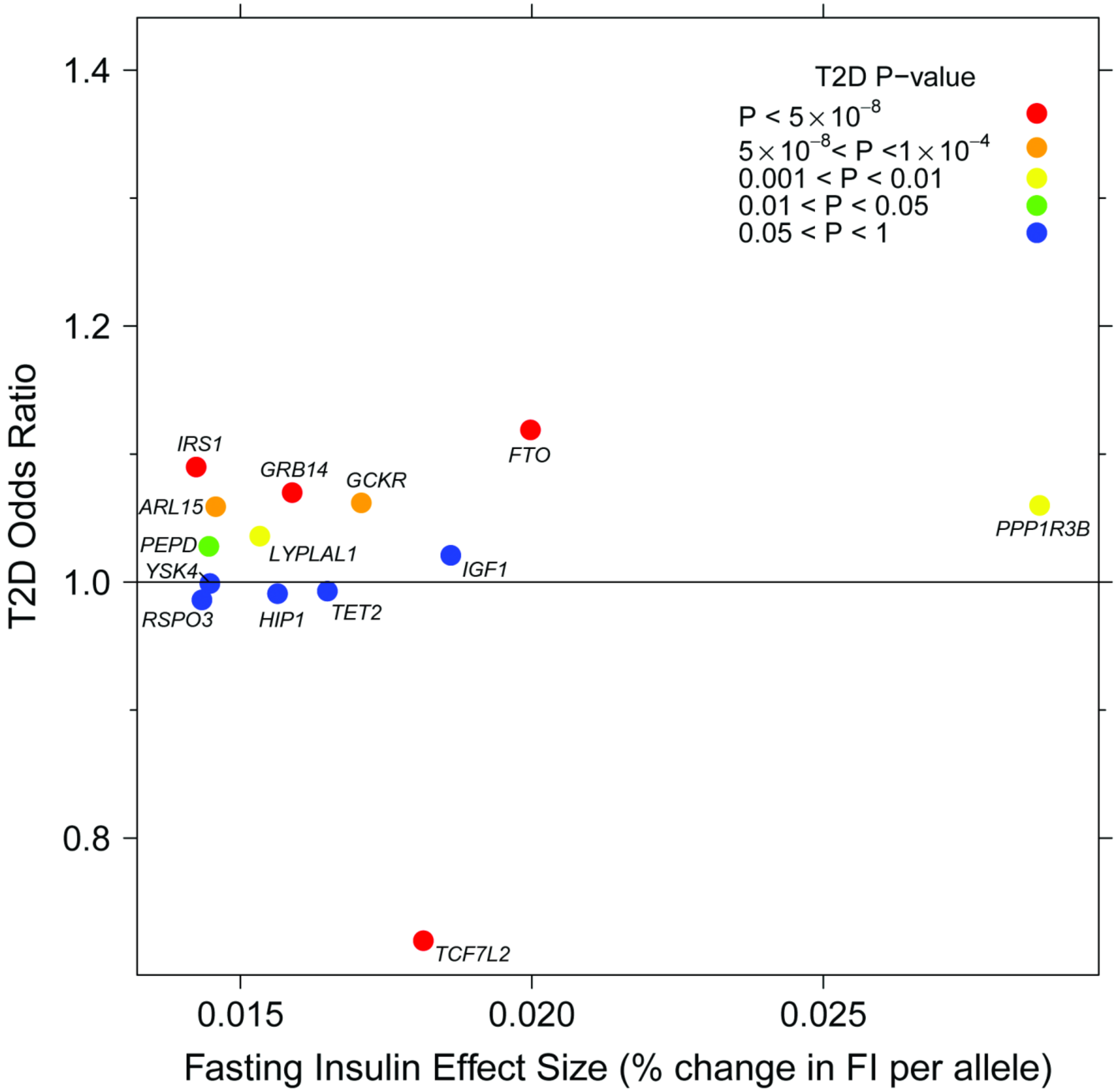


Figure 4.4. Per-allele beta coefficients for fasting insulin concentration vs. odds ratios for T2D.

4.3.1.3 2-h Glucose

In a total sample size of 42,854 individuals, we identified four additional loci to be associated with 2hGlu (Table 4.1, Figure 4.10, Figure 4.14), including a signal near ERAP2 and three signals near loci previously associated with FG (Dupuis et al. 2010) (GCK), HDL-cholesterol (Teslovich et al. 2010) (PPP1R3B) and T2D (Voight et al. 2010) (IGF2BP2), as well as confirming the five previous associations (Saxena et al. 2010). To determine whether these associations reflected differences in the response to a glucose challenge, or whether they were driven by effects on FG, we also performed analyses adjusted for FG. No additional loci were identified as genome-wide significant after adjustment for FG, although the GCK association with 2hGlu was severely attenuated after adjustment (ß=0.04(SE=0.016) mmol/L/allele; P=0.005, Vs. ß=0.1(0.016) mmol/L/allele; P=5.3x10^{-11} in the FG-unadjusted model) suggesting that the association with 2hGlu is driven, at least in part, by a primary association with FG (Table ST2e). The association of SNPs near GCK now with both FG and 2hGlu suggests a generalized raising of the glucose set point, consistent with inactivating mutations of GCK that cause MODY (Fajans et al. 2001). As for FG, when 2hGlu models were adjusted for BMI, no systematic differences were observed, although again the IGF2BP2 SNP rs7651090 reached genome-wide significance (Table 4.1).

Eight of the 9 SNPs associated with 2hGlu at genome-wide levels of significance were also associated with T2D (q<0.05) (Figure 4.2), although the 2hGlu-raising alleles at PPP1R3B, GCKR and VPS13C/C2CD4A/B were associated with lower risk of T2D (Figure 4.5), consistent with their association with lower FG levels (Table 4.1, Table ST2a).

In addition to those which were genome-wide significant in individual SNP analyses we identified 3 regions (containing 6 genes) showing association with 2hGlu in gene-based analyses. These included the HKDC1 gene also identified in gene-based analyses for FG, as well as an association signal at CRHR1 (P=2x10^{-6}) (Table ST3c); mostly driven by the lead SNP in this gene (rs17762954), which approached genome-wide significance (P=7.4x10^{-7}). CRHR1, together with GIPR, another gene with variants associated with 2hGlu, belongs to the family of class B GPCRs and is highly expressed in pancreatic ß-cells, where stimulation of the receptor potentiates insulin secretion in response to glucose (Schmid et al. 2011).

Fasting Glucose

IKBKAP (inhibitor of kappa light polypeptide gene enhancer in β-cells, kinase complex-associated protein): the protein encoded by this gene is a scaffold protein that binds IKKs and NF-kappa-B inducing kinase (NIK), assembling them into different active complexes. Splicing mutations in this gene lead to familial dysautonomia (Anderson et al. 2001). Also mapping to this region are ***C9orf4, C9orf5*** **and** ***C9orf6*****,** ***MIR32*** (microRNA 32, unknown function), as well as ***ACTL7A*** (actin-like 7A) and ***ACTL7B*** (actin-like 7B).

WARS (tryptophanyl-tRNA synthetase) catalyzes the aminoacylation of tRNA(trp) with tryptophan. The intronic SNP rs3783347 was associated with *WARS* expression in liver: the glucose-raising allele associated with lower mRNA expression (age- and sex-adjusted *P* =4.19x10-5) and is in perfect LD (r2=1, D'=1) with a 3'UTR SNP in *SLC25A47* (rs3736952) and in modest LD (r2=0.3, D'=1) with nonsynonymous Arg135Leu (qualified as tolerated by SIFT and probably damaging by polyphen). Nearby ***YY1*** (YY1 transcription factor), codes for a zinc-finger transcription factor involved in regulating a broad set of promoters. It has been suggested that YY1-regulated transcription is linked to glucose metabolism via O-GlcNAcylation (Hiromura et al. 2003).

KL (klotho): rs576674 lies ~36kb upstream of *KL,* which encodes a type-I membrane protein related to beta-glucosidases. Variation in *KL* has been associated with insulin regulation, insulin resistance phenotypes and cardiovascular disease in some studies (Rhee et al. 2006; Shimoyama et al. 2009; Oguro et al. 2010; Paroni et al. 2012) but *KL* variants were not associated with diabetes risk (Freathy et al. 2006). The various SNPS in these studies are all in weak LD with rs576674 (r^2<0.125). Variation in KL is also associated with bone metabolism and may play a role in associations of energy metabolism with bone metabolism (Mullin et al. 2005; Zarrabeitia et al. 2007).

TOP1 (topoisomerase (DNA) I). rs6072275 is intronic in *TOP1* and lies in a large region of high LD in Europeans, which includes the plausible biological candidate ***LPIN3*** (lipin 3). In mice, a related homolog *Lpin1* is associated with fatty liver dystrophy61, a phenotype similar to human lipodystrophy (loss of body fat, fatty liver, hypertriglyceridemia, and insulin resistance). *Lpin1* mRNA was expressed at high levels in adipose tissue and induced during differentiation of preadipocytes, suggesting that lipin is required for normal adipose tissue development while *LPIN2* has been suggested to be associated with T2D and glucose metabolism (Aulchenko et al. 2007). rs6072275 lies in the middle of a large CNV that extends from within the 3' end of *TOP1* to the 5' end of ***PLCG1*** (phospholipase C, gamma 1).

P2RX2 (purinergic receptor P2X, ligand-gated ion channel, 2). rs10747083 lies in a small CNV about 150kb upstream of five protein-coding genes, including ***P2RX2***, encoding one of a family of purinoceptors for ATP; ***GALNT9*** (UDP-N-acetyl-alpha-D-galactosamine:polypeptide Nacetylgalactosaminyltransferase 9 (GalNAc-T9), encoding one of a member of the UDP-N-acetylalpha-D-galactosamine polypeptide N-acetylgalactosaminyltransferase (GalNAc-T) family of enzymes and expressed specifically in the brain; ***FBRSL1*** (fibrosin-like 1); ***PXMP2*** (peroxisomal membrane protein 2, 22kDa); ***PGAM5*** (phosphoglycerate mutase family member 5), and within 184 kb downstream, ***POLE*** (polymerase (DNA directed), epsilon) and ***LOC100130238*** (hypothetical LOC100130238) a miscRNA.

LOC728489. rs3829109 is in low LD with a well-established locus for inflammatory bowel disease. Two recent publications reported *CARD9* SNP rs10781499 (r^2=0.29) to be associated with ulcerative colitis (Anderson et al. 2011), and CARD9/SNAPC4 rs4077515 (r^2=0.27) to be associated with Crohn's disease and ulcerative colitis (Franke et al. 2010; McGovern et al. 2010).Several genes are located in the region, but few with high plausibility for a role in glycemia.

AMT encodes the mitochondrial aminomethyltransferase which is a critical component of the glycine cleavage system. Depending upon the *AMT* transcript, rs11715915 is located in 3'UTR or within coding regions, where it causes a synonymous amino acid change. This SNP is also located downstream of ***TCTA*** (T-cell leukemia translocation altered) which has no known metabolic function and upstream of ***RHOA*** (ras homolog family member A). RHOA is a signaling molecule involved actin cytoskeleton stability and reorganization (Tang et al. 2012) that binds and activates Rho kinase (ROCK), a regulator of insulin transcription (Nakamura et al. 2006) and action (Furukawa et al. 2005) that is differentially regulated in T2D (Chun et al. 2011) and hypothesized to play a role in glucose homeostasis (Furukawa et al. 2005).

GLS2 encodes liver-expressed glutaminase 2, which is required for hydrolysis of glutamine. rs2657879 causes a benign (polyphen) amino acid change (L581P) in the GLS2 protein. The GLS2 protein is highly expressed (human protein atlas) by both liver and pancreas and it has been demonstrated in liver tumours that alterations in the balance of GLS2:GLS1 (the kidney-specific homologue) activity are important for regulating glutamate metabolism (Yuneva et al. 2012). The other gene in this region ***SPRYD4*** (SPRY domain containing 4) has no known function in metabolism.

RREB1 (ras responsive element binding protein 1) encodes a zinc finger transcription factor, with rs17762454 lying in an intron in the gene. The protein product of *RREB1* binds to RAS-responsive elements (RREs) of gene promoters, including the calcitonin gene promoter. The role of *RREB1* in energy metabolism is not known. An uncorrelated SNP at this locus (rs675209) was associated with serum urate levels (P=1.0x10^{-9}) in a GWAS of serum urate, gout and cardiovascular disease risk factors (Yang et al. 2010b). Another gene at this locus, ***SSR1*** (signal sequence receptor, alpha) encodes a glycosylated endoplasmic reticulum membrane receptor associated with protein translocation across the ER membrane. Reactome pathway analysis places this gene in a module with key roles in the synthesis and function of insulin, insulin-like growth factors and ghrelin, making this gene a plausible biological candidate at this locus. (REACTOME: REACT_15380). A third gene at this locus, ***CAGE1*** encodes cancer antigen 1. *CAGE1* has an unknown role in metabolism.

Fasting Insulin

TET2 (formerly *KIAA1546*) encodes the tet oncogene family member 2, isoform b which catalyzes the conversion of methylcytosine to 5-hydroxymethylcytosine. The enzyme is involved in myelopoiesis, and defects in this gene have been associated with several myeloproliferative disorders (NCBI RefSeq). Perhaps more relevant to glycemic regulation is ***PPA2***, which encodes the inorganic pyrophosphatase 2 isoform 1 precursor. Its protein product is localized to the mitochondrion; it has high homology to members of the inorganic pyrophosphatase family, including the signature sequence essential for its catalytic activity (NCBI RefSeq). Pyrophosphatases catalyze the hydrolysis of pyrophosphate to inorganic phosphate.

HIP1 encodes the huntingtin interacting protein 1, a membrane-associated protein that colocalizes with huntingtin. It is ubiquitously expressed with the highest level in brain. Loss of normal huntingtin-HIP1 interaction in Huntington disease may contribute to a defect in membrane-cytoskeletal integrity in the brain. Of interest to insulin action, *HIP1* is involved in clathrin-mediated endocytosis and trafficking. Mice transgenic for the mutated form of huntingtin develop diabetes (Hurlbert et al. 1999; Bjorkqvist et al. 2005); however, though Hip1/Hip1r double-knockout mice have severe vertebral defects, suffer from dwarfism and die in early adulthood, they do not show any fasting glucose abnormalities (Bradley et al. 2007). The lead SNP (rs1167800) is only 104 bp away from a nearby missense SNP (rs1167801), encoding a Q to H amino acid change; however, LD between them is low (r^2=0.196).

FAM13A (Family with sequence similarity 13, member A1) encodes a protein with unknown function. Previous GWAS for the study of lung function measures75 and chronic obstructive pulmonary disease (Cho et al. 2010) described variants in *FAM13A1* that affect these traits. ***SPP1,*** encoding osteopontin, a secreted matrix glycoprotein and pro-inflammatory cytokine involved in cell-mediated immunity is within 1Mb. Mice exposed to high fat diet show increased circulating osteopontin and over-expression of Spp1 in the macrophages recruited into adipose tissue improved insulin sensitivity (Nomiyama et al. 2007), while SPP1 was highly expressed in obese twins relative to their non-obese siblings (Pietilainen et al. 2008). Recent work linked osteopontin to beta cell function through the GIP pathway (Lyssenko et al. 2011). In carriers of the *GIPR* variant associated with impaired glucose and GIP-stimulated insulin secretion, osteopontin levels were lower compared to non-carriers. In addition, both GIP and osteopontin prevented cytokine-induced apoptosis and osteopontin-stimulated cell proliferation of functional beta cell mass.

PEPD* (**Peptidase D) encodes a member of the peptidase family. The protein forms a homodimer that hydrolyzes dipeptides or tripeptides with C-terminal proline or hydroxyproline residues. The enzyme serves an important role in the recycling of proline, and may be rate limiting for collagen production. ***CEBPA gene *(*CCAAT/enhancer binding protein (C/EBP) alpha) is ~100kb downstream of the lead SNP, is a transcription factor expressed in adipose tissue regulates a number of genes involved in lipid and glucose metabolism genes and a SNP in low LD with our lead SNP was previously associated with triglyceride levels (Olofsson et al. 2008). The cells from CEBPA (-/-) mice show a complete absence of insulin-stimulated glucose transport, secondary to reduced gene expression and tyrosine phosphorylation for the insulin receptor and IRS1 (Wu et al. 1999). CEBPA also modulates expression of leptin by binding to the promoter of the gene (Hollenberg et al. 1997) and our lead SNP showed modest association with BMI in previous GIANT meta-analyses (*P*=0.005).

YSK4 (Sps1/Ste20-related kinase homolog) contains rs1530559 in an intron. This gene has no known function in human energy metabolism. Three other genes at this locus also have no known role in energy metabolism, including ***RAB3GAP1*** (RAB3 GTPase activating protein subunit 1 (catalytic), encoding the catalytic subunit of a Rab GTPase activating protein. Mutations in this gene are associated with Warburg micro syndrome; ***CCNT2*** (cyclin T2), belonging to the highly conserved cyclin family, whose members are characterized by a dramatic periodicity in protein abundance through the cell cycle; and ***ACMSD*** (aminocarboxymuconate semialdehyde decarboxylase), involved in the de novo synthesis pathway of NAD from tryptophan. *ACMSD* has been implicated in the pathogenesis of several neurodegenerative disorders.

2h Glucose

ERAP2 (Endoplasmic reticulum aminopeptidase 2) Aminopeptidases hydrolyze N-terminal amino acids of

proteins or peptide substrates. The lead SNP was strongly associated with *ERAP2* expression in liver (*P*=1.1 x10^{-55}) in lymphoblastoid cell lines in individuals from the CEU (*P*=8x10^{-21}) and YRI samples (*P*=2x10^{-15}). Also near to this lead SNP is the ***LNPEP*** gene: Leucyl/cystinyl aminopeptidase, which is widely expressed and well characterised in muscle and fat cells. In response to insulin, LNPEP translocates to the cell surface and co-localises with GLUT482. Although the role it plays in insulin action is unknown, this translocation is impaired in individuals with T2D82. The ***PCSK1*** gene is also within 500kb of the lead SNP, although on the other side of a recombination hotspot (Figure 4.14).

Table 4.2. Genes nearest to loci associated with glycemic traits. Information regarding the most likely biological candidate genes near each associated locus for fasting glucose, fasting insulin, and 2-h glucose.

4.3.2 Fine-mapping of established loci

Higher SNP density around previously established loci did not generally yield stronger associations or more plausible functional variants (Table ST4). For fasting glucose, only for four of the 16 loci did we observe markedly more significant SNPs or larger effect size than the previous lead SNP: PROX1, GCK, ADRA2A and VPS13C/C2CD4A/B (Table ST4). Regional plots for these loci are shown in Figure 4.21. While the new lead SNP near ADRA2A was not markedly more significant than the previous lead, the effect size is almost double that of the previous lead SNP (Table ST4). However, this and other new lead SNPs were without more plausible functionality. The new lead SNP at VPS13C/C2CD4A/B, previously associated with proinsulin (Strawbridge et al. 2011), is far more significant and of larger effect size than the previous (ß=0.0273(SE=0.0035) mmol/L/allele; P=4.8x10^{-15} Vs. ß=0.0057(0.0036) mmol/L/allele; P=0.111; r^2=0.27). For FI, another SNP downstream of IGF1 was found to be more significant than the previous lead and with a larger effect size, although with no known functionality (Table ST4; Figure 4.21). For 2hGlu, again, another SNP at VPS13C/C2CD4A/B was more significant than the previous lead (Table ST4; Figure 4.21) and had previously been associated with diabetes in Chinese individuals (Cui et al. 2011).

4.3.3 Pathway analysis

Next, we explored whether glycemic loci were enriched for connectivity between genes representing particular pathways or molecular processes. To do this, we used GRAIL software (Raychaudhuri et al. 2009a) and investigated both an excess of connectivity between the established loci (genome-wide significant) and then between established loci and those loci that did not reach genome-wide significance, but which showed a lower level of association (P<0.0005) (Methods 4.2.10.3). We aimed to establish whether there were any biologically relevant genes amongst this longer list of suggestively-associated loci. This more liberal threshold yielded 218, 155, and 100 regions for FG, FI and 2hGlu, respectively. To further assess whether our loci represented common biological pathways, we used MAGENTA to undertake gene-set enrichment analyses (Methods 4.2.10.4).

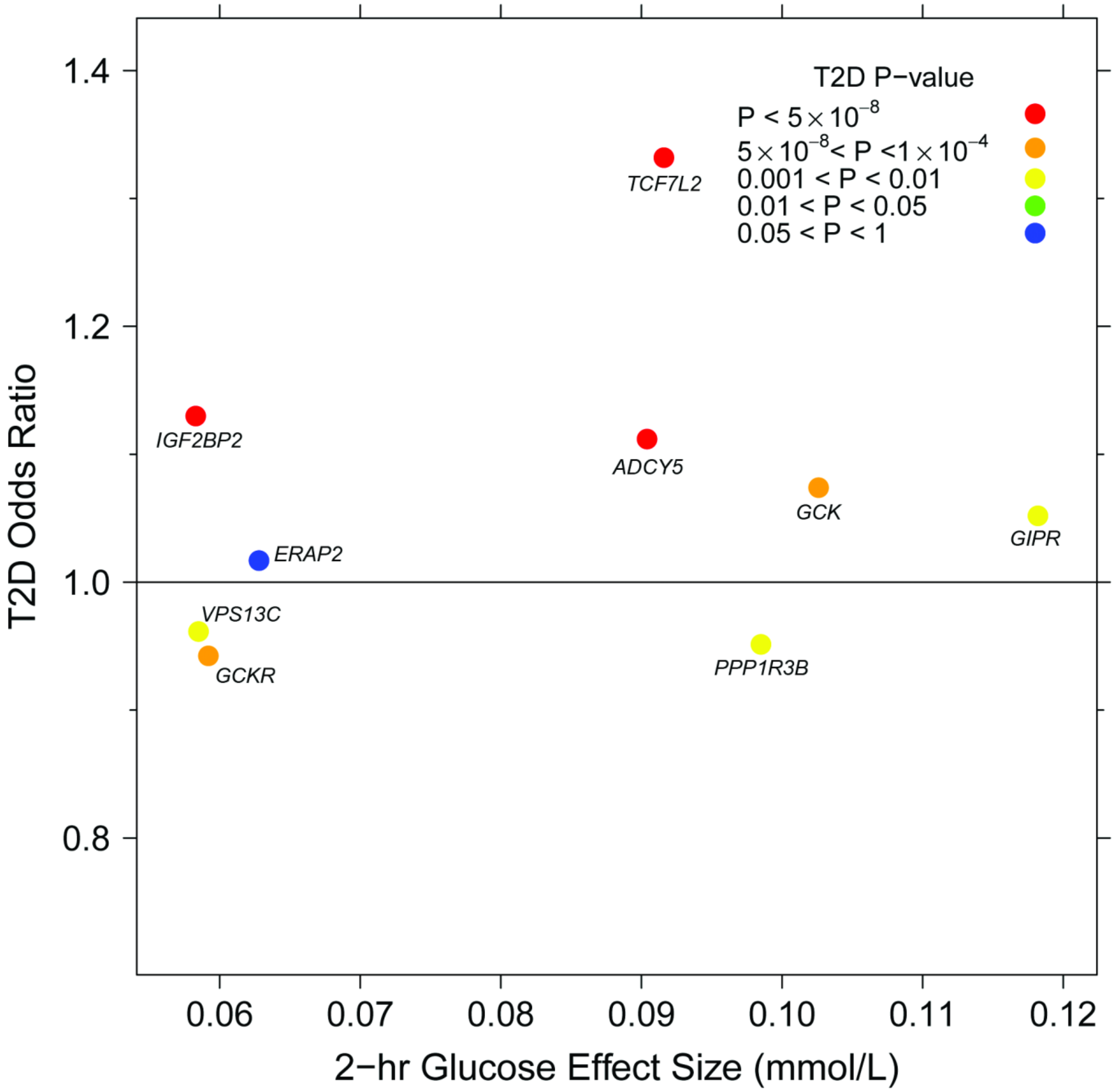


Figure 4.5 Per-allele beta coefficients for 2-hr glucose concentration vs. odds ratios for T2D.

We found that genes near the 36 FG-associated loci had a high degree of connectivity (Methods 4.2.10.3), with 8 genes demonstrating highly significant similarity to genes in other loci at a P_{grail}<0.01 level, connected by keywords such as “glucose”, “insulin”, “pancreatic”, and “diabetes” (Table ST5a and Figure 4.26), and more than expected by chance ($P_{permutation}$=0.003). We observed less connectivity amongst the genome-wide significant FI loci than among those for FG, with no genes reaching P_{grail}<0.01. The same was true for 2hGlu: only one out of nine genes reaching P_{grail}<0.01 ($P_{permutation}$=0.07) (Table 5c).

Among the list of 218 suggestively-associated FG loci (P<0.0005) we observed 13 genes to be connected to those in the genome-wide significant loci at P_{grail}<0.01, more than expected by chance ($P_{permutation}$=0.003) (Table ST6a). These included genes such as GLP1R (P=3.3x10^{-7}) (a glucagon receptor that mediates the GLP-1 incretin effect and stimulates insulin release), IRS2 (P=6.9x10^{-5}; central to development and maintenance of ß- cell mass and function (Withers et al. 1998; Withers et al. 1999)) and the INS gene (P=2.5x10^{-6}; the insulin gene encoding proinsulin). The presence of these and other genes support our conjecture that many of the SNPs that did not reach genome-wide significance are likely to represent “real” associations. Of the 155 suggestively-associated loci for FI (adjusted for BMI), we observed seven to be connected to the genome-wide significant loci at P_{grail}<0.01; more than expected by chance ($P_{permutation}$=0.002), and these included INSR (P_{grail}=1.5x10^{-4}; encoding insulin receptor precursor), CD36 (P_{grail}=0.001; previously implicated in insulin resistance46), GCG (P_{grail}=0.008; glucagon gene), and HNF1A (P_{grail}=0.005; mutations in which are associated with MODY3 (Yamagata et al. 1996)) (Table ST6b). Of the 100 suggestively-associated loci for 2hGlu (P<0.0005), we found three to reach P_{grail}<0.01 ($P_{permutation}$=0.014) and the gene highlighted as most biologically connected to the genome-wide significant loci was again HNF1A (P_{grail}=3.4x10^{-4}) (Table ST6c).

Using MAGENTA, we identified four pathways that were enriched for FG associations: GOTERM pathways lens development in camera-type eye (P=0.004), PANTHER processes gut mesoderm development (P=0.009), other steroid Metabolism (P=0.02), and KEGG MODY pathway (P=0.03), although these were no longer significant after removing the lead genes (P>0.05), which were all known FG loci: PROX1 for eye and gut, G6PC2 and GCK for steroid and MODY pathways, respectively.

4.3.4 Directional consistency of associations between discovery and follow-up studies among SNPs not reaching genome-wide significance

Given the wealth of biologically plausible genes in loci not reaching genome-wide significance (Tables ST6a-c) and the deviation of the observed distribution from the expected in QQ plots even after removing all established loci (Figure 4.15 - Figure 4.18), we hypothesized that additional loci not reaching genome-wide significance were likely to represent “real” associations with small effects. To establish the presence of such “real” associations that did not reach genome-wide significance, we compared SNP associations in discovery studies (those included in the original meta-analyses for 42,078 (FG)/34,230 (FI)/15,252 (2hGlu) individuals (Dupuis et al. 2010; Saxena et al. 2010)) with those in “follow-up” studies (including newly available GWAS and studies genotyped on the Metabochip (consisting of 85,710 (FG)/69,240 (FI)/27,602 (2hGlu) individuals)). We identified all SNPs which had a nominally significant association (P<0.05) in the follow-up studies alone and, for these SNPs, performed a binomial test of whether more SNPs than expected by chance (50%) had a consistent direction of effect with that observed in the discovery analyses. We were also able to compare among SNPs submitted by different consortia to follow-up on associations with a range of traits (Figure 4.22 - Figure 4.25).

For each trait, evaluation of the 66,000 Metabochip follow-up SNPs demonstrated a significant excess of SNPs showing directionally consistent associations (P<0.05) compared to that expected by chance (FG: $P_{binomial}$=5.01x10^{-12}; FI: $P_{binomial}$=7.58x10^{-13}; FI (adjusted for BMI): $P_{binomial}$=9.76x10^{-9}; 2hGlu: $P_{binomial}$=2.37x10^{-6}; Table ST7). FDR analyses suggested that a number of these nominal associations in the follow-up studies are true positives for FG and FI in particular (23%: q<0.77 at P=0.05; Table ST7). As expected, SNPs submitted to follow up on each respective trait showed a particular excess of consistent associations (FG: $P_{binomial}$=1.30x10^{-4}; FI: $P_{binomial}$=2.69x10^{-4}; FI (adjusted for BMI): $P_{binomial}$ P=4.71x10^{-5}; 2hGlu: $P_{binomial}$=7.67x10^{-6}; Table ST7; Figure 4.22 - Figure 4.25), and FDR analyses suggested that a higher proportion of these nominal associations were likely to be true positives than for SNPs following up on other traits (Table ST7). Interestingly, when we evaluated consistency of association with FI (between discovery and follow-up) among SNPs submitted to the Metabochip by other consortia, SNPs submitted by GIANT (anthropometric traits) ($P_{binomial}$=1.52x10^{-8}) and by GLGC (lipid traits) ($P_{binomial}$=1.15x10^{-6}) also demonstrated a marked excess (Table ST7; Figure 4.23). From these consortia, SNPs submitted for follow-up of BMI ($P_{binomial}$=4.66x10^{-6}) and triglyceride associations ($P_{binomial}$=6.17x10^{-12}) showed particular enrichment of consistent directionality amongst nominally significant SNPs (Table ST7; Figure 4.23). As expected, when we performed the same test for FI adjusted for BMI, the observed enrichment among SNPs submitted by GIANT and GLGC was attenuated (Table ST7; Figure 4.24), although interestingly the SNPs submitted to follow up on associations with triglycerides in discovery analyses remained the most significant (P=3.18x10^{-7}, Table ST7; Figure 4.24). Of the 3,353 SNPs submitted for follow-up of triglyceride associations, 158 SNPs showed nominal significance (P<0.05) in follow-up studies and consistent direction of association with FI (adjusted for BMI) in both discovery and follow-up (Table ST7). In 139 (88%) of these SNPs the FI-raising alleles were associated with higher levels of triglycerides. This finding is consistent with the positive correlations between FI and triglyceride associations observed amongst the genome-wide significant FI loci (Figure 4.2).

4.4 Discussion

In the current meta-analysis of ~66,000 Metabochip follow-up SNPs, in up to 133,010 European individuals without diabetes we identified a large number of new loci that are associated with FG, FI and 2hGlu. In total, 53 loci influencing glycemic traits have now been validated. These loci explain 4.8%, 1.2% and 1.7% of the variance in each of these traits (FG, FI and 2hGlu, respectively). Of these 53 loci, 33 are also associated with increased T2D risk (q<0.05), extending the overlap between glycemic and T2D loci. Given the current DIAGRAM effective sample size of 106,953 individuals, we can exclude an effect on T2D of 1.04 with 80% power for alleles more frequent than 5%, effectively confirming that the overlap is incomplete and that many loci associated with glycemic traits have no discernible effect on T2D (Figure 4.2, Figure 4.3, Figure 4.4, Figure 4.5).

Previously, we had detected only two loci associated with FI, and hypothesized that this might be due to a different genetic architecture of this trait compared to FG, with potentially smaller effect sizes, lower frequency alleles or greater environmental influence on FI (Dupuis et al. 2010). In the current, much more extensive, follow-up meta-analysis totaling up to 108,557 individuals (compared to 62,264 individuals previously), we expanded the number of loci associated with this trait to 19. Of note was the effect of BMI-adjustment on our ability to detect additional loci (five non-overlapping with unadjusted results), demonstrating that BMI-adjustment removed a substantial proportion of the variance in FI, therefore facilitating the identification of genetic associations. We also noted that some of the loci influencing FI uncovered after BMI-adjustment are likely to have been negatively confounded in previous efforts: at some loci the FI-raising allele was nominally associated with lower BMI (potentially via insulin resistance attenuating the anabolic effects of insulin), and given the positive correlation between BMI and FI, it is likely that this association had previously masked their effect on FI. FI loci showed directionally consistent association with lipid levels (HDL and triglycerides); that is, the FI raising allele was associated with lower HDL and higher triglyceride levels: a hall-mark combination observed in insulin resistant individuals. We also observed some overlap between FI loci and those associated with abdominal obesity (Figure 4.2). Jointly, these data suggest links of these FI loci to insulin resistance-related phenotypes. Indeed, some of the FI loci identified such as IRS1 and PPARG are classically known to exert effects on insulin action or sensitivity (Spiegelman 1998; White 1998). Additionally, SNPs nominated for the Metabochip to follow up on their association with triglycerides had directionally consistent associations with FI (in discovery and follow-up) more often than expected by chance ($P=1.66x10^{-11}$). This remained true for their associations with FI after adjustment for BMI ($P=1.28x10^{-8}$).

There are now 36 established FG loci, many of which contain compelling biological candidate genes with plausible causality, including those encoding transcription factors with known roles in pancreas development (e.g. PDX1, FOXA2, PROX1, GLIS3) and genes involved in ß-cell function and insulin secretion pathways (SLC2A2, GCK, PCSK1). For 2hGlu, only nine loci have been established to date, which is likely reflecting the smaller sample size available and consequently reduced power.

Comparing the consistency of the direction of associations for glycemic traits between “discovery” and “follow-up” studies suggests that we are observing more directionally consistent associations than expected by chance (Figure 4.22 - Figure 4.25). This, combined with the excess of biologically plausible genes amongst the borderline loci (Table ST6a-c), suggests that beyond the genome-wide significant loci there is a more extensive list of loci still likely to contain “real” associations. Indeed some of these loci are implicated by gene-based analyses, which identify genes with compelling biological credentials. For FI, these analyses revealed additional loci with previously suggested links to insulin resistance: PPARD and PTEN, both of which were nominally associated in both stages ($P<0.05$). These results lend further support to the proposal that a long tail of common variants of small effect size are likely to account for a significant proportion of the variance of complex traits (Park et al. 2010; Yang et al. 2010a).

Of note is the number of glycemic loci associated with other metabolic traits ($q<0.05$: 34 of 53) and also at genome-wide levels of significance ($P<5x10^{-8}$) (14 of 53) (Figure 4.2), potentially implicating pleiotropic effects. Further support for this notion comes from the analysis of loci nominated for the Metabochip by other consortia and their associations with glycemic traits (Figure 4.22 - Figure 4.25). Indeed, some of the loci associated with glycemic traits at genome-wide significance levels were not originally nominated to the Metabochip for follow-up by MAGIC (Table 4.1.) Metabochip data available across all contributing consortia will facilitate systematic exploration of these correlated phenotypes with more sophisticated statistical methods for joint analysis (Kim et al. 2009; Kim and Xing 2009; Curtis et al. 2012), yielding greater insight into the underlying pathways and genetic networks they represent. As data from human genetic networks accrues, we will be better placed to test whether there is support for the notion of “hub” genes, that is, genes highly connected with others in the network and proposed by experiments in *C.elegans* to act as buffers for genetic variation and that could act as modifier genes for many different disorders (Lehner et al. 2006).

Here, we present a large number of genome-wide significant loci influencing glycemic traits, many of which with a compelling biological basis for their association with glycemia, as well as a number of loci not previously implicated in glycemic regulation, and for which fine-mapping and functional follow-up will expand and improve our understanding. Use of the Metabochip for deep follow-up has also suggested additional loci to be involved in glycemic regulation that, due to insufficient sample size and power, did not reach genome-wide significance. Consideration of such loci in future studies will better exploit data from GWAS and complimentary approaches and further improve our biological understanding of glycemic control and the etiology of diabetes.

4.5 Supplementary Figures and Tables

All supplemental tables are available online at:

http://www.umich.edu/~welchr/magic_supplemental/supplemental_tables.xls

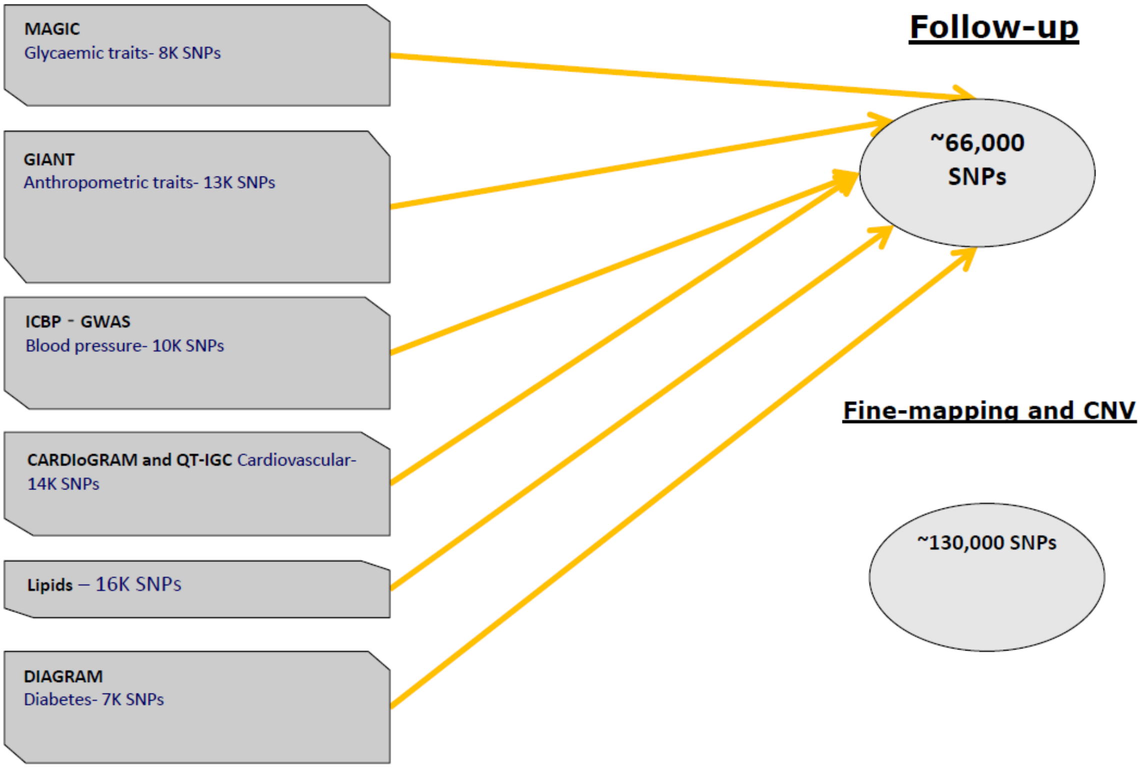


Figure 4.6. Metabochip design. Consortia and the number of SNPs they submitted to be followed-up using the Metabochip is shown.

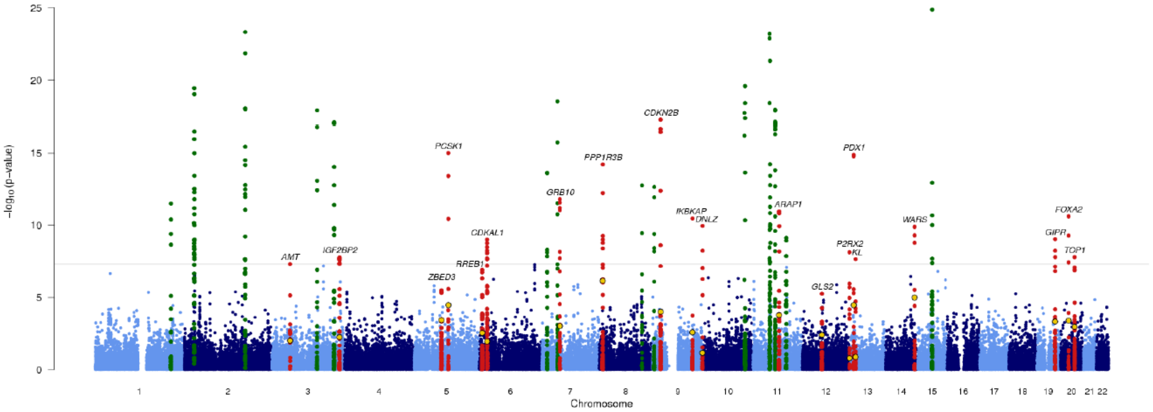


Figure 4.7. Manhattan plot of fasting glucose associations. Points shown in yellow mark the P-value for these SNPs for association with FG in the original discovery meta-analysis (Dupuis et al. 2010). Novel lead SNPs (+/- 500kb) are highlighted red, and known loci in green. P-values were cropped at P<1x10^{-25}. ZBED3, RREB1, and GLS2 were genome-wide significant only after adjustment for BMI, but are also highlighted red above.

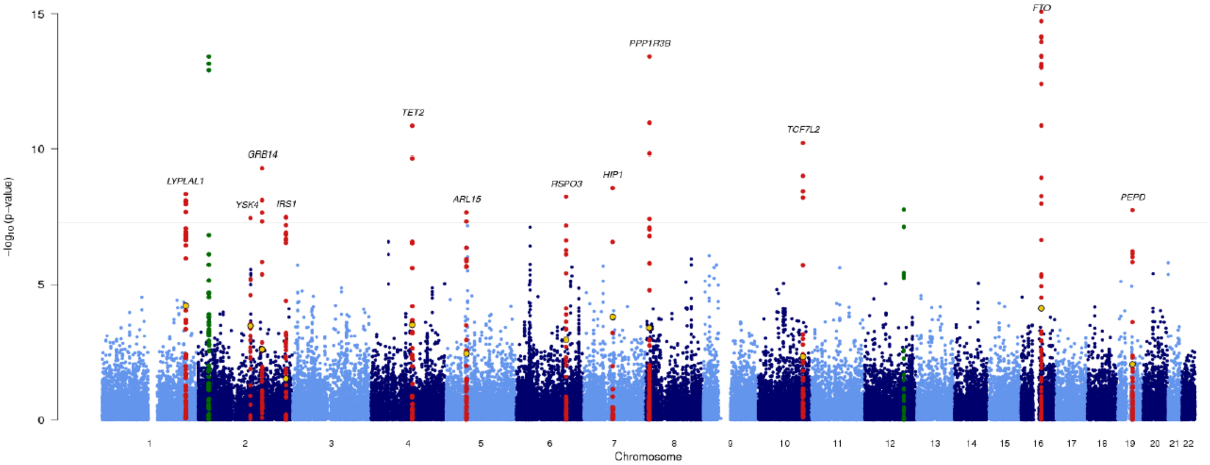


Figure 4.8. Manhattan plot of fasting insulin associations. Novel lead SNPs (+/- 500kb) are highlighted red, and known loci in green. Points shown in yellow mark the P-value for these SNPs for association with FI in the original discovery meta-analysis (Dupuis et al. 2010).

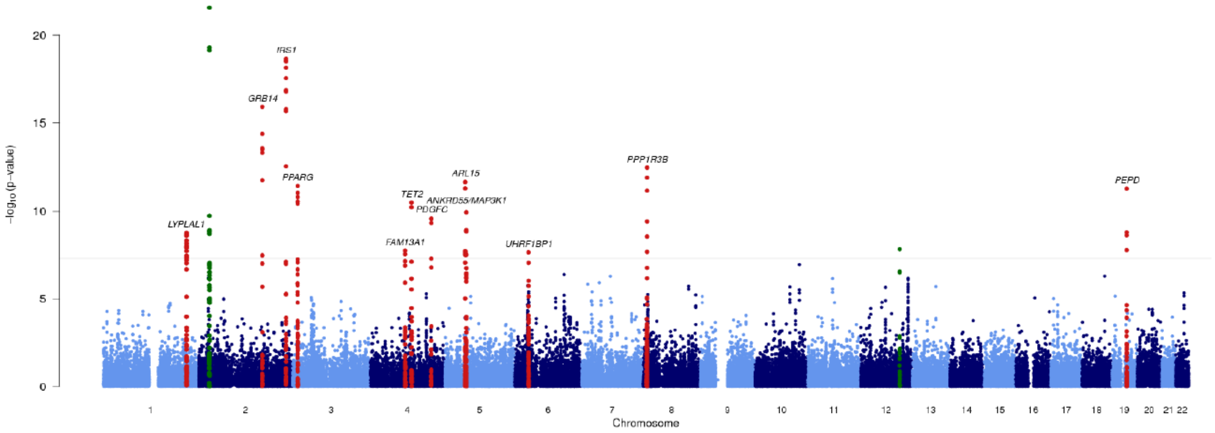


Figure 4.9. Manhattan plot of fasting insulin (adjusted for BMI) associations. Novel lead SNPs (+/- 500kb) are highlighted red, and known loci in green. Points shown in yellow mark the P-value for these SNPs for association with FI in the original discovery meta-analysis (Dupuis et al. 2010).

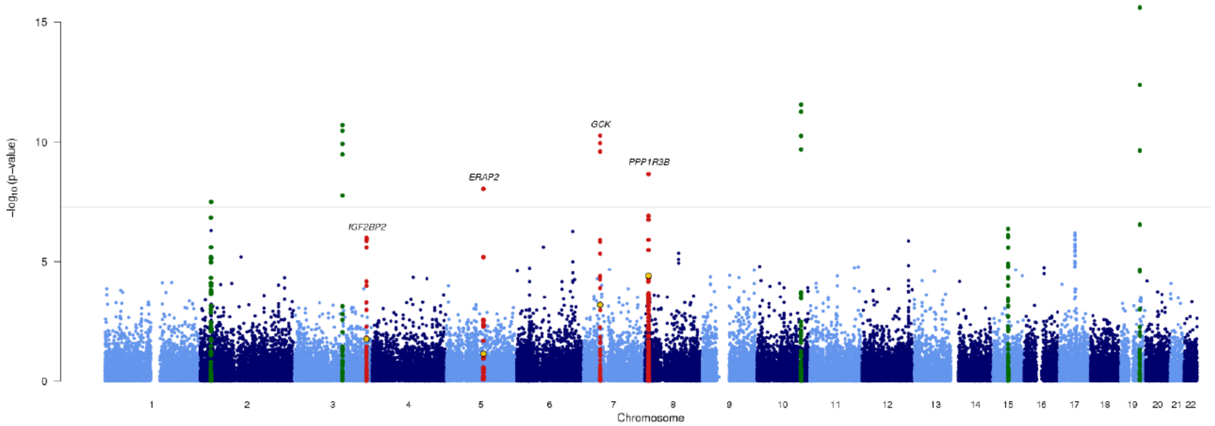
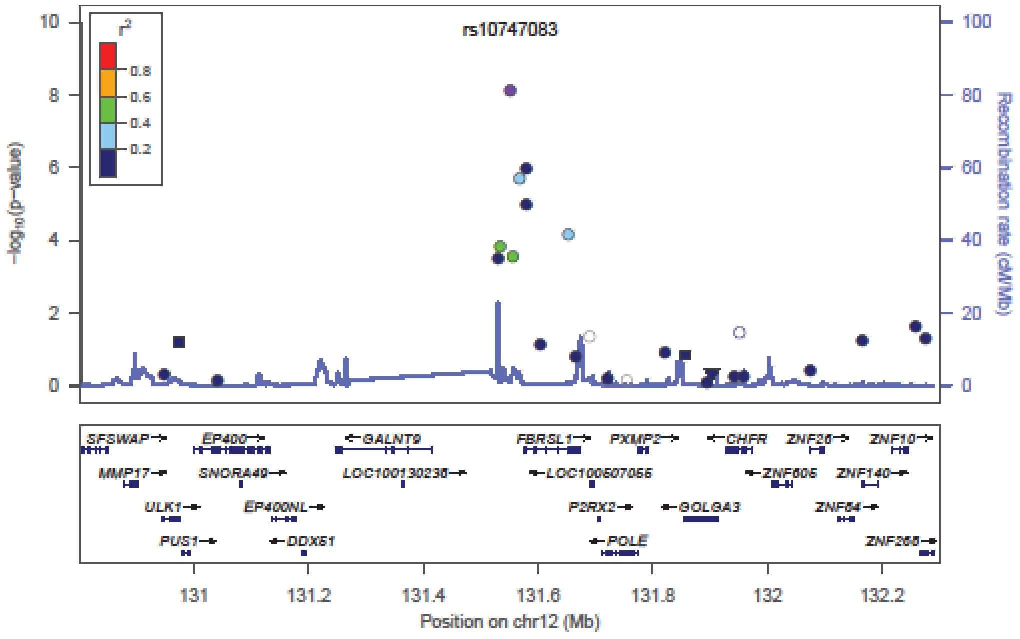
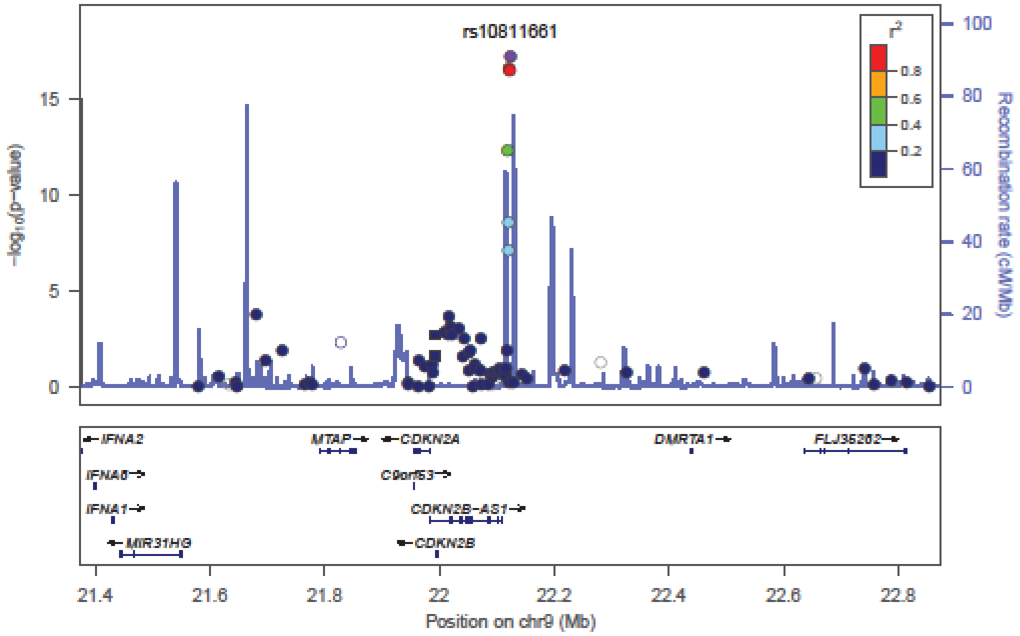
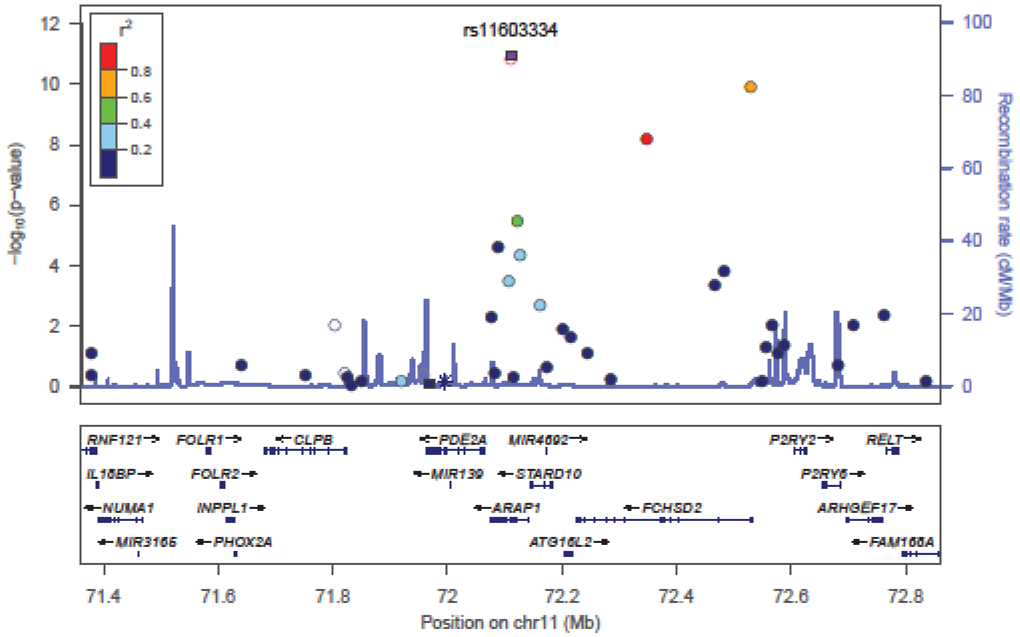
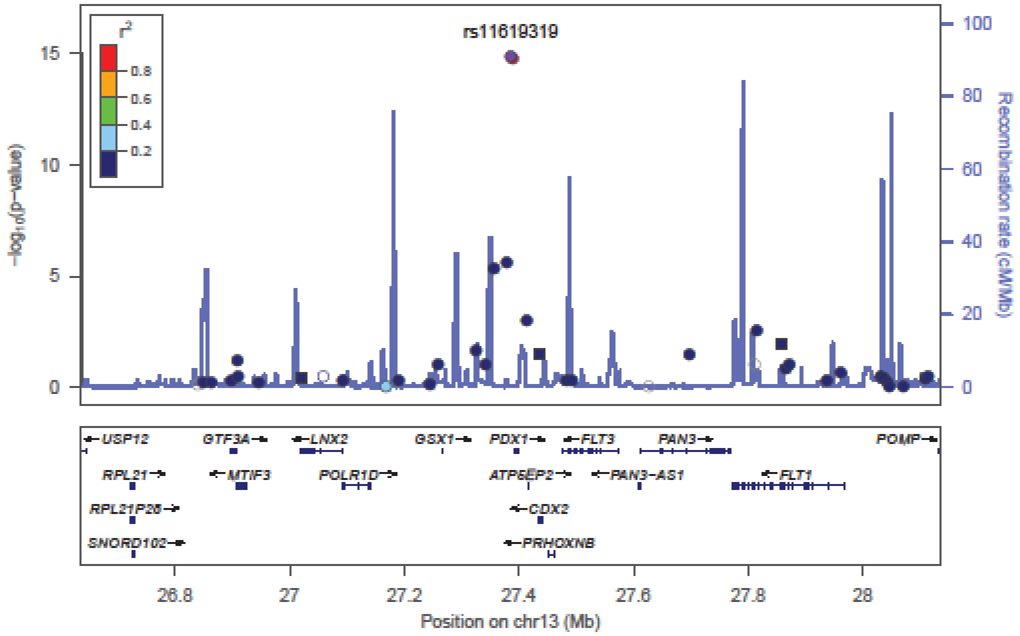


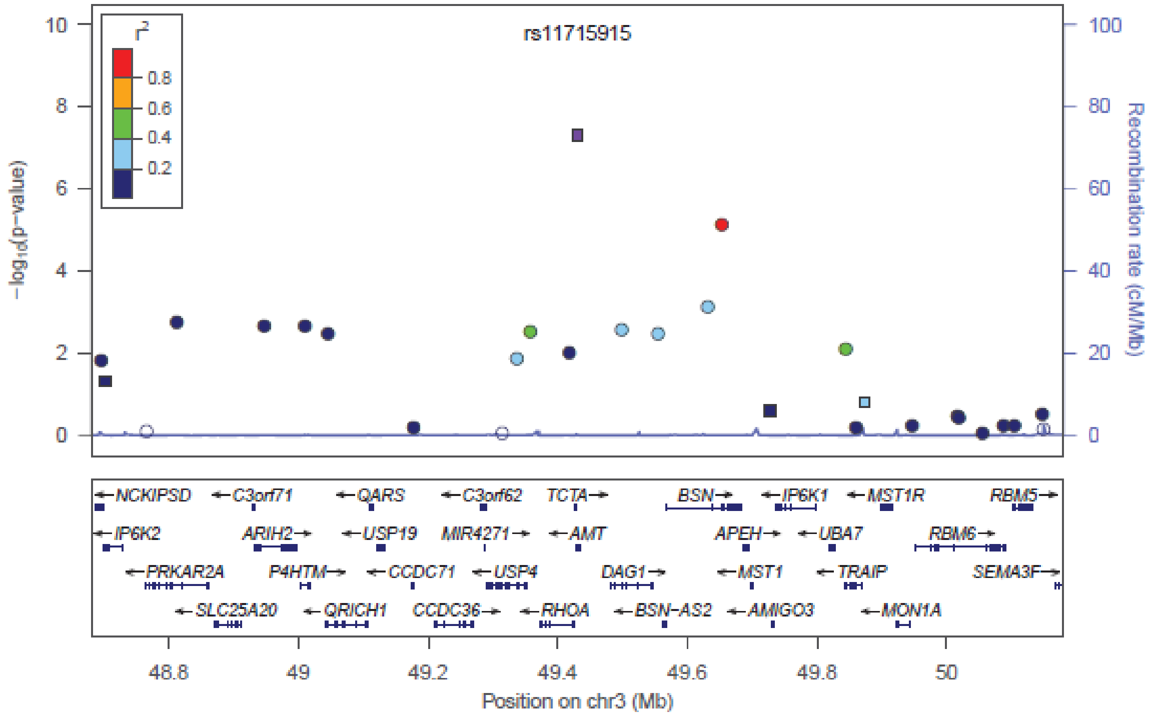
Figure 4.10. Manhattan plot of 2-h glucose associations. Novel lead SNPs (+/- 500kb) are highlighted red, and known loci in green. Points shown in yellow mark the P-value for these SNPs for association with 2hGlu in the original discovery meta-analysis (Saxena et al. 2010).

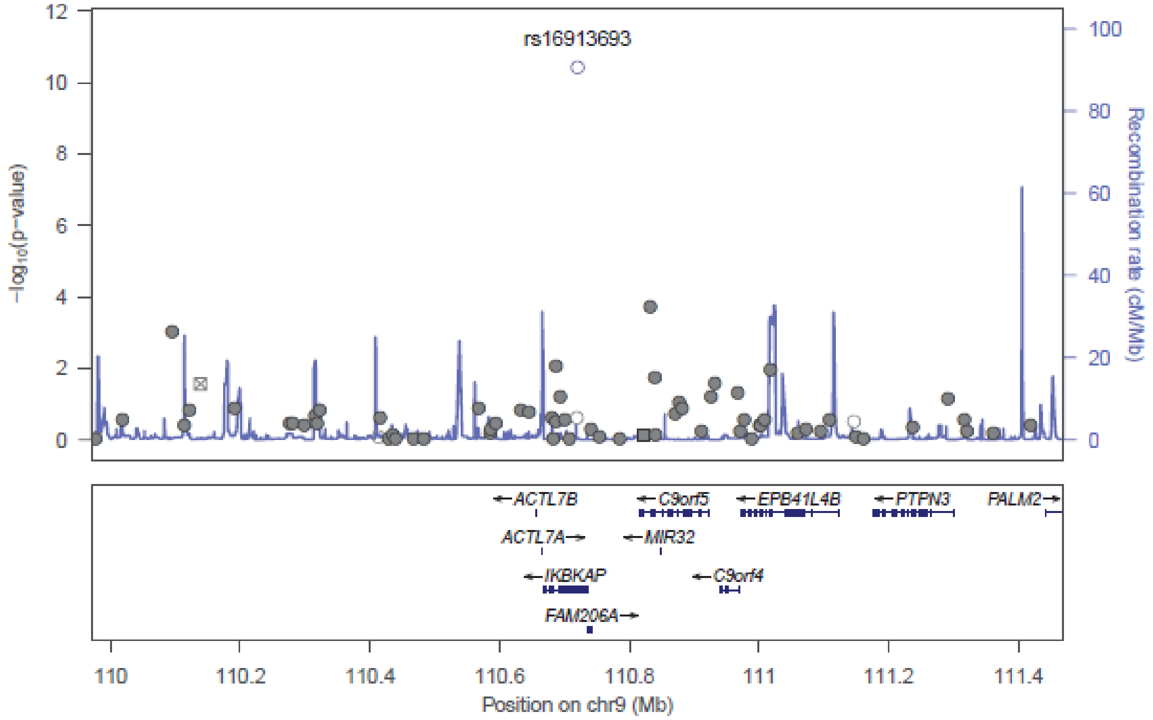

rs10747083
r²
0.8
0.6
0.4
0.2
−log₁₀(p−value)
10
8
6
4
2
0
Recombination rate (cM/Mb)
100
80
60
40
20
0
SFSWAP
EP400
GALNT9
FBRSL1
PXMP2
CHFR
ZNF26
ZNF10
MMP17
SNORA49
LOC100130238
LOC100507055
ZNF605
ZNF140
ULK1
EP400NL
P2RX2
GOLGA3
ZNF84
PUS1
DDX51
POLE
ZNF268
131
131.2
131.4
131.6
131.8
132
132.2
Position on chr12 (Mb)

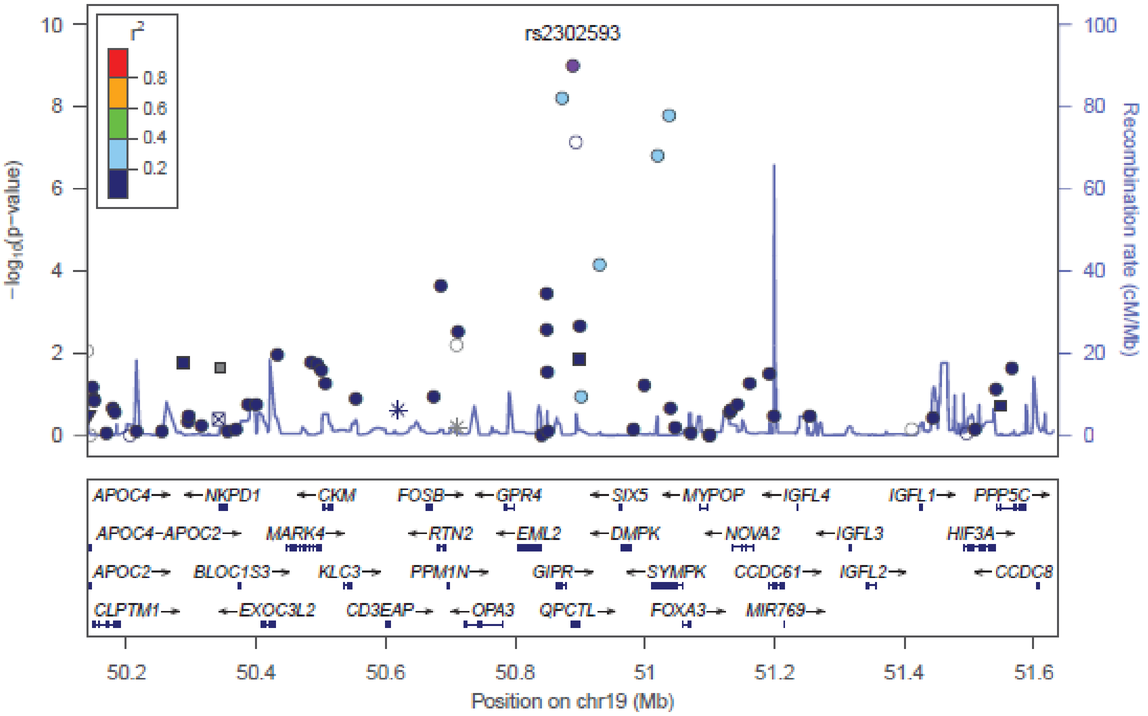

rs10811661
r²
0.8
0.6
0.4
0.2
−log₁₀(p−value)
15
10
5
0
Recombination rate (cM/Mb)
100
80
60
40
20
0
IFNA2
MTAP
CDKN2A
DMRT1
FLJ35282
IFNA8
C9orf53
IFNA1
CDKN2B−AS1
MIR31HG
CDKN2B
21.4
21.6
21.8
22
22.2
22.4
22.6
22.8
Position on chr9 (Mb)

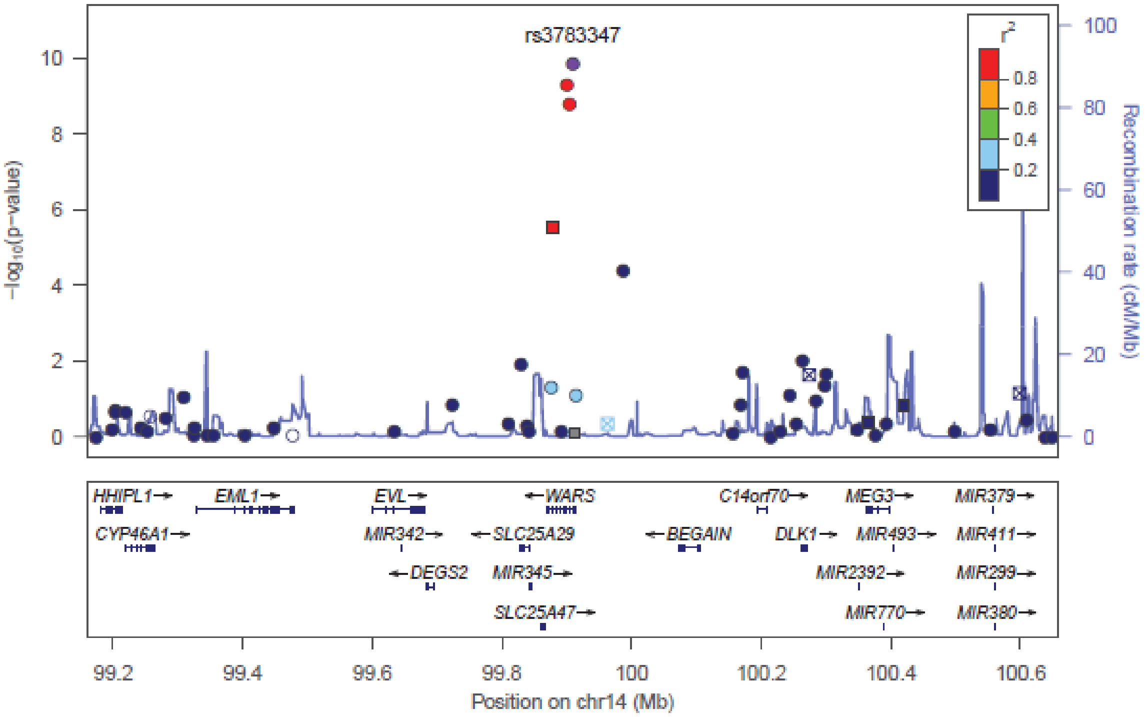

rs11603334
r^2
0.8
0.6
0.4
0.2
$-\log_{10}$(p-value)
12
10
8
6
4
2
0
100
80
60
40
20
0
Recombination rate (cM/Mb)
RNF121
FOLR1
CLPB
PDE2A
MIR4692
P2RY2
RELT
IL18BP
FOLR2
MIR139
STARD10
P2RY6
NUMA1
INPPL1
ARAP1
FCHSD2
ARHGEF17
MIR3166
PHOX2A
ATG16L2
FAM168A
71.4
71.6
71.8
72
72.2
72.4
72.6
72.8
Position on chr11 (Mb)

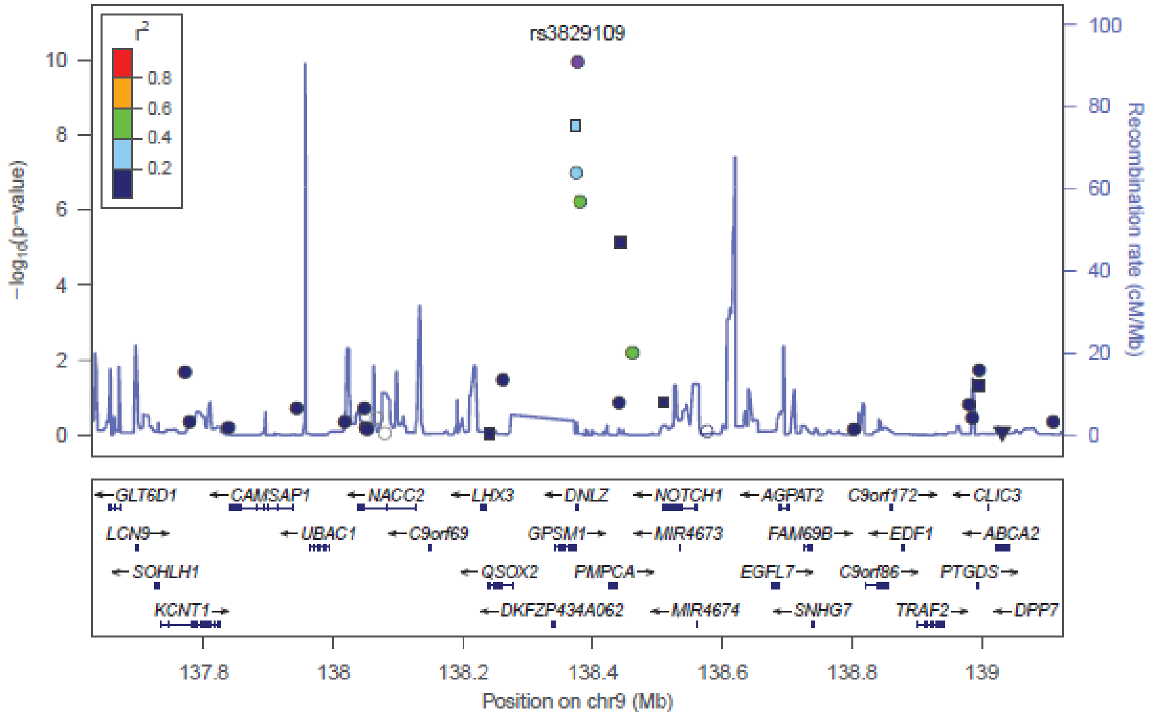

rs11619319
r^2
0.8
0.6
0.4
0.2
$-\log_{10}$(p-value)
15
10
5
0
100
80
60
40
20
0
Recombination rate (cM/Mb)
USP12
GTF3A
LNX2
GSX1
PDX1
FLT3
PAN3
POMP
RPL21
MTIF3
POLR1D
ATP5EP2
PAN3-AS1
FLT1
RPL21P28
CDX2
SNORD102
PRHOXNB
26.8
27
27.2
27.4
27.6
27.8
28
Position on chr13 (Mb)

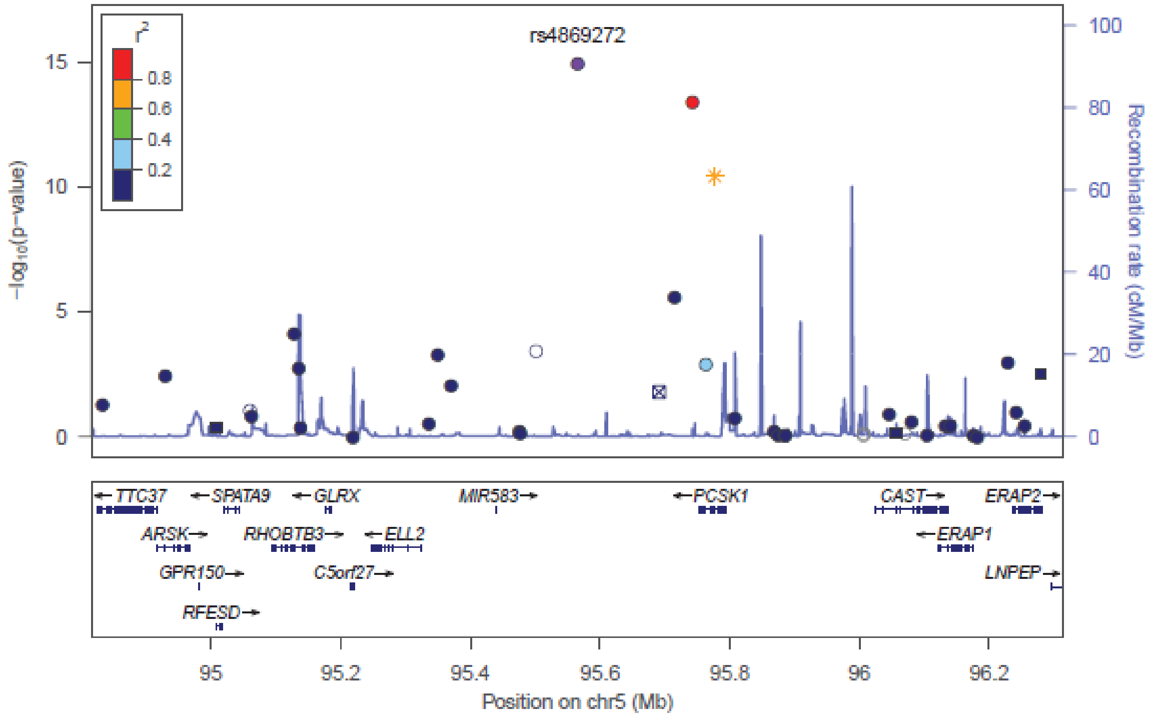

rs11715915
r2
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
NCKIPSD
C3orf71
QARS
C3orf62
TCTA
BSN
IP6K1
MST1R
RBM5
IP6K2
ARIH2
USP19
MIR4271
AMT
APEH
UBA7
RBM6
PRKAR2A
P4HTM
CCDC71
USP4
DAG1
MST1
TRAIP
SEMA3F
SLC25A20
QRICH1
CCDC36
RHOA
BSN-AS2
AMIGO3
MON1A
48.8
49
49.2
49.4
49.6
49.8
50
Position on chr3 (Mb)

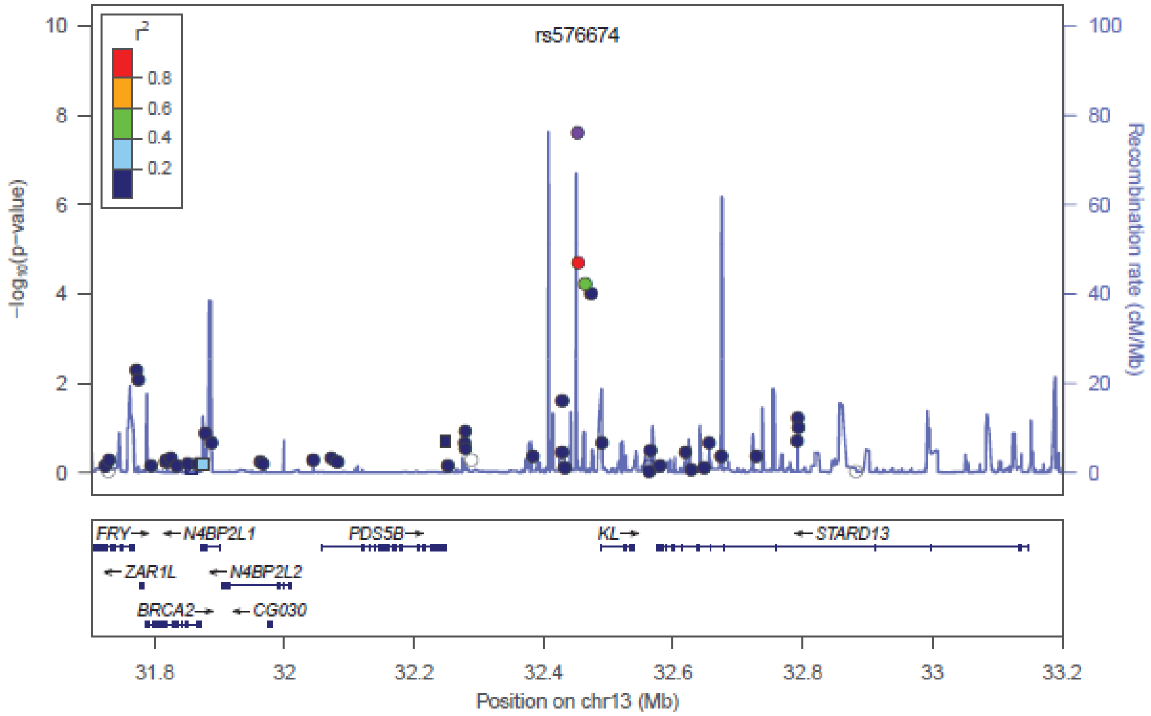

rs16913693
−log10(p−value)
Recombination rate (cM/Mb)
ACTL7B
C9orf5
EPB41L4B
PTPN3
PALM2
ACTL7A
MIR32
IKBKAP
C9orf4
FAM206A
110
110.2
110.4
110.6
110.8
111
111.2
111.4
Position on chr9 (Mb)

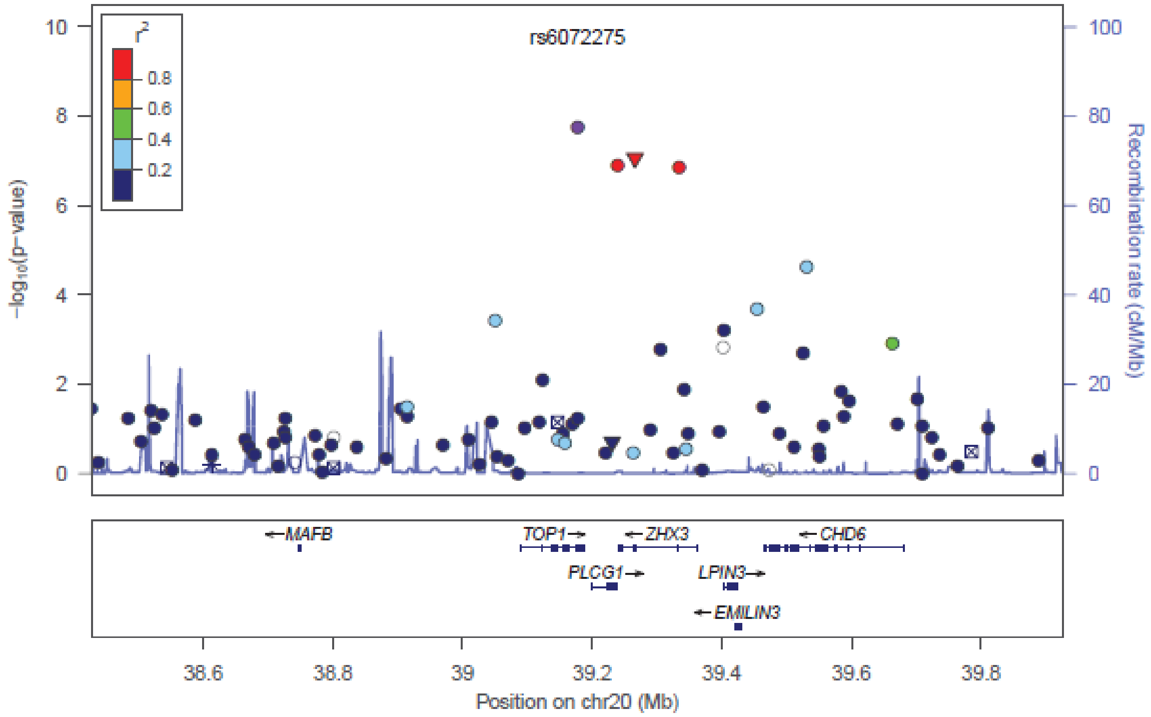

rs2302593
r2
0.8
0.6
0.4
0.2
10
8
6
4
2
0
−log10(p−value)
100
80
60
40
20
0
Recombination rate (cM/Mb)
APOC4→
←NKPD1
←CKM
FOSB→
←GPR4
←SIX5
←MYPOP
←IGFL4
IGFL1→
PPP5C→
APOC4−APOC2→
MARK4→
←RTN2
←EML2
←DMPK
←NOVA2
←IGFL3
HIF3A→
APOC2→
BLOC1S3→
KLC3→
PPM1N→
GIPR→
←SYMPK
CCDC61→
IGFL2→
←CCDC8
CLPTM1→
←EXOC3L2
CD3EAP→
←OPA3
QPCTL→
FOXA3→
MIR769→
50.2
50.4
50.6
50.8
51
51.2
51.4
51.6
Position on chr19 (Mb)

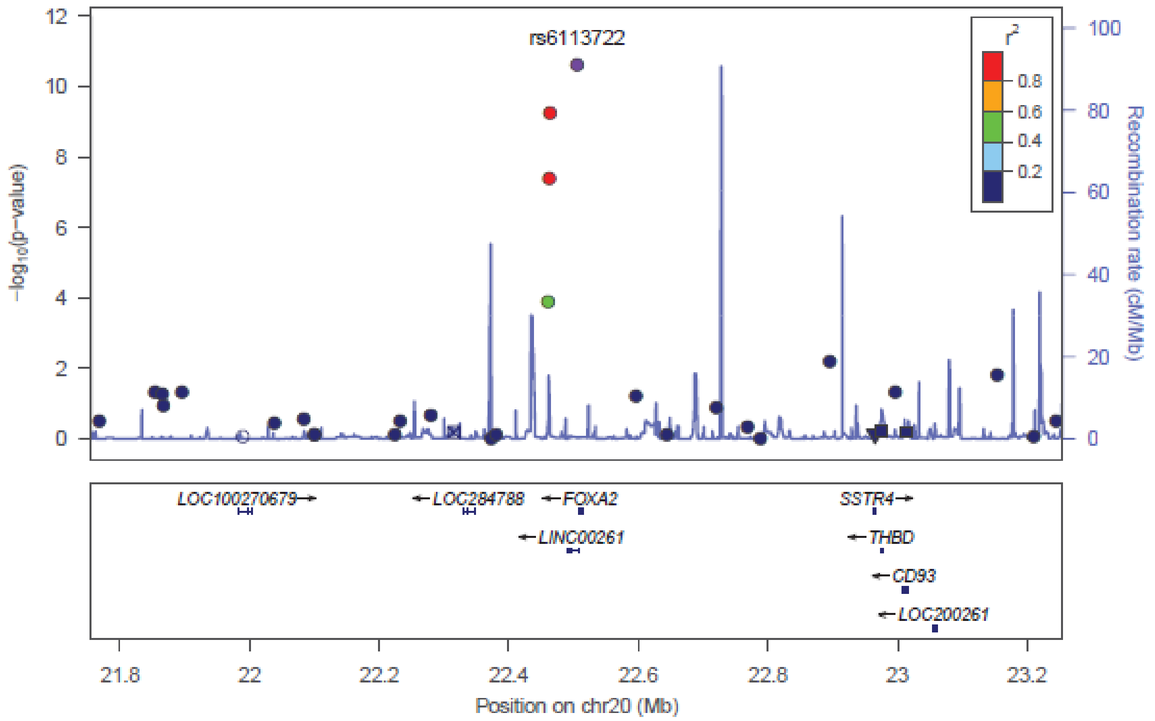

rs3783347
r2
0.8
0.6
0.4
0.2
10
8
6
4
2
0
−log10(p−value)
100
80
60
40
20
0
Recombination rate (cM/Mb)
HHIPL1→
EML1→
EVL→
←WARS
C14orf70→
MEG3→
MIR379→
CYP46A1→
MIR342→
←SLC25A29
←BEGAIN
DLK1→
MIR493→
MIR411→
←DEGS2
MIR345→
MIR2392→
MIR299→
SLC25A47→
MIR770→
MIR380→
99.2
99.4
99.6
99.8
100
100.2
100.4
100.6
Position on chr14 (Mb)

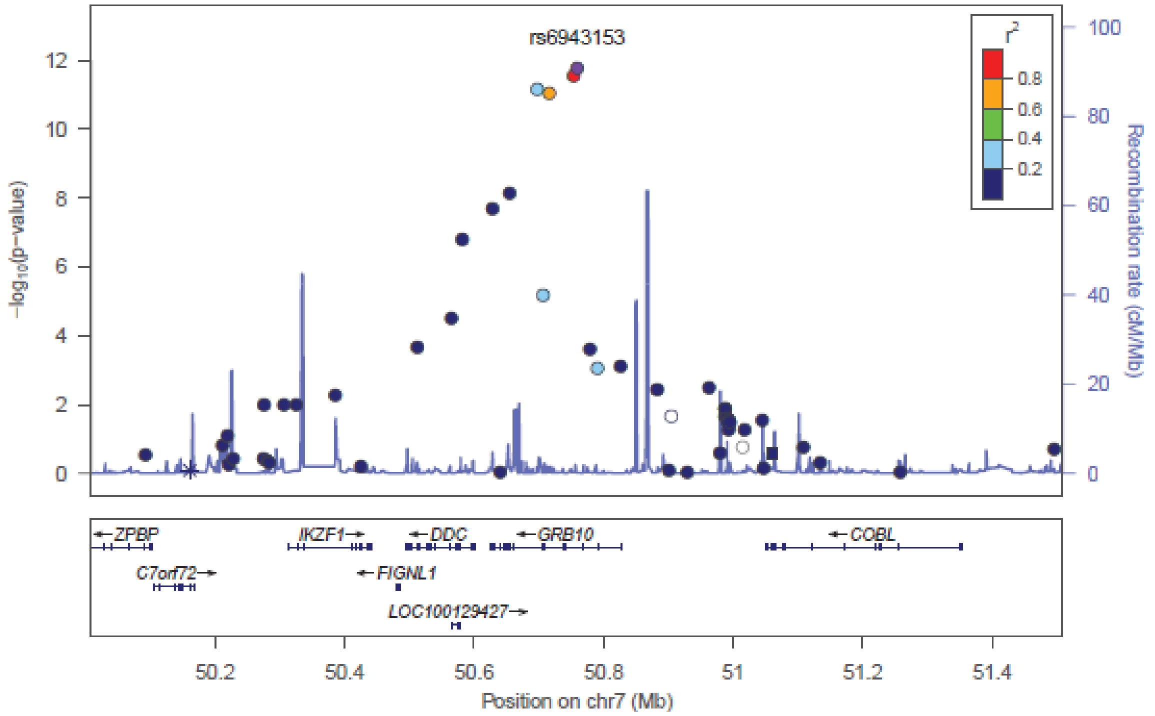

rs3829109
r2
0.8
0.6
0.4
0.2
−log10(p-value)
10
8
6
4
2
0
Recombination rate (cM/Mb)
100
80
60
40
20
0
GLT6D1
CAMSAP1
NACC2
LHX3
DNLZ
NOTCH1
AGPAT2
C9orf172
CLIC3
LCN9
UBAC1
C9orf69
GPSM1
MIR4673
FAM69B
EDF1
ABCA2
SOHLH1
QSOX2
PMPCA
EGFL7
C9orf86
PTGDS
KCNT1
DKFZP434A062
MIR4674
SNHG7
TRAF2
DPP7
137.8
138
138.2
138.4
138.6
138.8
139
Position on chr9 (Mb)

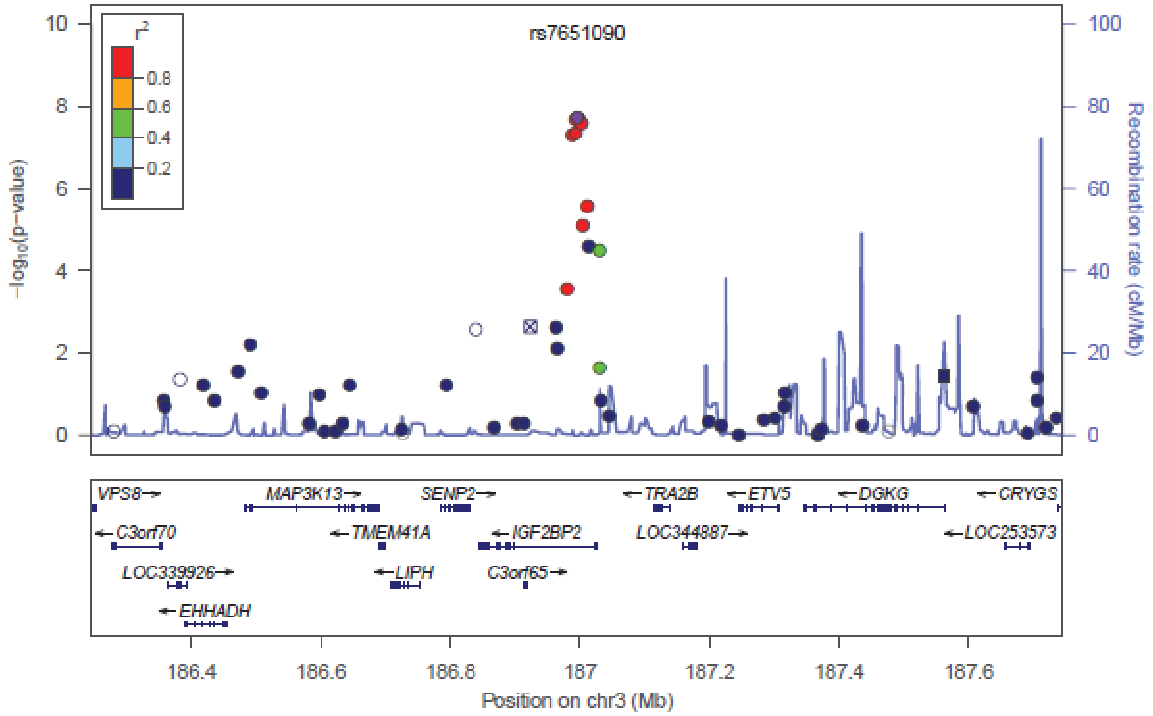

rs4869272
r2
0.8
0.6
0.4
0.2
−log10(p-value)
15
10
5
0
Recombination rate (cM/Mb)
100
80
60
40
20
0
TTC37
SPATA9
GLRX
MIR583
PCSK1
CAST
ERAP2
ARSK
RHOBTB3
ELL2
ERAP1
GPR150
C5orf27
LNPEP
RFESD
95
95.2
95.4
95.6
95.8
96
96.2
Position on chr5 (Mb)

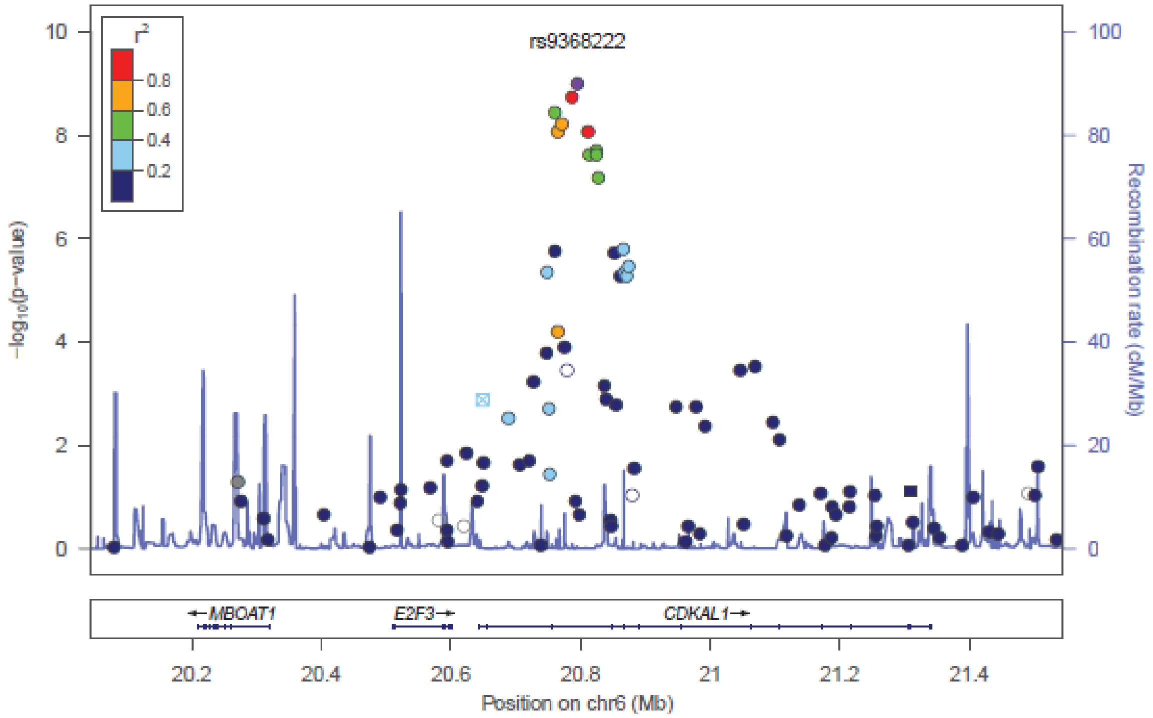

rs576674
r2
0.8
0.6
0.4
0.2
−log10(p−value)
10
8
6
4
2
0
Recombination rate (cM/Mb)
100
80
60
40
20
0
FRY→
←N4BP2L1
PDS5B→
KL→
←STARD13
←ZAR1L
←N4BP2L2
BRCA2→
←CG030
31.8
32
32.2
32.4
32.6
32.8
33
33.2
Position on chr13 (Mb)

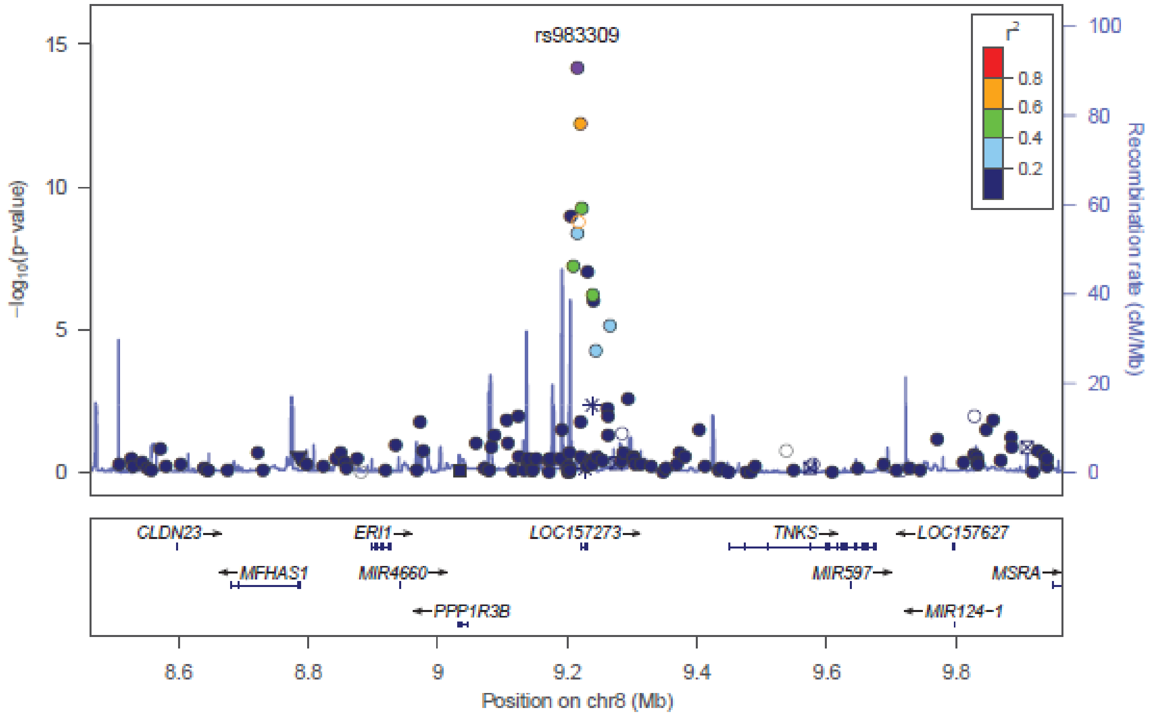

rs6072275
r2
0.8
0.6
0.4
0.2
−log10(p−value)
10
8
6
4
2
0
Recombination rate (cM/Mb)
100
80
60
40
20
0
←MAFB
TOP1→
←ZHX3
←CHD6
PLCG1→
LPIN3→
←EMILIN3
38.6
38.8
39
39.2
39.4
39.6
39.8
Position on chr20 (Mb)

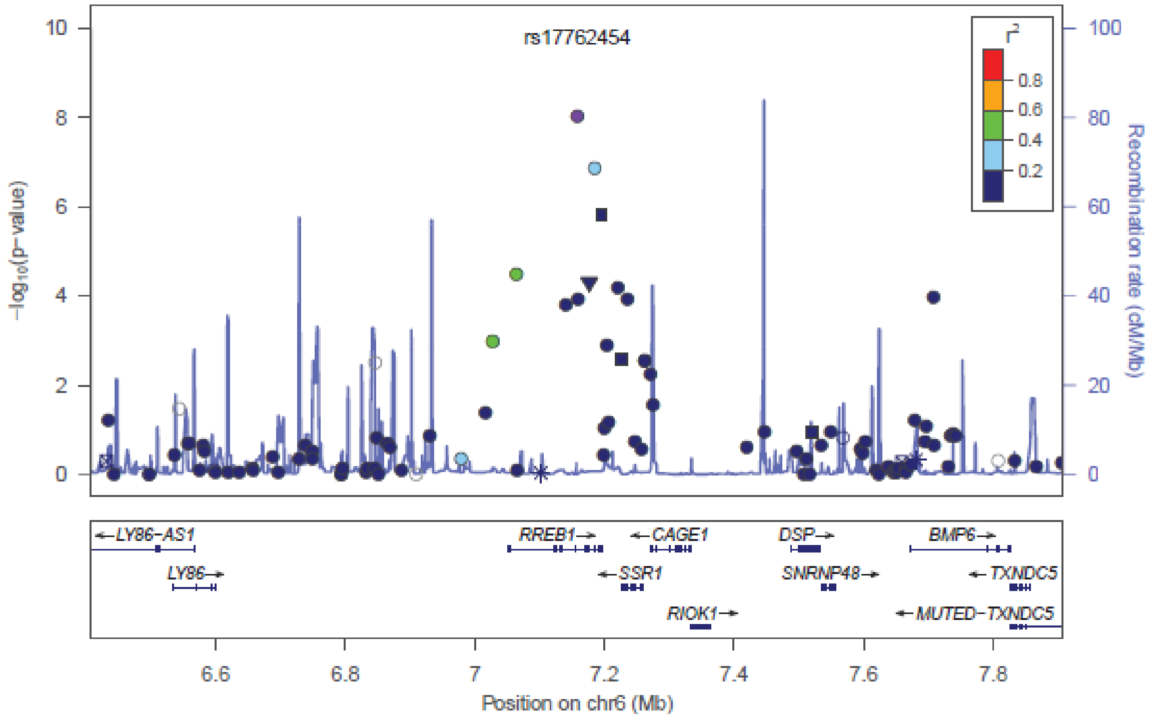

rs6113722
r2
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
LOC100270679
LOC284788
FOXA2
LINC00261
SSTR4
THBD
CD93
LOC200261
Position on chr20 (Mb)
21.8
22
22.2
22.4
22.6
22.8
23
23.2


rs6943153
r2
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
ZPBP
IKZF1
DDC
GRB10
COBL
C7orf72
FIGNL1
LOC100129427
Position on chr7 (Mb)
50.2
50.4
50.6
50.8
51
51.2
51.4


rs7651090
r²
0.8
0.6
0.4
0.2
−log₁₀(p−value)
Recombination rate (cM/Mb)
0
2
4
6
8
10
0
20
40
60
80
100
VPS8→
MAP3K13→
SENP2→
←TRA2B
←ETV5
←DGKG
←CRYGS
←C3orf70
←TMEM41A
←IGF2BP2
LOC344887→
←LOC253573
LOC339926→
←LIPH
C3orf65→
←EHHADH
186.4
186.6
186.8
187
187.2
187.4
187.6
Position on chr3 (Mb)


rs9368222
r²
0.8
0.6
0.4
0.2
−log₁₀(p−value)
Recombination rate (cM/Mb)
0
2
4
6
8
10
0
20
40
60
80
100
←MBOAT1
E2F3→
CDKAL1→
20.2
20.4
20.6
20.8
21
21.2
21.4
Position on chr6 (Mb)


rs983309
r²
0.8
0.6
0.4
0.2
−log₁₀(p−value)
15
10
5
0
Recombination rate (cM/Mb)
100
80
60
40
20
0
CLDN23→
ERI1→
LOC157273→
TNKS→
←LOC157627
←MFHAS1
MIR4660→
MIR597→
MSRA→
←PPP1R3B
←MIR124−1
8.6
8.8
9
9.2
9.4
9.6
9.8
Position on chr8 (Mb)


rs17762454
r²
0.8
0.6
0.4
0.2
−log₁₀(p−value)
10
8
6
4
2
0
Recombination rate (cM/Mb)
100
80
60
40
20
0
←LY86−AS1
RREB1→
←CAGE1
DSP→
BMP6→
LY86→
←SSR1
SNRNP48→
←TXNDC5
RIOK1→
←MUTED−TXNDC5
6.6
6.8
7
7.2
7.4
7.6
7.8
Position on chr6 (Mb)

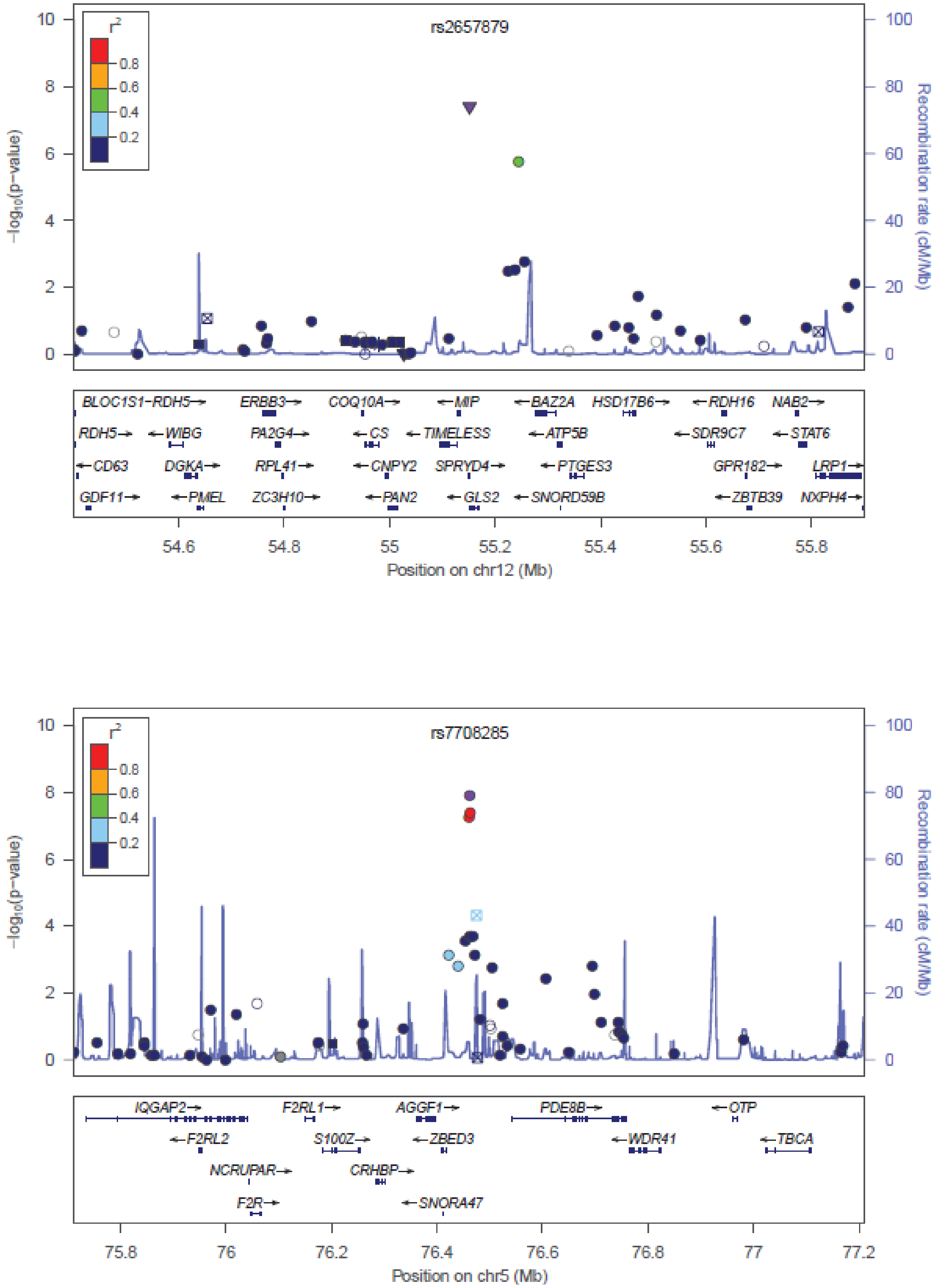
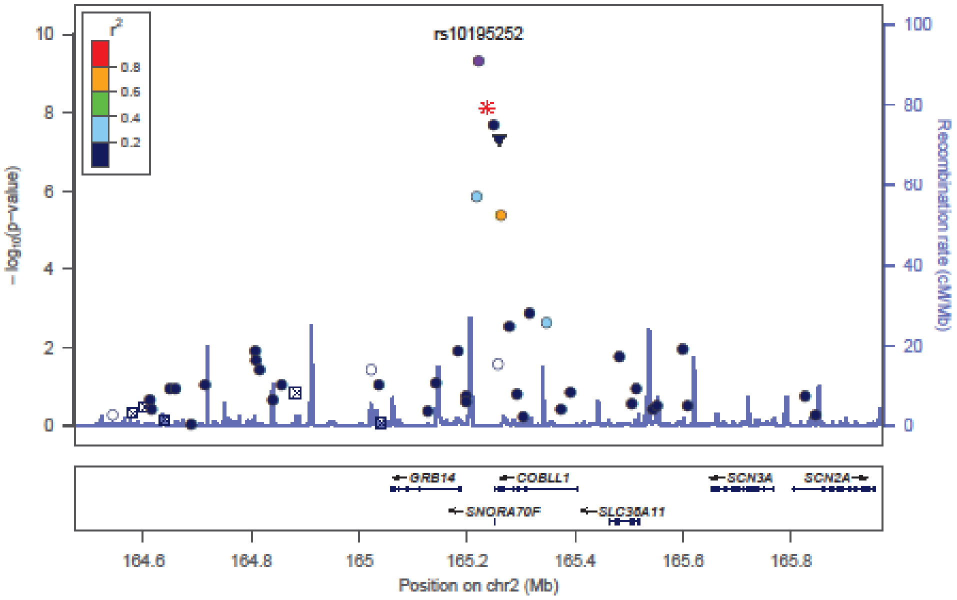
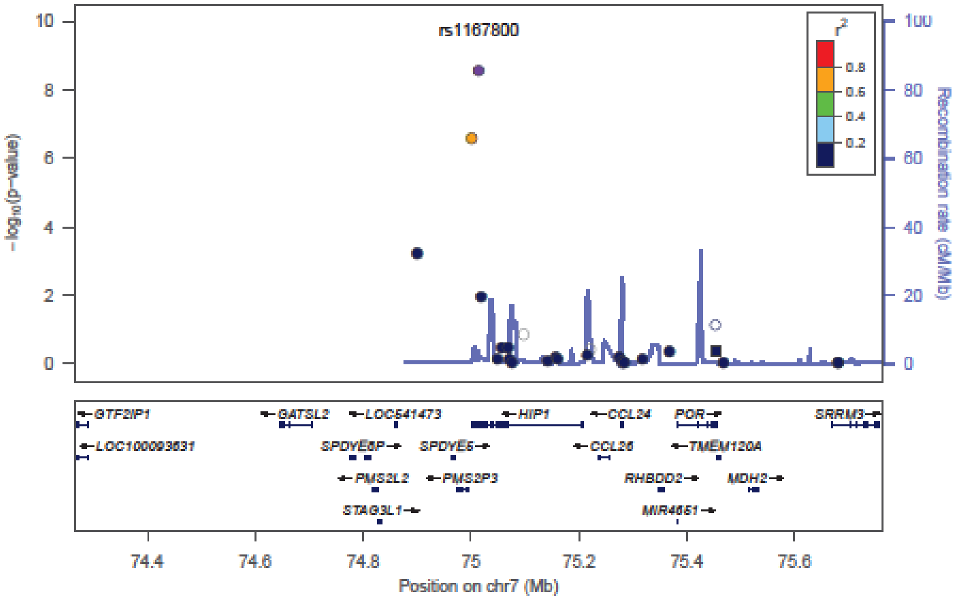
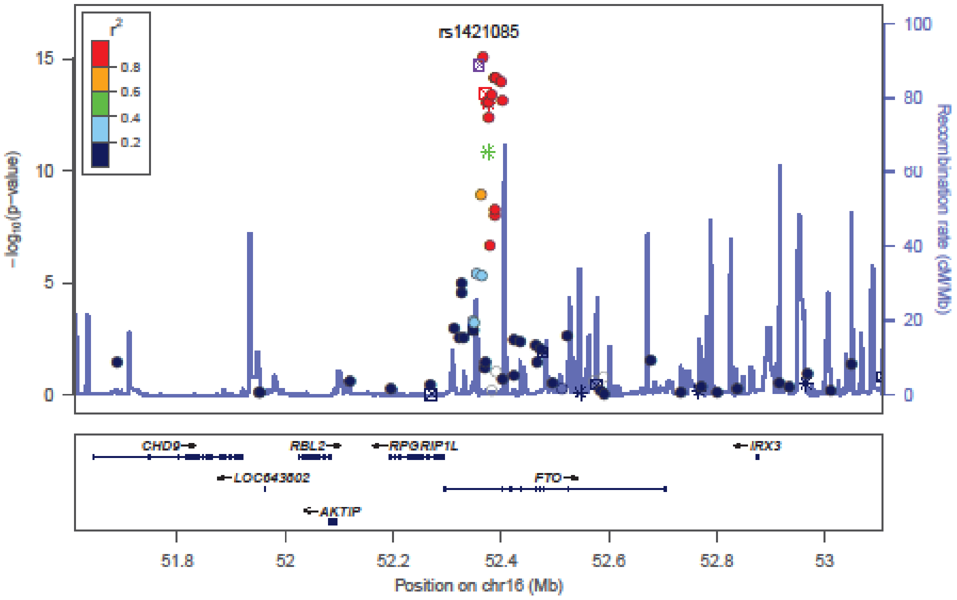
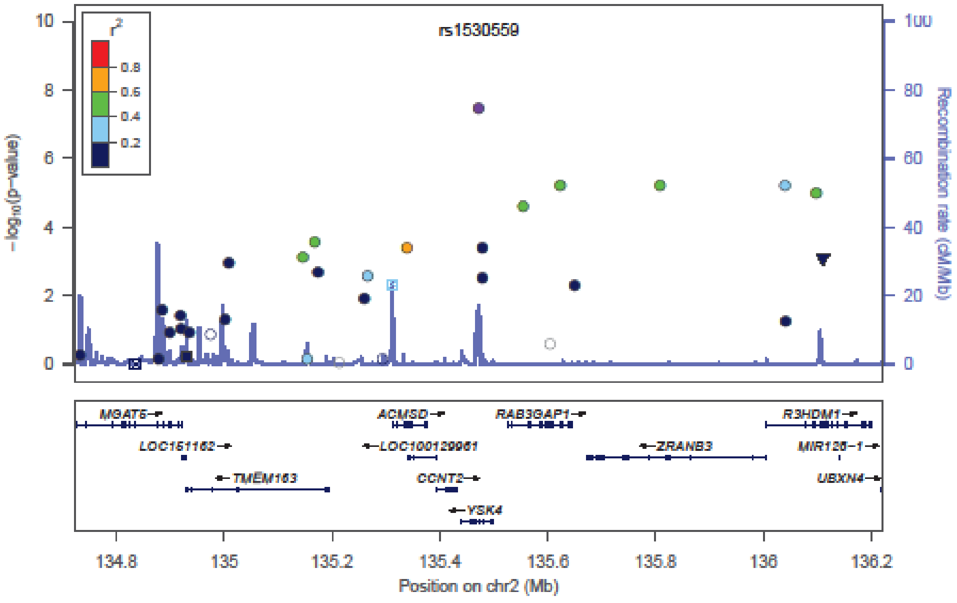


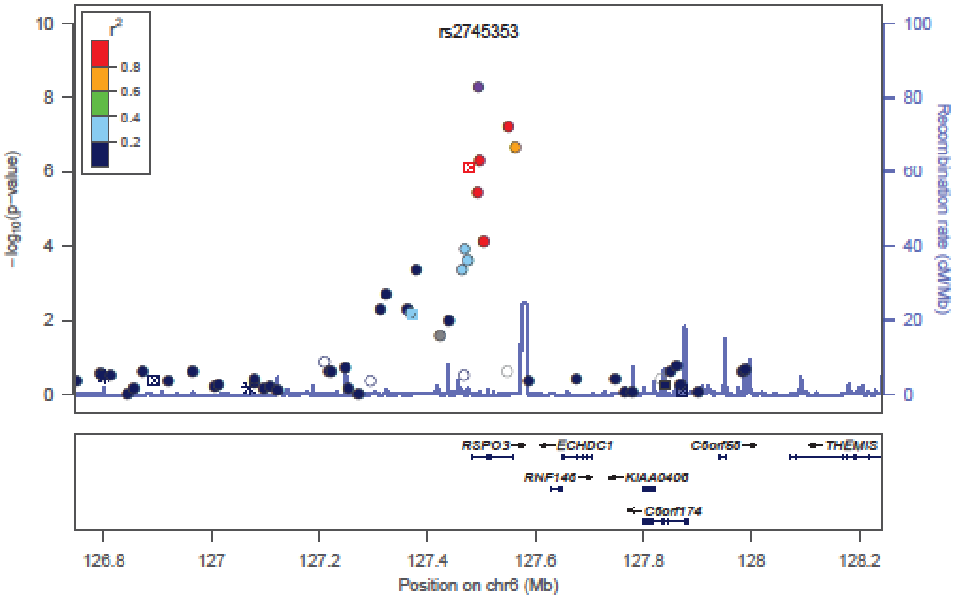
Figure 4.11. Regional association plots for fasting glucose associated loci.

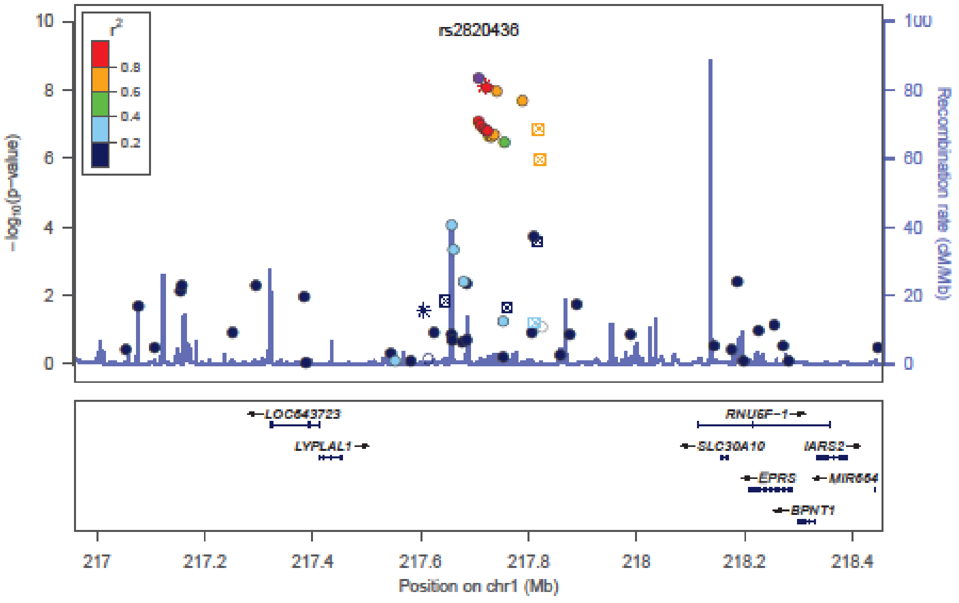

rs10195252
r2
0.8
0.6
0.4
0.2
10
8
6
4
2
0
−log10(p-value)
100
80
60
40
20
0
Recombination rate (cM/Mb)
GRB14
COBLL1
SCN3A
SCN2A
SNORA70F
SLC38A11
164.6
164.8
165
165.2
165.4
165.6
165.8
Position on chr2 (Mb)

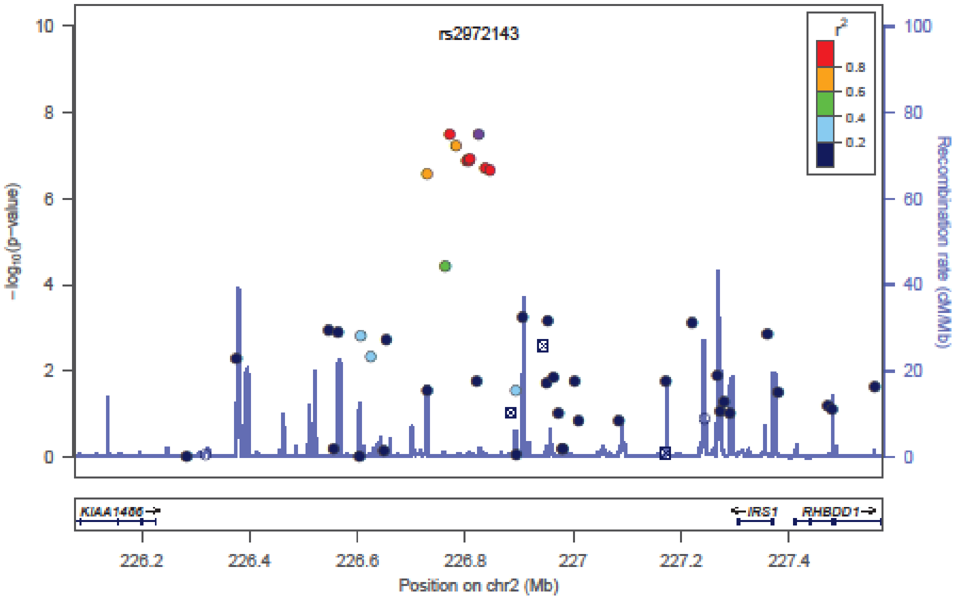

rs1167800
r2
0.8
0.6
0.4
0.2
10
8
6
4
2
0
−log10(p-value)
100
80
60
40
20
0
Recombination rate (cM/Mb)
GTF2IP1
GATSL2
LOC541473
HIP1
CCL24
POR
SRRM3
LOC100093631
SPDYE8P
SPDYE5
CCL26
TMEM120A
PMS2L2
PMS2P3
RHBDD2
MDH2
STAG3L1
MIR4651
74.4
74.6
74.8
75
75.2
75.4
75.6
Position on chr7 (Mb)

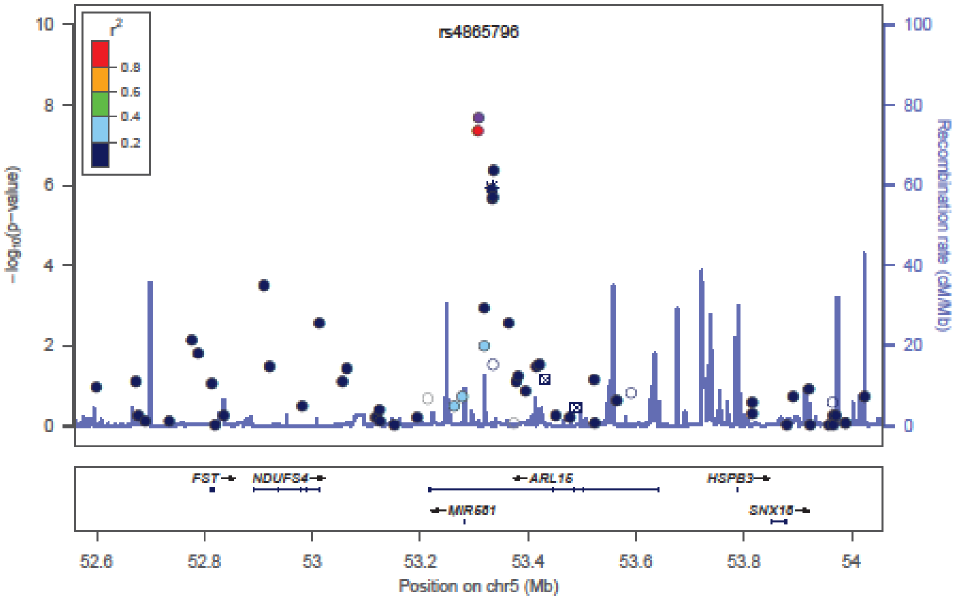

rs1421085
r²
0.8
0.6
0.4
0.2
15
10
5
0
−log₁₀(p-value)
100
80
60
40
20
0
Recombination rate (cM/Mb)
CHD9
RBL2
RPGRIP1L
IRX3
LOC643802
FTO
AKTIP
51.8
52
52.2
52.4
52.6
52.8
53
Position on chr16 (Mb)

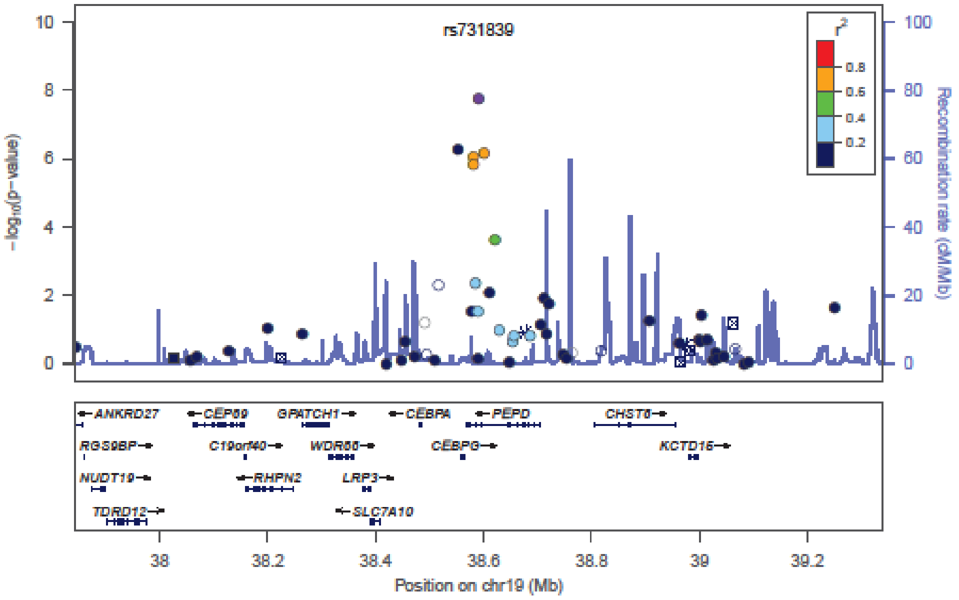

rs1530559
r²
0.8
0.6
0.4
0.2
10
8
6
4
2
0
−log₁₀(p-value)
100
80
60
40
20
0
Recombination rate (cM/Mb)
MGAT5
ACMSD
RAB3GAP1
R3HDM1
LOC151162
LOC100129961
ZRANB3
MIR128-1
TMEM163
CCNT2
UBXN4
YSK4
134.8
135
135.2
135.4
135.6
135.8
136
136.2
Position on chr2 (Mb)

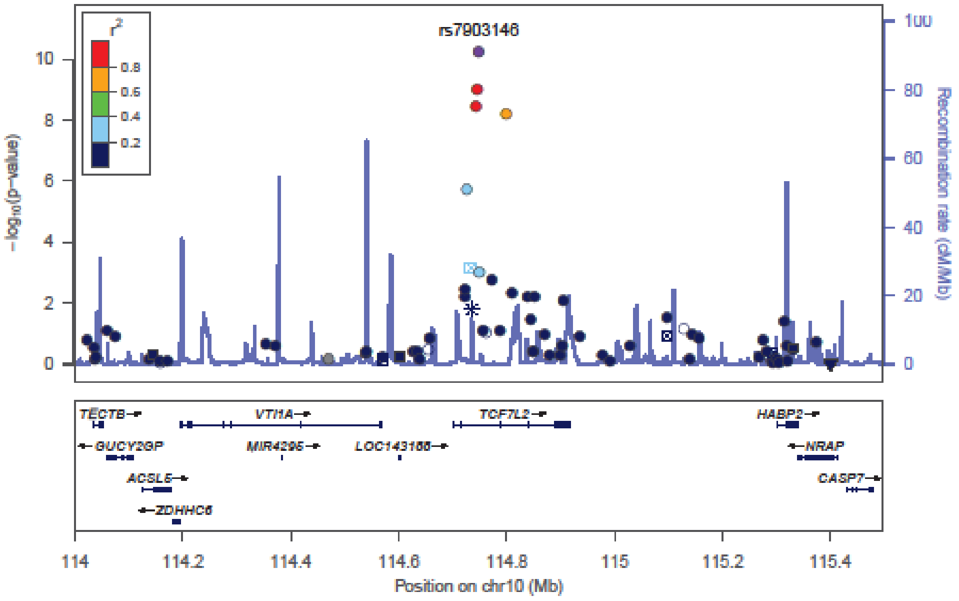

rs2745353
r²
0.8
0.6
0.4
0.2
−log₁₀(p−value)
10
8
6
4
2
0
Recombination rate (cM/Mb)
100
80
60
40
20
0
RSPO3
ECHDC1
C6orf58
THEMIS
RNF146
KIAA0408
C6orf174
126.8
127
127.2
127.4
127.6
127.8
128
128.2
Position on chr6 (Mb)


rs2820436
r²
0.8
0.6
0.4
0.2
−log₁₀(p−value)
10
8
6
4
2
0
Recombination rate (cM/Mb)
100
80
60
40
20
0
LOC643723
LYPLAL1
RNU6F−1
SLC30A10
IARS2
EPRS
MIR664
BPNT1
217
217.2
217.4
217.6
217.8
218
218.2
218.4
Position on chr1 (Mb)


rs2972143
r2
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
0
2
4
6
8
10
0
20
40
60
80
100
KIAA1486
IRS1
RHBDD1
226.2
226.4
226.6
226.8
227
227.2
227.4
Position on chr2 (Mb)


rs4865796
r2
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
0
2
4
6
8
10
0
20
40
60
80
100
FST
NDUFS4
ARL15
MIR581
HSPB3
SNX18
52.6
52.8
53
53.2
53.4
53.6
53.8
54
Position on chr5 (Mb)


rs731839
$-\log_{10}(p\text{-value})$
10
8
6
4
2
0
Recombination rate (cM/Mb)
100
80
60
40
20
0
r^2
0.8
0.6
0.4
0.2
ANKRD27
CEP89
GPATCH1
CEBPA
PEPD
CHST8
RGS9BP
C19orf40
WDR88
CEBPG
KCTD15
NUDT19
RHPN2
LRP3
TDRD12
SLC7A10
38
38.2
38.4
38.6
38.8
39
39.2
Position on chr19 (Mb)


rs7903146
$-\log_{10}(p\text{-value})$
10
8
6
4
2
0
Recombination rate (cM/Mb)
100
80
60
40
20
0
r^2
0.8
0.6
0.4
0.2
TECTB
VTI1A
TCF7L2
HABP2
GUCY2GP
MIR4295
LOC143188
NRAP
ACSL5
CASP7
ZDHHC6
114
114.2
114.4
114.6
114.8
115
115.2
115.4
Position on chr10 (Mb)

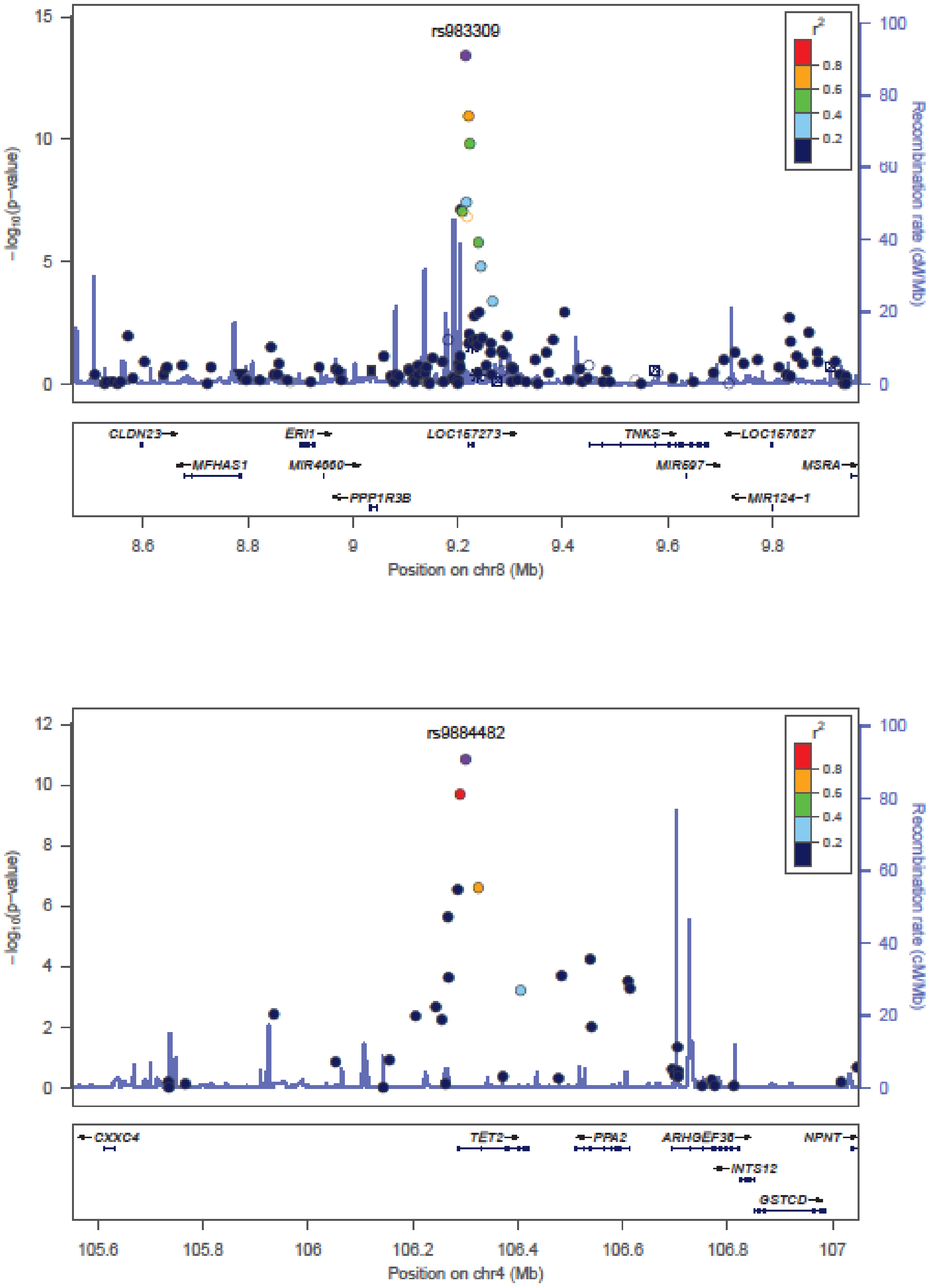
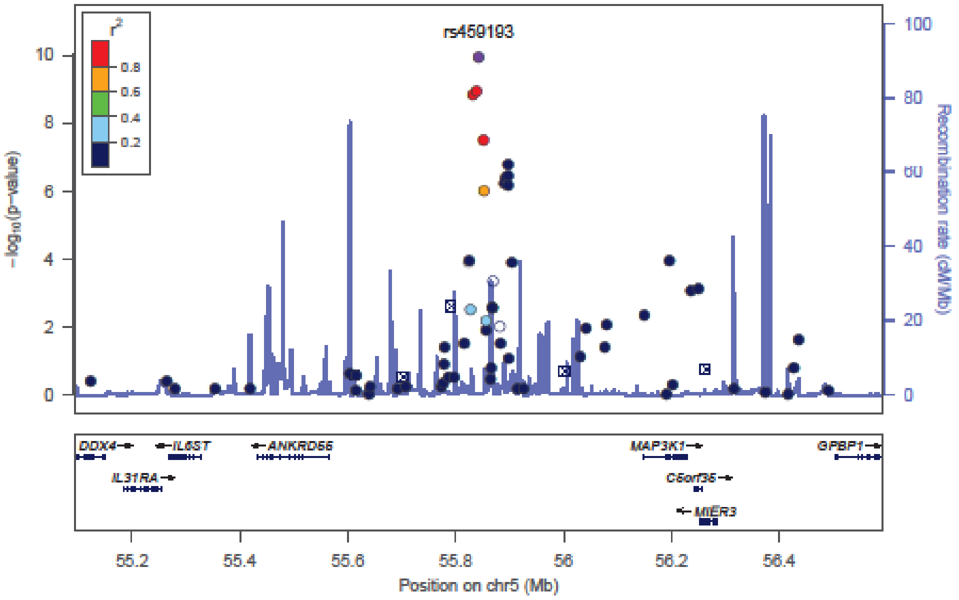
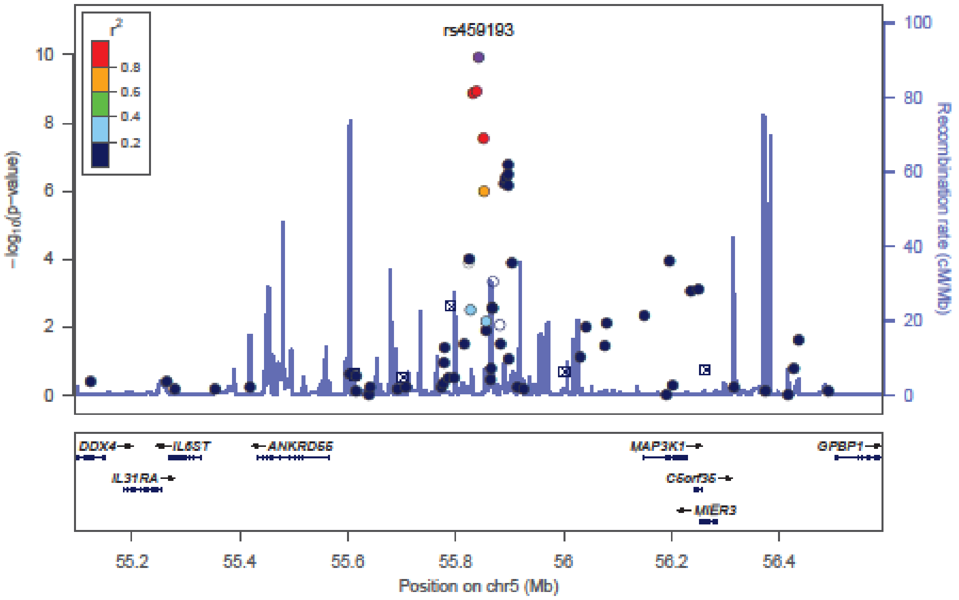


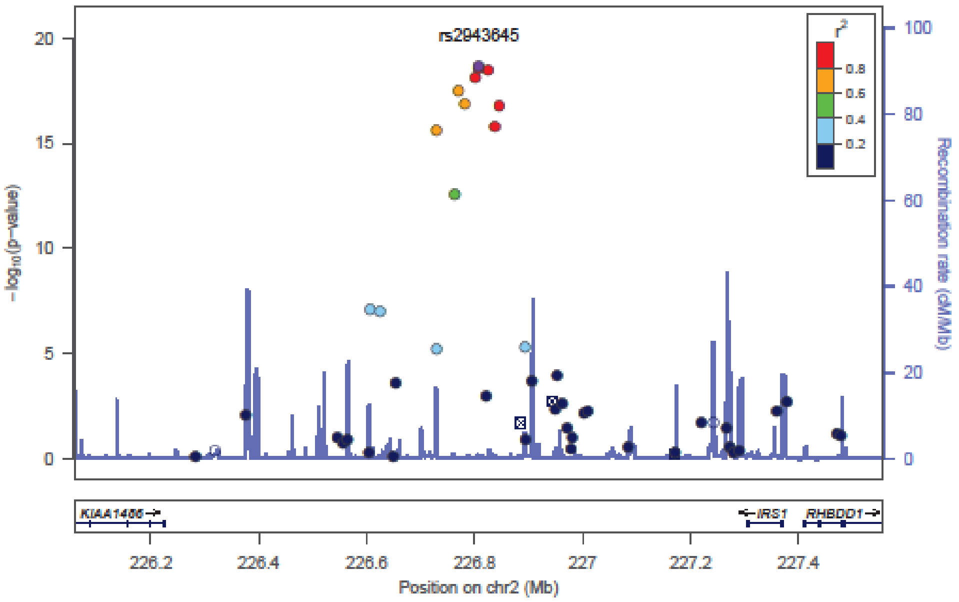
Figure 4.12. Regional association plots for fasting insulin associated loci.

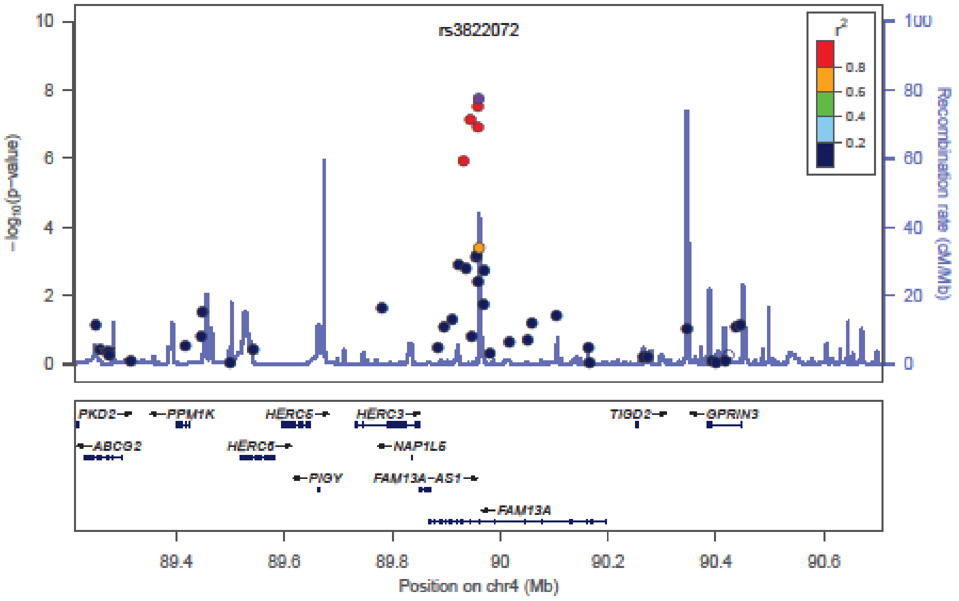


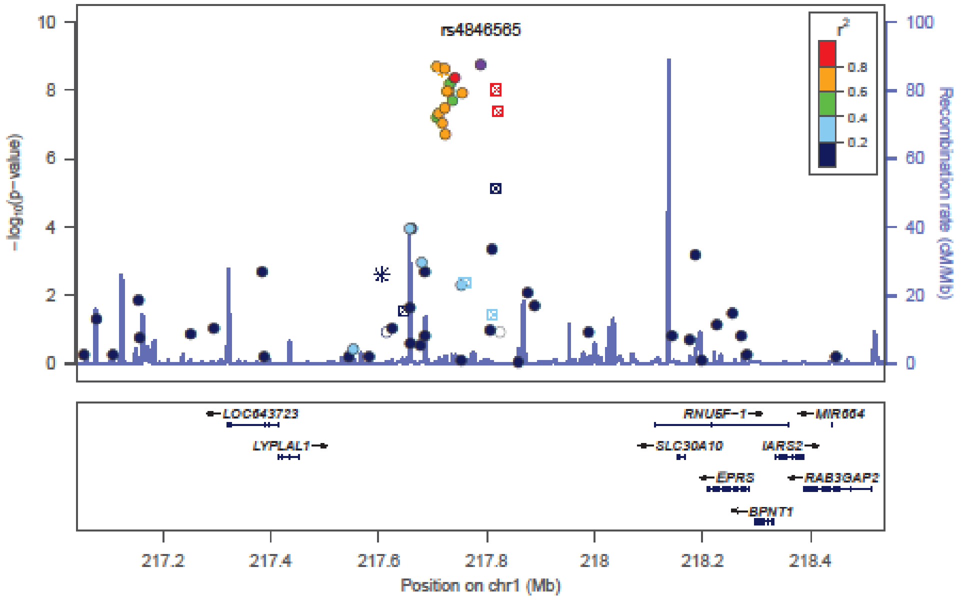
rs459193
r²
0.8
0.6
0.4
0.2
−log₁₀(p−value)
0
2
4
6
8
10
Recombination rate (cM/Mb)
0
20
40
60
80
100
DDX4
IL6ST
ANKRD55
IL31RA
MAP3K1
C5orf35
MIER3
GPBP1
55.2
55.4
55.6
55.8
56
56.2
56.4
Position on chr5 (Mb)

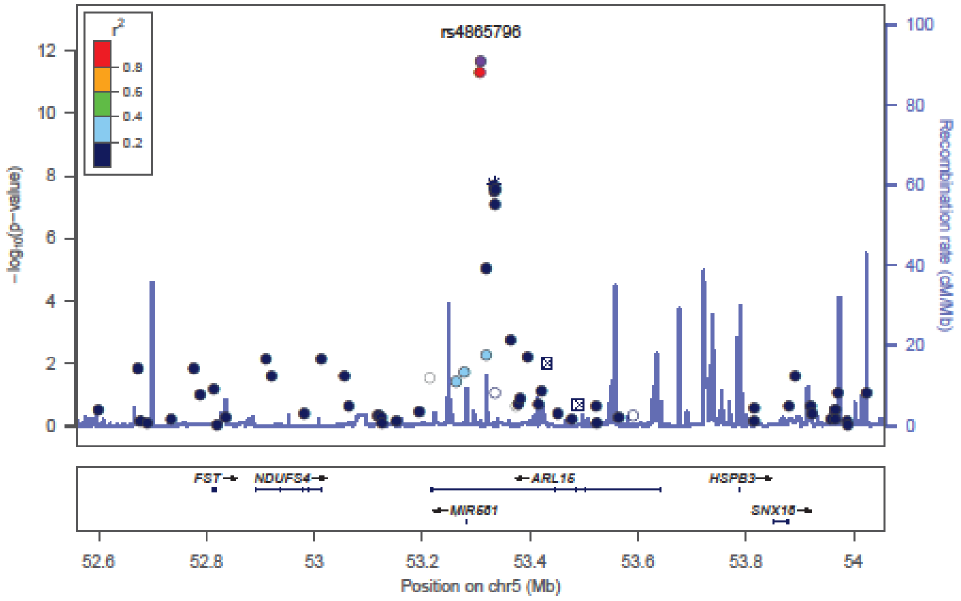


rs459193
r²
0.8
0.6
0.4
0.2
−log₁₀(p−value)
0
2
4
6
8
10
Recombination rate (cM/Mb)
0
20
40
60
80
100
DDX4
IL6ST
ANKRD55
IL31RA
MAP3K1
C5orf35
MIER3
GPBP1
55.2
55.4
55.6
55.8
56
56.2
56.4
Position on chr5 (Mb)


rs2943645
r2
0.8
0.6
0.4
0.2
−log10(p−value)
0
5
10
15
20
Recombination rate (cM/Mb)
0
20
40
60
80
100
KIAA1486
IRS1
RHBDD1
226.2
226.4
226.6
226.8
227
227.2
227.4
Position on chr2 (Mb)


rs3822072
r2
0.8
0.6
0.4
0.2
−log10(p−value)
0
2
4
6
8
10
Recombination rate (cM/Mb)
0
20
40
60
80
100
PKD2
PPM1K
HERC5
HERC3
TIGD2
GPRIN3
ABCG2
HERC6
NAP1L5
PIGY
FAM13A−AS1
FAM13A
89.4
89.6
89.8
90
90.2
90.4
90.6
Position on chr4 (Mb)


rs4846565
r^2
0.8
0.6
0.4
0.2
10
8
6
4
2
0
-log$_{10}$(p-value)
100
80
60
40
20
0
Recombination rate (cM/Mb)
LOC643723
LYPLAL1
RNU5F-1
MIR664
SLC30A10
IARS2
EPRS
RAB3GAP2
BPNT1
217.2
217.4
217.6
217.8
218
218.2
218.4
Position on chr1 (Mb)


rs4865796
r^2
0.8
0.6
0.4
0.2
12
10
8
6
4
2
0
-log$_{10}$(p-value)
100
80
60
40
20
0
Recombination rate (cM/Mb)
FST
NDUFS4
ARL15
HSPB3
MIR581
SNX18
52.6
52.8
53
53.2
53.4
53.6
53.8
54
Position on chr5 (Mb)

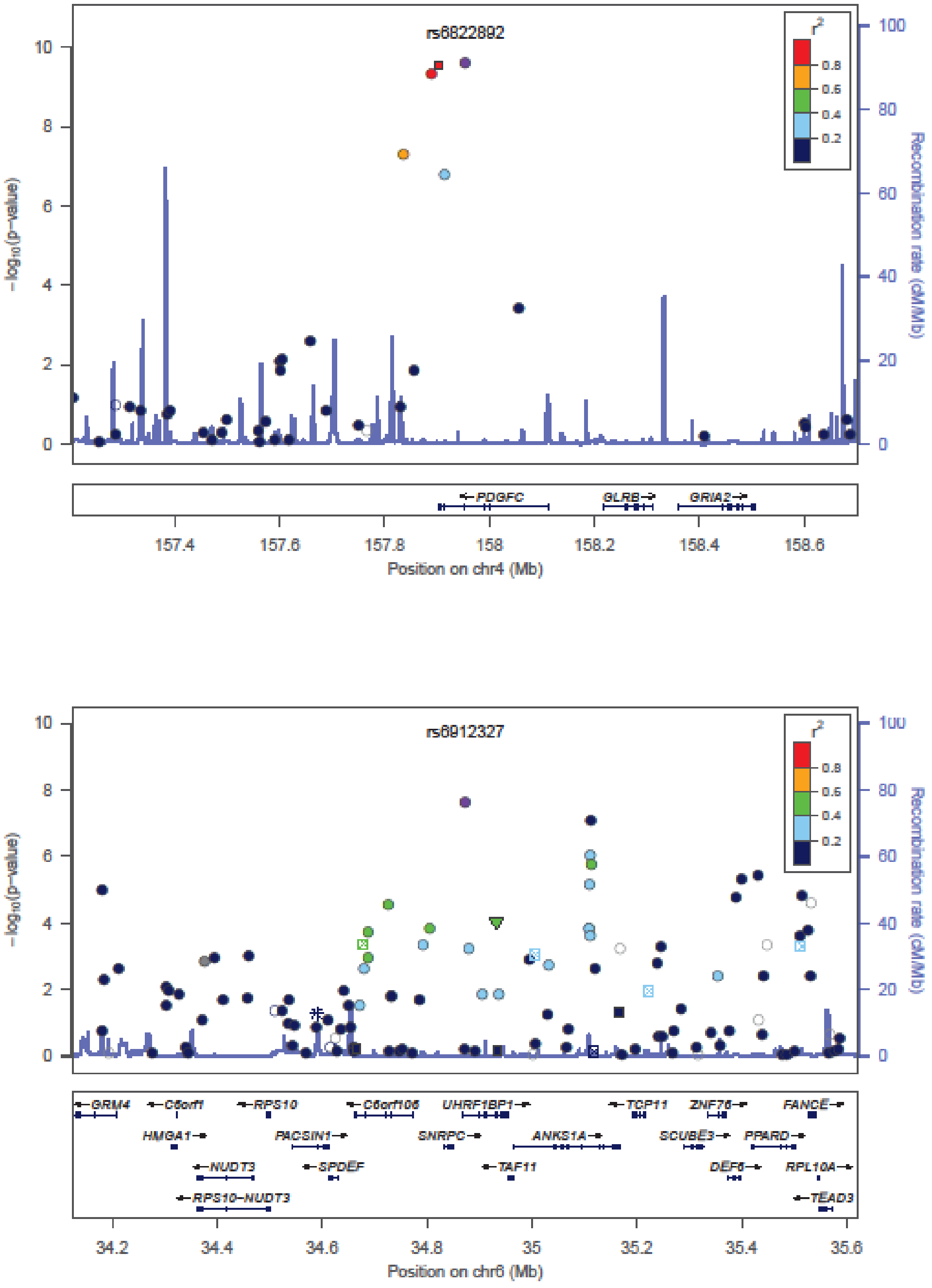
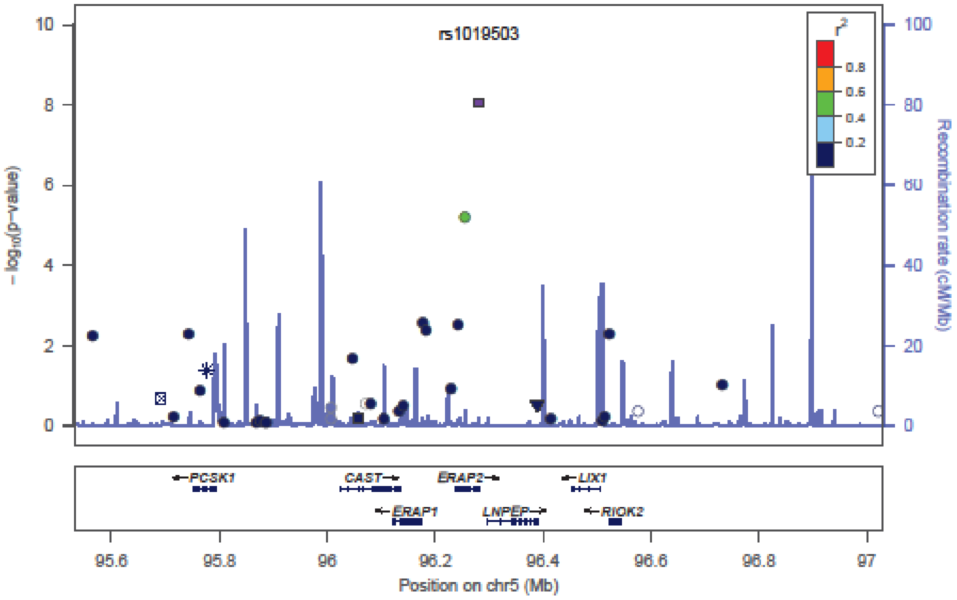
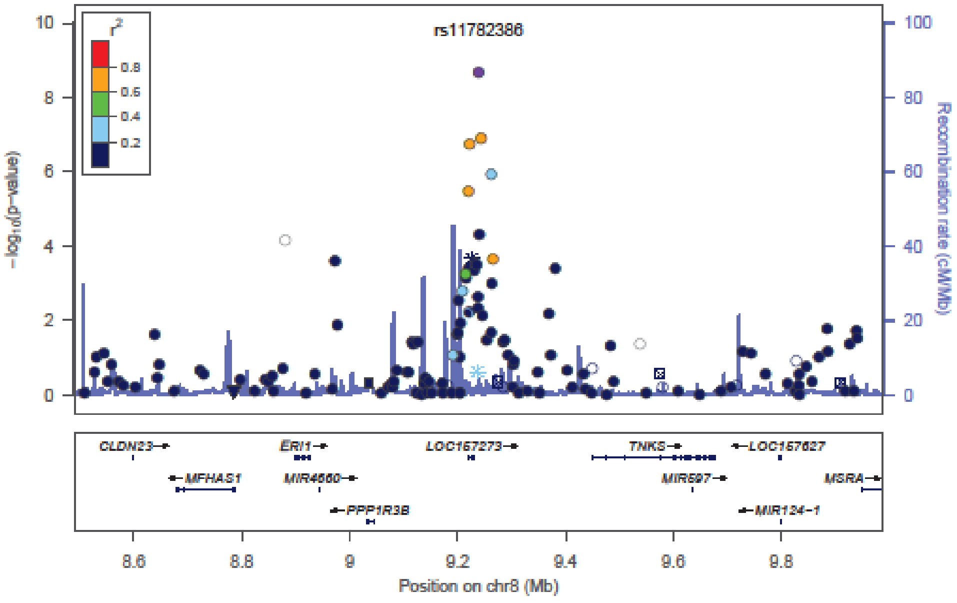


Figure 4.13. Regional association plots for fasting insulin (adjusted for BMI) associated loci.


rs1019503
r²
0.8
0.6
0.4
0.2
– log₁₀(p-value)
0
2
4
6
8
10
Recombination rate (cM/Mb)
0
20
40
60
80
100
PCSK1
CAST
ERAP2
LIX1
ERAP1
LNPEP
RIOK2
95.6
95.8
96
96.2
96.4
96.6
96.8
97
Position on chr5 (Mb)


rs11782386
r²
0.8
0.6
0.4
0.2
– log₁₀(p-value)
0
2
4
6
8
10
Recombination rate (cM/Mb)
0
20
40
60
80
100
CLDN23
ERI1
LOC157273
TNKS
LOC157627
MFHAS1
MIR4660
MIR597
MSRA
PPP1R3B
MIR124-1
8.6
8.8
9
9.2
9.4
9.6
9.8
Position on chr8 (Mb)

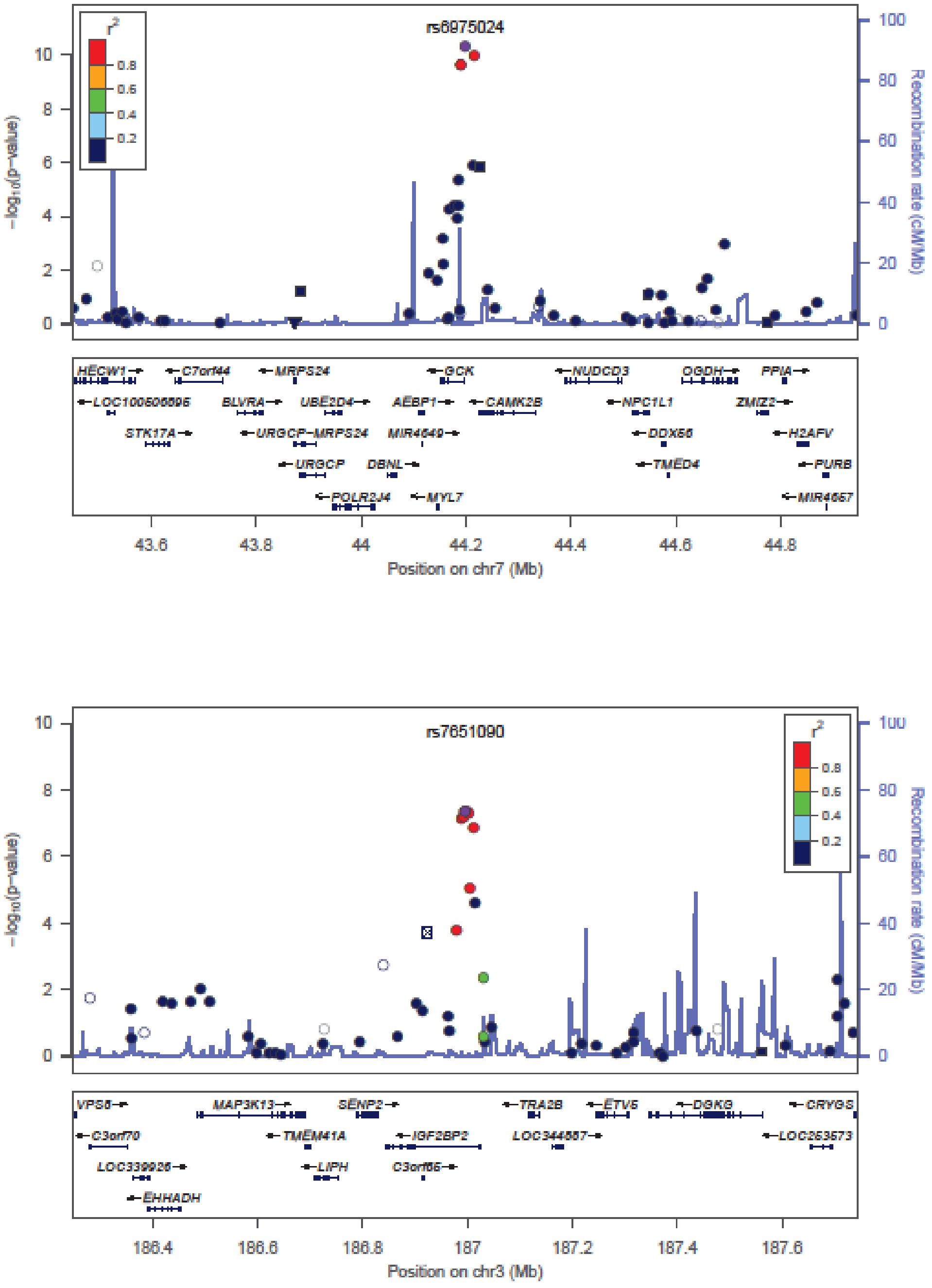


Figure 4.14. Regional association plots for 2-h glucose associated loci.

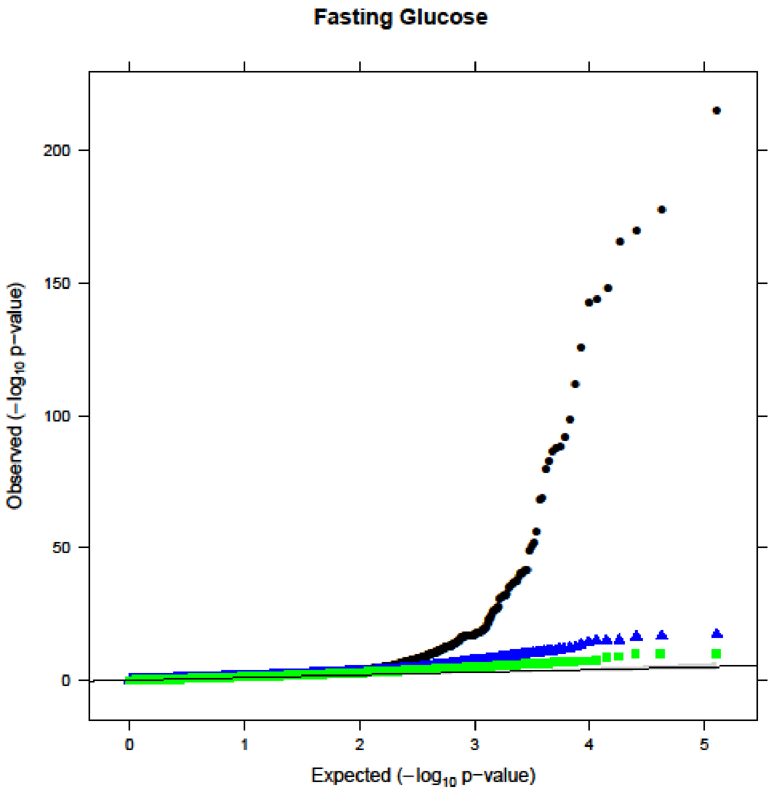


Figure 4.15. QQ plot for fasting glucose. Black dots show all Metabochip follow-up SNPs, Blue triangles show observed associations after removal of previously established signals, while green squares shows observed associations after removal of all genome-wide significant SNPs.

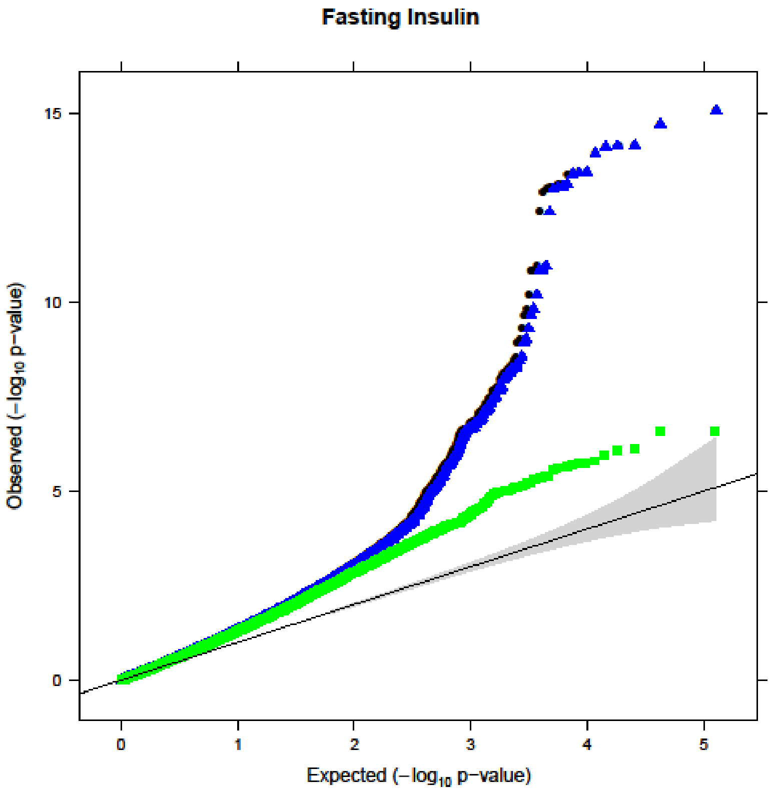


Figure 4.16. QQ plot for fasting insulin. See Figure 4.15 for legend.

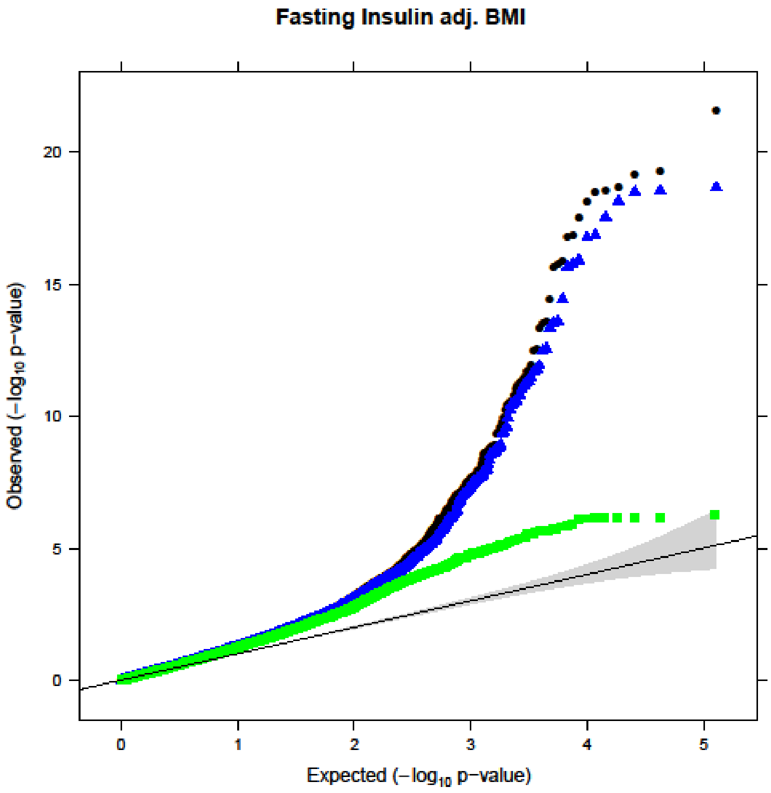


Figure 4.17. QQ plot for fasting insulin adjusted for BMI. See Figure 4.15 for legend.

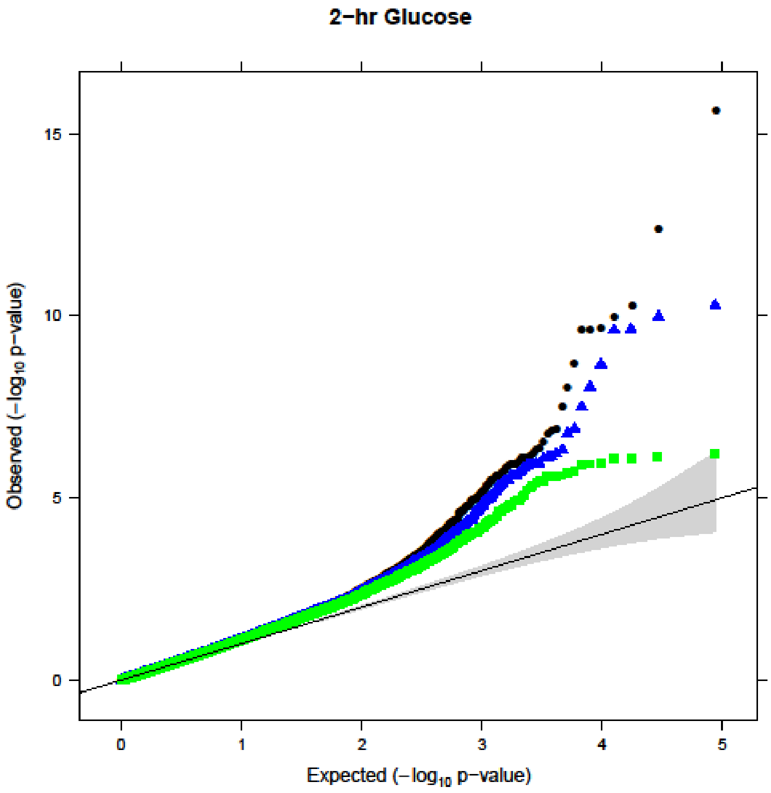


Figure 4.18. QQ plot for 2-h glucose. See Figure 4.15 for legend.

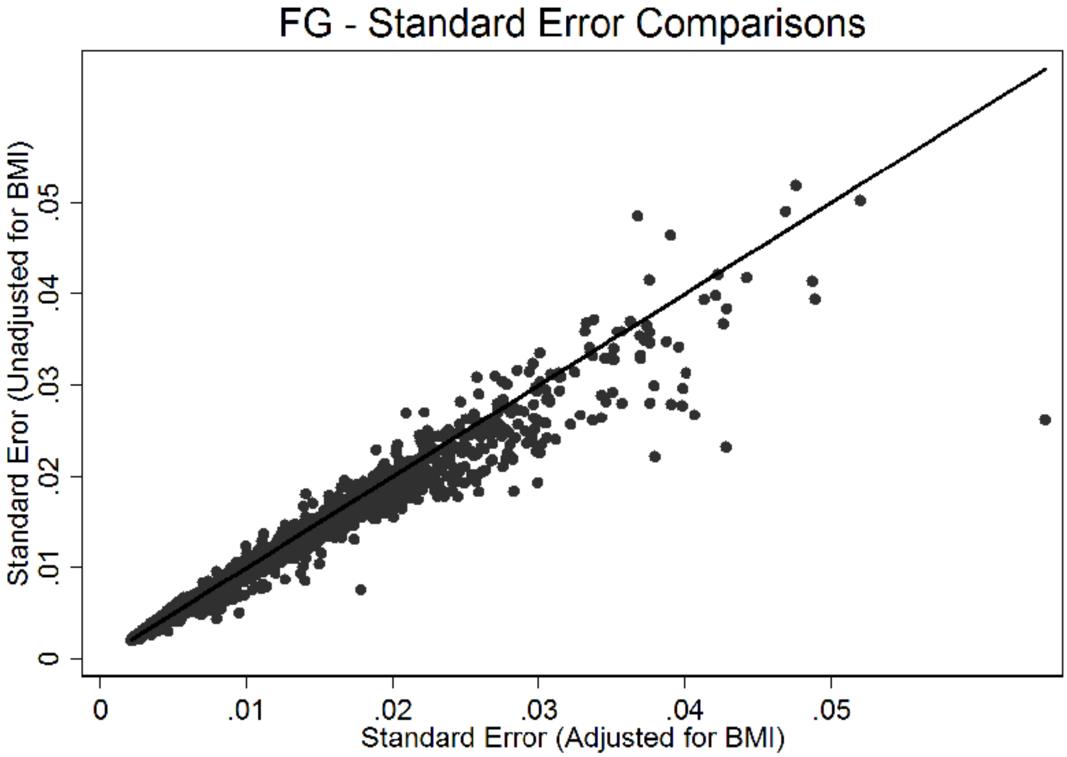


Figure 4.19. Comparison of the standard errors of FG effect estimates of all~66,000 SNPs between models unadjusted (uniform) and adjusted for BMI.

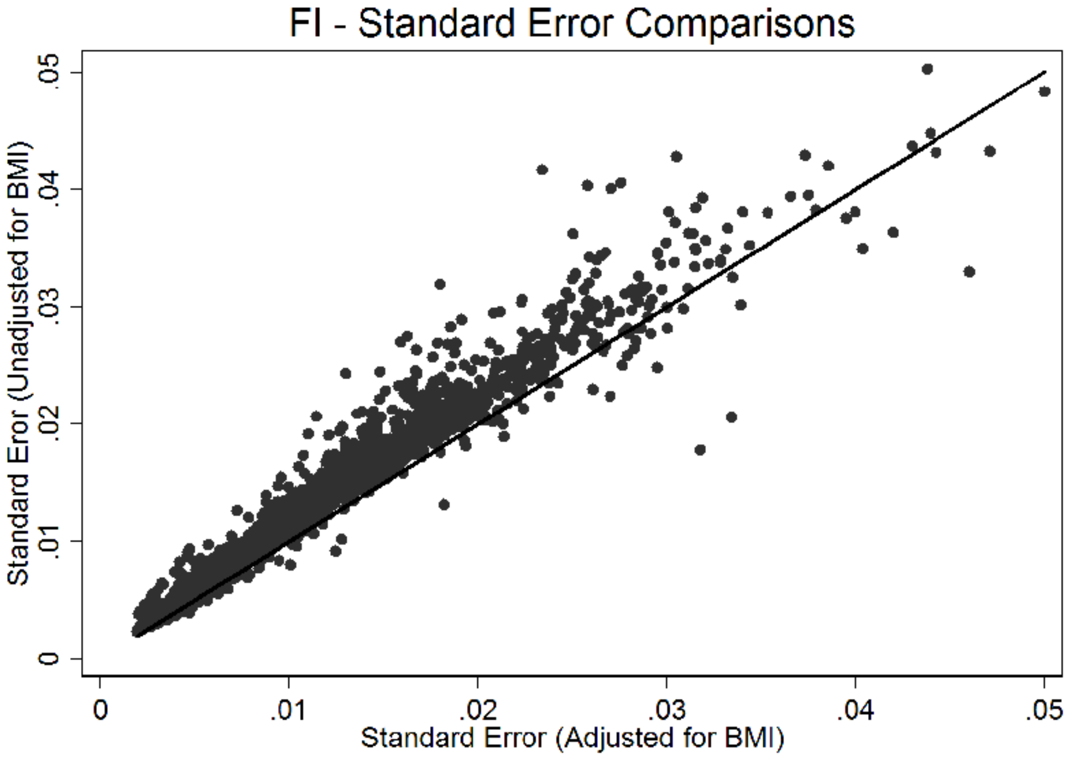


Figure 4.20. Comparison of the standard errors of FI effect estimates of all ~66,000 SNPs between models unadjusted (uniform) and adjusted for BMI. It can be observed that the standard errors are systematically smaller for BMI adjusted models.

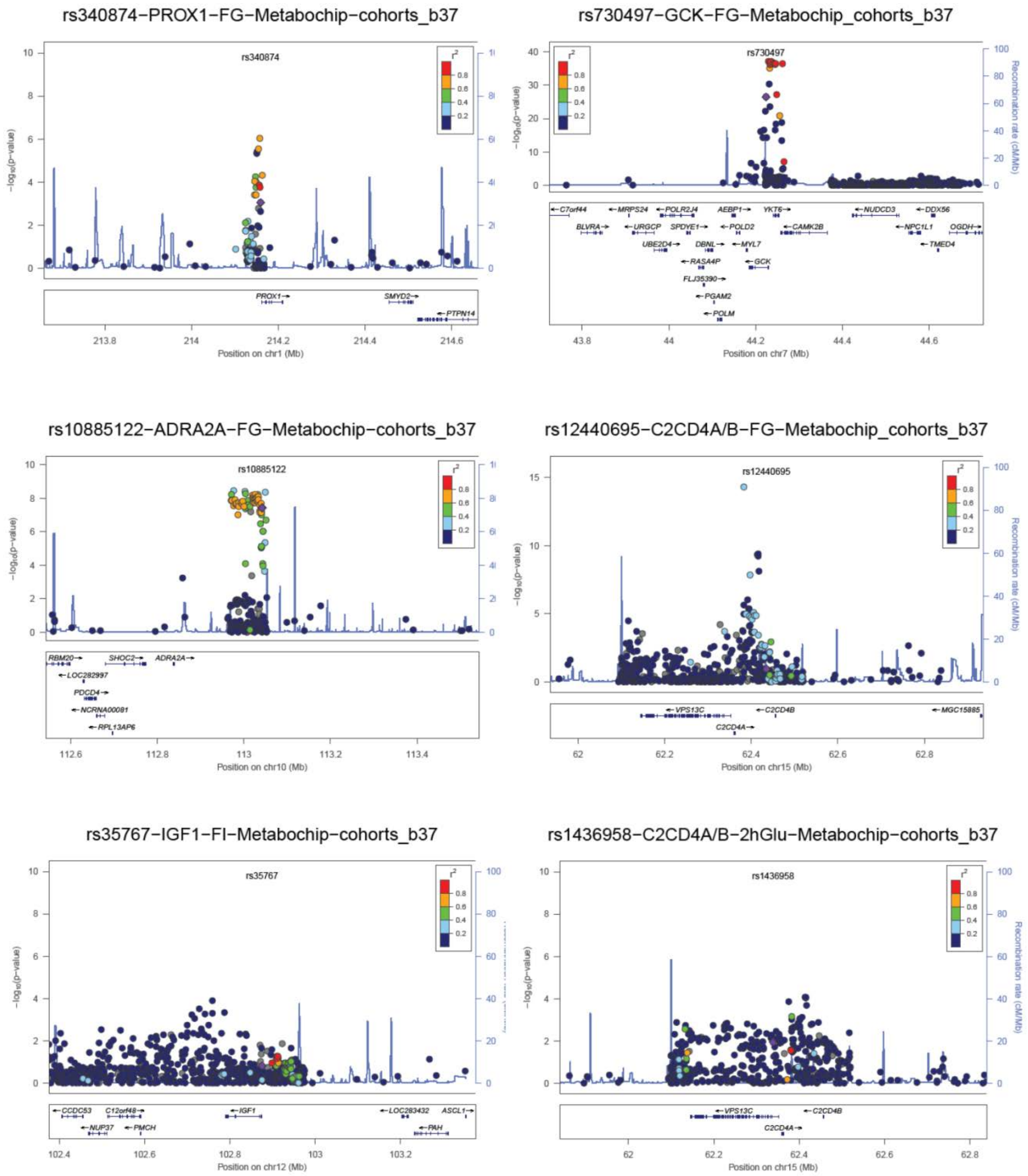


Figure 4.21. Regional plot of fine-mapping data for FG at PROX1, GCK, ADRA2A, VPS13C/C2CD4A/B, and for FI at IGF1 as well as VPS13C/C2CD4A/B for 2hGlu. Previous lead SNPs are shown in purple, and it can be seen that a number of SNPs are more significant than that previously considered to be the lead. New lead SNPs are shown in Table ST4.

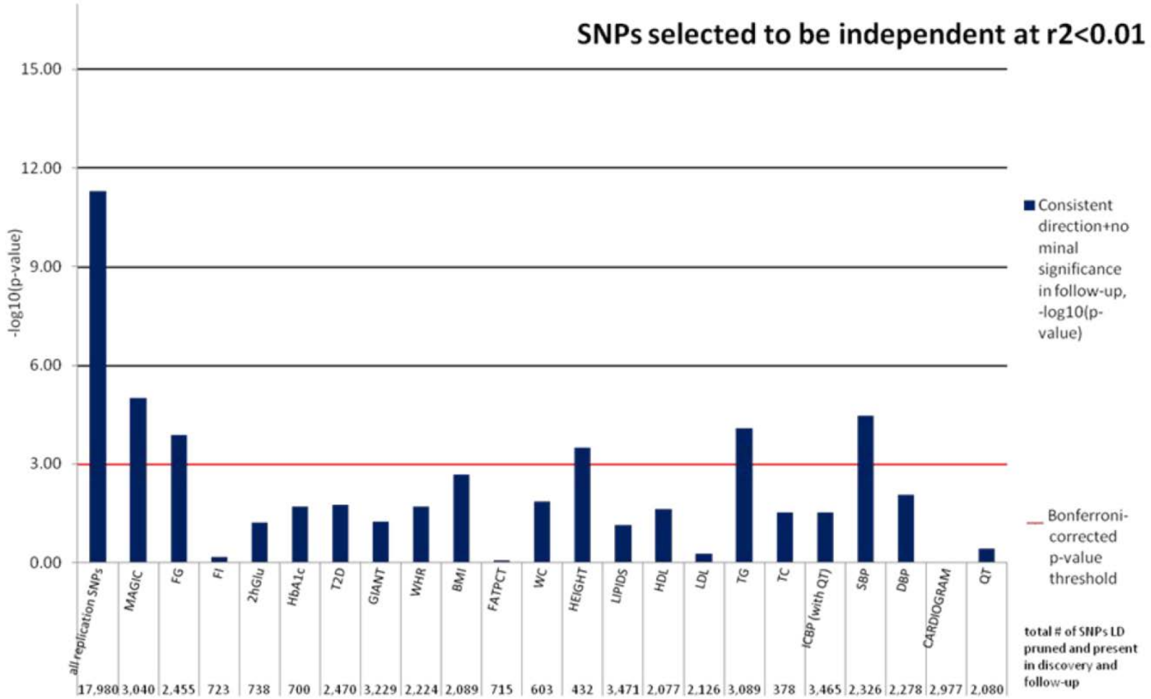


Figure 4.22. FG association directional consistency. SNP lists submitted by each consortium are detailed on the x-axis and on the y-axis -log10 p-values for the binomial tests of 1) consistent direction of FG association between the discovery and follow-up studies in blue, and 2) consistent direction and nominal significance (P< 0.05) in follow-up studies alone in red.

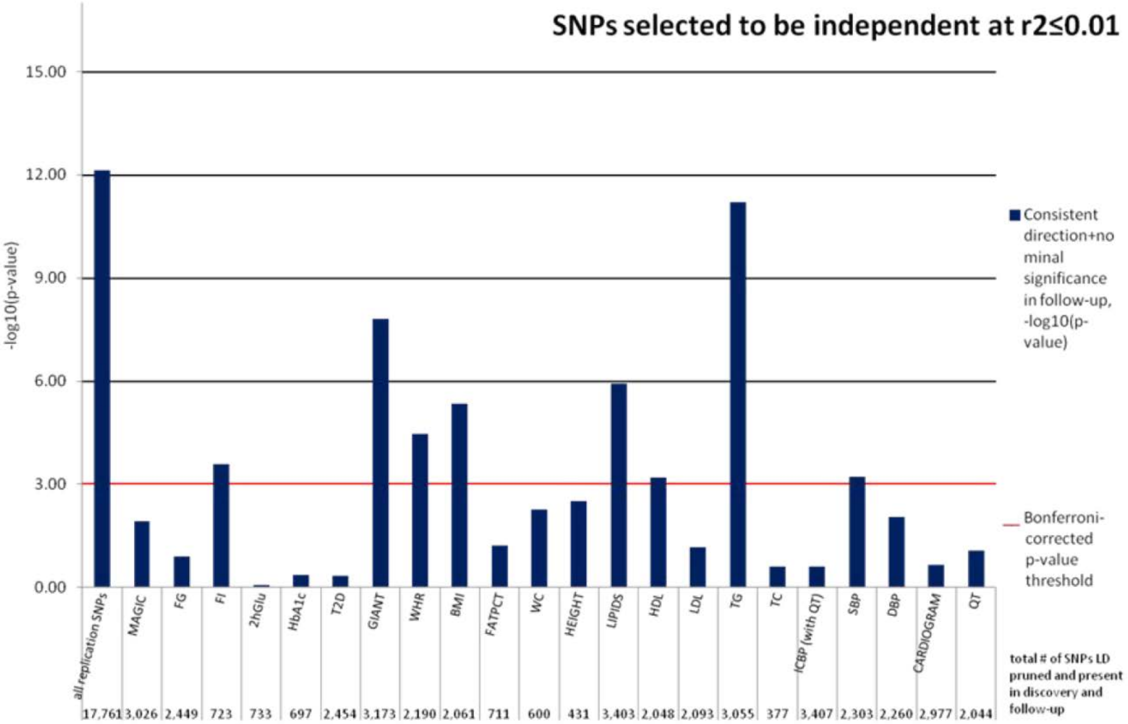


Figure 4.23. FI association directional consistency. SNP lists submitted by each consortium are detailed on the x-axis and on the y-axis -log10 p-values for the binomial tests of 1) consistent direction of FI association between the discovery and follow-up studies in blue, and 2) consistent direction and nominal significance (P< 0.05) in follow-up studies alone in red.

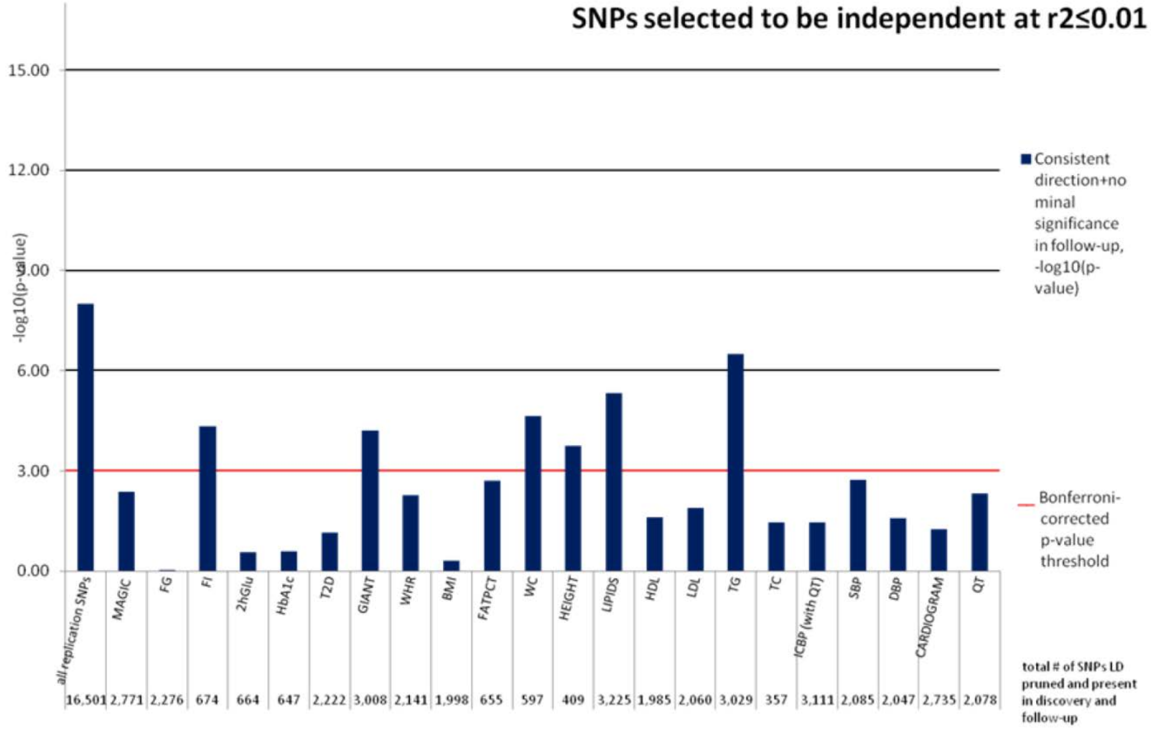


Figure 4.24. FI (BMI-adjusted) association directional consistency. SNP lists submitted by each consortium are detailed on the x-axis and on the y-axis -log10 p-values for the binomial tests of 1) consistent direction of FI association between the discovery and follow-up studies in blue, and 2) consistent direction and nominal significance (P< 0.05) in follow-up studies alone in red.

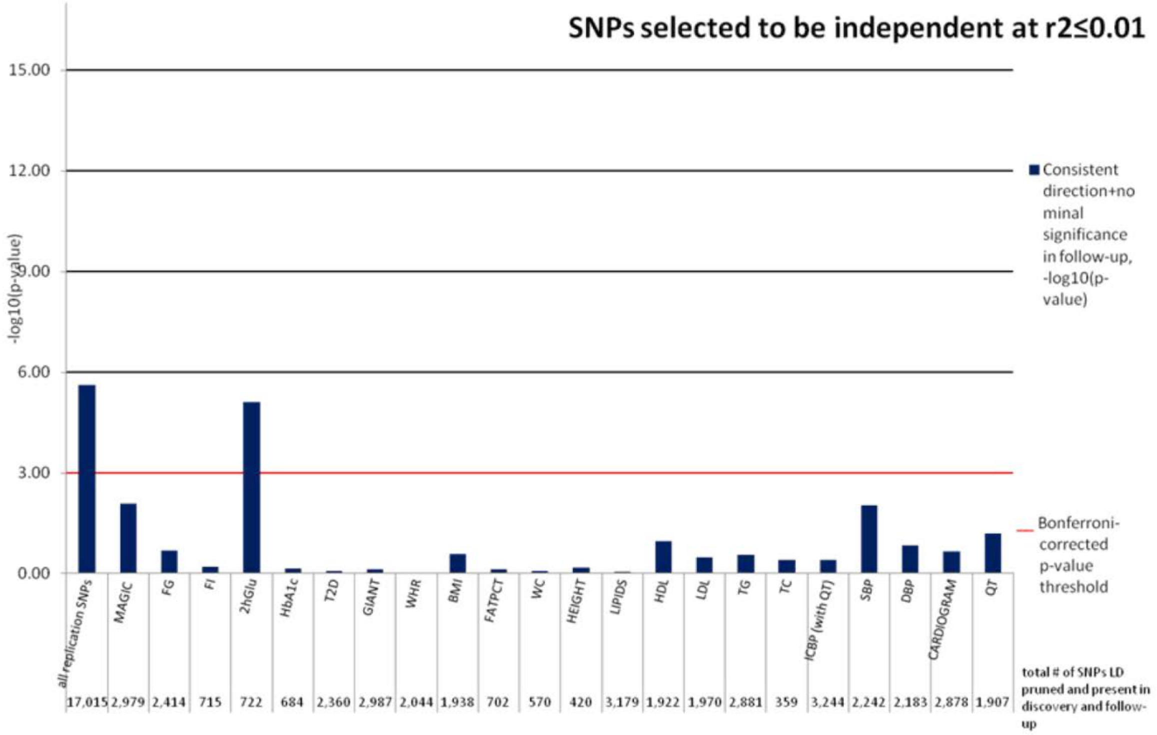


Figure 4.25. 2hGlu signal enrichment. SNP lists submitted by each consortium are detailed on the x-axis and on the y-axis -log10 p-values for the binomial tests of 1) consistent direction of 2hGlu association between the discovery and follow-up studies in blue, and 2) consistent direction and nominal significance (P< 0.05) in follow-up studies alone in red.

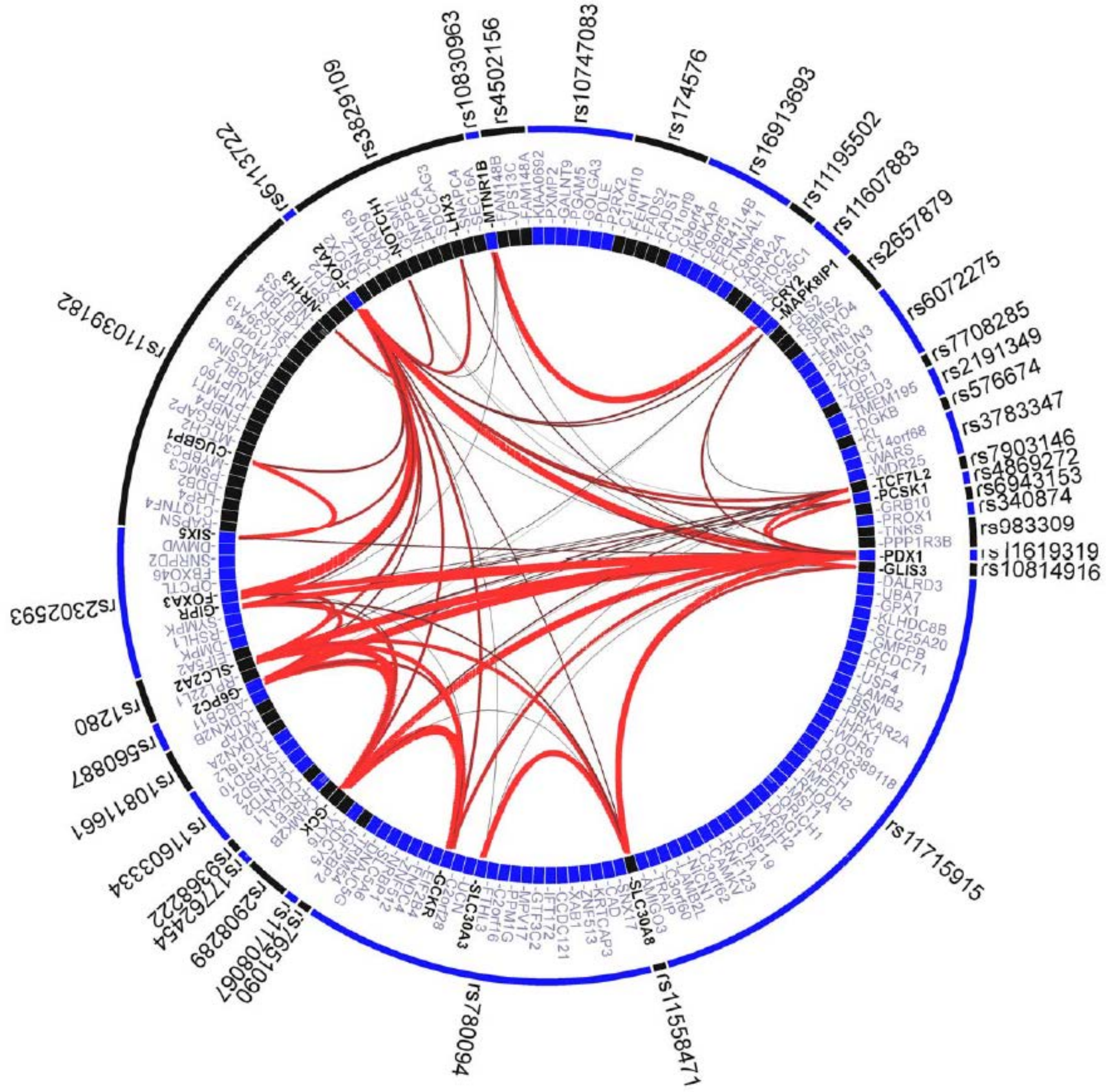


Figure 4.26. GRAIL Connectivity plot for FG. Each genome-wide locus for FG is plotted and significant connections (P<0.05) based on pubmed abstracts (pre-2007) shown as red lines.

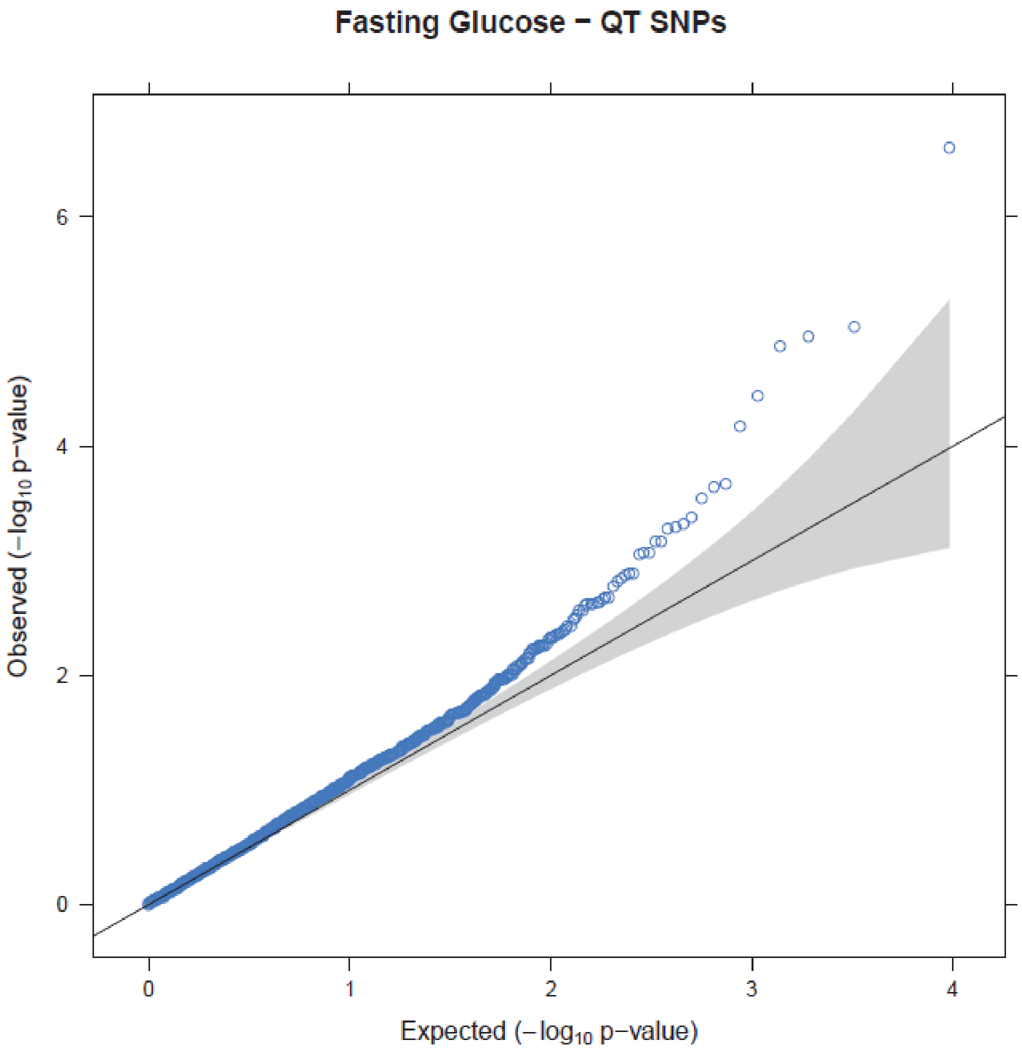


Figure 4.27. QQ plot for fasting glucose association with only QT interval SNPs.

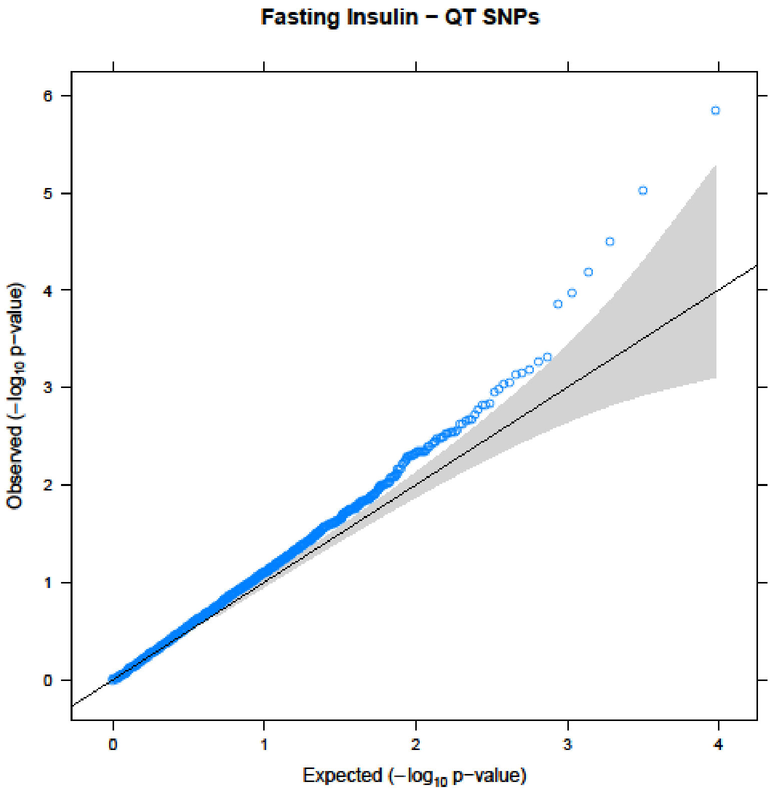


Figure 4.28. QQ plot for fasting insulin association with only QT interval SNPs.

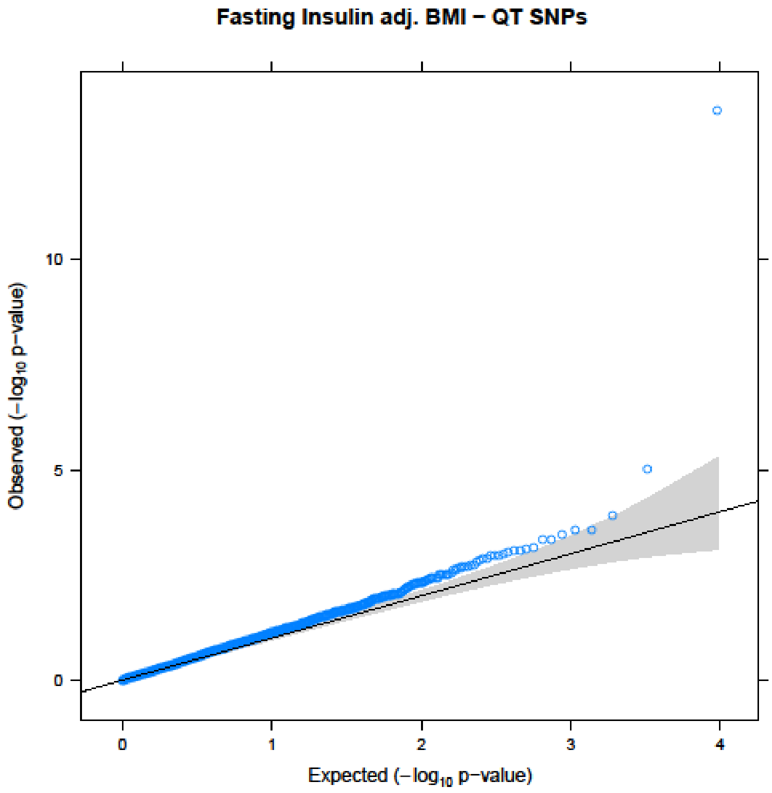


Figure 4.29. QQ plot for fasting insulin (adjusted for BMI) association with only QT interval SNPs.

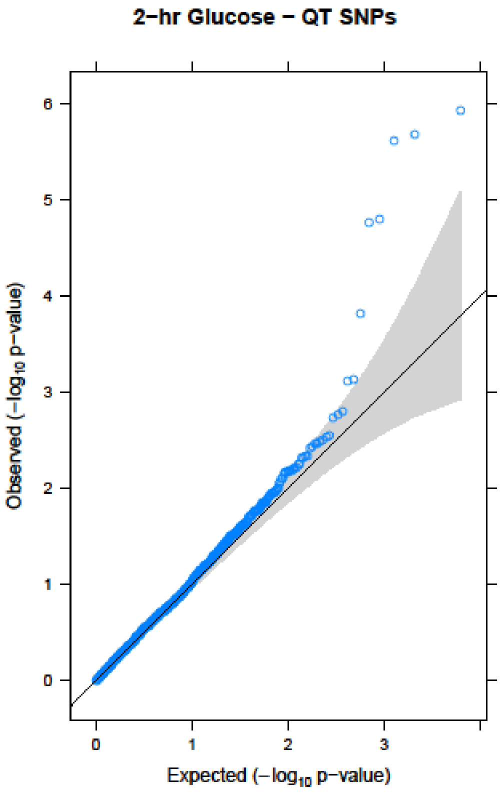


Figure 4.30. QQ plot for 2-h glucose association with only QT interval SNPs.

Chapter 5 ChIP-Enrich: Gene set enrichment testing for ChIP-seq data

5.1 Introduction

Genome-wide high-throughput experiments, whether assessing differential gene expression, differential epigenetic marks, or transcription factor binding, often result in hundreds or thousands of genes of interest. A common approach to bridging the gap from individual genes to biological pathways is gene set enrichment testing, in which the genes identified as biologically interesting from an experiment are tested for statistically significant overlap with pre-defined biologically-related sets of genes (e.g., Gene Ontology (GO)) (Dennis et al. 2003; Curtis et al. 2005; Rivals et al. 2007). Gene set enrichment testing was originally developed for microarray gene expression data, (Tavazoie et al. 1999) with the underlying assumption that, in the absence of a specific enrichment, each gene category would contain the same proportion of associated genes, i.e. that genes in any given GO term or other gene set are no more likely to be identified than other genes. Subsequently, gene set enrichment testing has been applied to a wide variety of whole genome experimental data including chromatin immunoprecipitation followed by deep sequencing (ChIP-seq) experiments. In ChIP-seq, biologically interesting genes are defined as those genes with a “peak” in a region of the genome designated as part of their potential regulatory domain. We term this region from which we predict a gene could be regulated as the *locus* of the gene, and we term the length in base pairs of the locus as the *locus length*.

ChIP-seq peaks may follow a number of potential models of spatial distribution across the human genome. Considering those peaks that represent specific (true) protein binding, a range of possible binding patterns could exist (Barski et al. 2007; Wang et al. 2012). In one possible pattern, a protein specifically binds close to transcription start sites (TSSs); locus length does not influence the probability that a gene has a peak, as each gene has approximately the same length of regulatory space (Figure 5.2, top left). A second possible pattern is that a protein serves as an enhancer and binds distal to the TSS. The probability of a peak being assigned to a gene might increase with locus length, as peaks could be assigned to neighboring genes that they do not regulate, but the exact relationship would depend on factors such as how genes are clustered in the genome and the average distance from the enhancer to the gene. In a third possible pattern, peaks observed in an experiment are false (e.g. experimental error, antibody with poor specificity), and we would expect these peaks might occur randomly throughout the genome, and thus the probability of a peak being assigned to a gene would be proportional to the locus length of the gene (Figure 5.2, bottom right). In addition, technical complexities in calling ChIP-seq peaks can result in unobserved (false negative) peaks. Sequence mappability varies across the human genome, and therefore peaks are more likely to be identified in highly mappable gene loci (Rozowsky et al. 2009; Cheung et al. 2011). Together, the different patterns of observed peaks and false negative peaks likely lead to a complex relationship between observable protein binding, locus length, mappability, and the enrichment of biologically-related gene sets.

The most common statistical test employed for gene set enrichment is Fisher's exact test, and many tools are based on this or a closely related test [ConceptGen (Sartor et al. 2010), DAVID (Dennis et al. 2003), OntoTools (Draghici et al. 2007; Khatri et al. 2007)]. Fisher's exact test is used to find significant associations between two categorical variables: in this case, whether a gene is identified in an experiment, and whether a gene belongs to a pre-defined gene set. One assumption of Fisher's exact test is that, in the absence of enrichment, genes in each gene set (e.g. GO term) are equally likely to have ≥1 peak. However, the locus length of genes vary by several orders of magnitude (and thus genes with longer locus length are more likely to be assigned a peak by chance). Furthermore, genes with longer (or shorter) locus length tend to belong to certain biological functions and processes (Ovcharenko et al. 2005) and thus Fisher's exact test will be biased when as a group, genes with longer locus length are more likely to be assigned a peak than genes with shorter locus lengths (Taher and Ovcharenko 2009).

To address this bias, (Taher and Ovcharenko 2009) proposed a binomial test where successes are defined as the number of binding events (e.g. ChIP-seq peaks) within the predicted regulatory domains (loci) of genes annotated to a GO term. To account for the relationship between probability of a binding event and locus length, they used the total length of the gene locus relative to the total genome length as the expected binomial probability. Their approach assumes that true binding sites are equally likely to occur anywhere in the genome, and they justify this assumption based on the fact that no comprehensive database is yet available containing all possible regulatory regions along the human, or other species', genome. This is equivalent to assuming that the expected number of peaks assigned to a gene is proportional to locus length. Genomic Regions Enrichment of Annotation Tool (GREAT) (McLean et al. 2010) implements this basic binomial test. The binomial test can detect enrichment of a GO term from a large number of peaks in a single or a few genes, and to overcome this limitation McLean et al. (2010) suggest comparing the results of the binomial test with a hypergeometric test (one-sided Fisher's exact) to detect gene set enrichment.

The assumptions of Fisher's exact test and the binomial approach represent two extremes that likely do not reflect the true nature of the relationship between locus length and the probability that a peak is assigned to a gene. Our goal was to develop a method (ChIP-Enrich) that empirically models this potentially complex relationship, rather than assuming either that peaks occur randomly across the genome or that each gene has an equal likelihood of being assigned at least one peak. We demonstrate both through simulation and permutations of experimental data that in the absence of enrichment, ChIP-Enrich exhibits the correct type 1 error rate. In contrast, Fisher's exact test and the binomial approach both show an inflated type 1 error rate under certain scenarios of no true enrichment. We compare the behavior of these tests using publicly available ChIP-seq experimental datasets. ChIP-Enrich is applicable to other types of genome-wide experiments that generate lists of genomic regions (such as MeDIP-seq) as their output, and we describe an R package *chipenrich* for researchers to apply our method to such experiments.

5.2 Methods

5.2.1 Definition of a gene locus

We define a gene as the region between the furthest upstream transcription start site (TSS) and furthest downstream transcription end site (TES) for that gene. The positions of the TSSs and TESs for each gene are extracted from the UCSC refFlat table (human genome build hg19.) We removed small nuclear RNAs as they follow different regulatory mechanisms than other genes and often reside within another gene. We define a gene locus in two manners for analyses. For primary analyses in which we test for gene set enrichment, we define a gene locus as the region spanning the midpoints between the TSSs of adjacent genes (Figure 5.1.) We calculate the midpoint for each peak, and assign the peak to the gene locus that its midpoint overlaps. This procedure is equivalent to assigning each peak to the gene of the nearest TSS. We chose this model for assigning peaks to genes because our ChIP-seq peak datasets are for transcription factors, and therefore may be more likely to bind and exert an effect on nearby TSSs. We compare this model (assigning peaks to the nearest TSS) to an commonly used method (assigning peaks to the nearest gene) in Supplementary Results 5.6.5.

To test whether GO terms are enriched when we consider only peaks near TSSs, we created an additional definition of a gene locus as the region within 1 kb of any of its TSSs. If TSSs from two adjacent genes are within 1kb of each other, we use the midpoint between the two TSSs as the boundary of the locus for each gene.

5.2.2 Quality control for Gene Ontology categories

We use a set of filters to create a quality set of Gene Ontology (GO) terms to use for enrichment testing. We extracted GO molecular functions, GO cellular components, and GO biological processes from Bioconductor species specific annotation packages and the GO.db R package. We first remove genes from each GO term that do not exist in our gene locus definitions, since these genes would never have a peak assigned to them and would therefore inflate the count of genes in a GO term without a peak. Each GO term is checked for duplicated Entrez gene IDs, and we retain one ID from each set of duplicated IDs. We limit our categories to those with greater than 30 genes, to avoid the abundance of small GO terms which often lack the power to detect enrichment even when present.

5.2.3 Mappability calculations

We define base pair mappability M_i as the average read mappability of all possible reads of size *K* that encompass a specific base pair location, *b*. Mappability files from the UCSC Genome Browser mappability track were used to calculate base pair mappability. The mappability track provides values for theoretical read mappability, or the number of places in the genome that could be mapped by a read that begins with the base pair location *b*. For example, a value of 1 indicates a *K*mer read beginning at *b* is mappable to one area in the genome. A value of 0.5 indicates a *K*mer read beginning at *b* is mappable to two areas in the genome. For our purposes, we are only interested in uniquely mappable reads; therefore, all reads with mappability less than 1 were set to 0 to indicate non-unique mappability. Then, base pair mappability is calculated as:

$$M_i = \left(\frac{1}{2K-1}\right) \sum_{j=i-K+1}^{i+(K-1)} M_{readj}$$

where M_i is the mappability of base pair *i*, and M_{readj} is mappability (from UCSC's mappability track) of read *j* where *j* is the start position of the *K* length read. We calculated base pair mappability for reads of lengths 24, 36, 40, 50, 75, and 100 base pairs for *Homo sapiens* (build hg19) and for reads of lengths 36, 40, 50, 75, and 100 base pairs for *Mus musculus* (build mm9). We define gene locus mappability as the average of all base pair mappability values for a gene locus.

5.2.4 ChIP-Enrich method for gene set enrichment testing of ChIP-seq data

We developed a logistic regression approach to simultaneously 1) adjust for the gene locus length and mappability, and 2) test for gene set enrichment. Our model is shown in Equation 5.1:

$$\log\frac{\pi_i}{1-\pi_i} = \beta_0 + \beta_1 \text{geneset} + f(\log_{10}(\text{locus length} * \text{mappability} + 1)) \qquad \textbf{Equation 5.1}$$

π_i is defined as the probability of gene *i* being assigned a peak. The π_i values are not observed directly; instead we observe only whether each gene was assigned one or more peaks. Only those genes annotated in GO and present within our locus definitions are considered. Our dependent variable is then a binary vector with 1 if the gene has one or more peaks assigned to it, and 0 otherwise. The parametric term *geneset* is also a binary vector, where 0 denotes that the gene does not belong in the set of genes being tested and 1 otherwise. We chose this coding as it results in a test of the enrichment of genes having one or more peaks within a set of genes. Alternate codings are possible, such as whether a gene had ≥2,3,4, or more peaks, although we do not explore this in our current work. The function *f(log$_{10}$ (locus length * mappability+1))* is a binomial smoothing spline term that takes into account both the locus length and the average mappability of each gene locus. The model is fit using a penalized likelihood maximization approach, where the smoothing penalty is the conventional squared second derivative penalty, and where the smoothing parameters are estimated using generalized cross-validation (Wood 2006; Wood 2011). We use the gam procedure of the R package mgcv to fit the model (Wood 2010). For each gene set, the model is fit and a p-value is computed by performing a Wald test on the *geneset* coefficient β_1. P-values are corrected for multiple testing using the false discovery rate approach (Benjamini and Hochberg 1995). We compared alternative choices of models in Supplementary Results 5.6.1.

5.2.5 Fisher's exact test for gene set enrichment testing of ChIP-seq data

We use Fisher's two-sided exact test to test for enrichment of genes with at least 1 assigned peak within GO terms. We used the R procedure *fisher.test* to perform the test. We considered only genes that are both 1) annotated in GO, and 2) present in our locus definitions (i.e. a peak could be assigned to a gene) for analysis.

5.2.6 Implementation of a binomial test for enrichment of ChIP-seq peaks within gene sets

We implemented a binomial test based on the test introduced by (Taher and Ovcharenko 2009) and previously implemented in GREAT (McLean et al. 2010). We use a one-sided binomial test for enrichment of peaks within the loci of genes annotated to a GO term defined by 3 parameters: 1) the count of peaks assigned to genes within a GO term, 2) the total number of peaks assigned to any GO term, and 3) the probability of observing a peak in a gene within a GO term, defined as the total length of the genes in the GO term divided by the sum of the lengths of the gene locus regions. The probability parameter represents the fraction of the genome covered by the gene loci within a GO term. We defined gene loci using the nearest TSS method (see Methods 5.2.1).

Our test differs in two ways from the (McLean et al. 2010) version. First, we remove genes from consideration that are not present within any GO term. This removes peaks assigned to genes that are un-annotated in GO, and therefore would artificially inflate the total number of peaks. This is relevant since genes annotated in GO are enriched for peaks relative to genes not annotated in any GO term for all eight transcription factors we studied (Supplementary Methods 5.5.6, Table 5.8.) Second, we use the sum of the gene locus regions as our total genome length, rather than the length of the non-gapped genome. This removes the gene loci not present in any GO term from the calculation of the total genome length, which would artificially decrease the binomial probability and inflate the significance of the test. It also allows for a consistent genome length when considering gene locus definitions that do not encompass the entire genome, or for genome lengths representing the mappable or sequence-able genome, rather than the entire non-gapped genome length. We investigate the effect of overestimating genome length on the binomial test in Supplementary Results 5.6.2.

5.2.7 Experimental ChIP-seq peak datasets

We selected 8 publicly available ChIP-seq peak datasets from the literature, spanning a range of read lengths, total number of peaks, and binding distribution relative to transcription start sites (Table 5.4, Figure 5.21). (Reddy et al. 2009; Heikkinen et al. 2011; Lee et al. 2011; Mokry et al. 2012). The dataset authors called peaks using a variety of calling methods (Cisgenome, MACS, and a custom Parzen window approach.) We converted datasets in hg18 coordinates to hg19 using the UCSC liftOver tool (Fujita et al. 2011).

5.2.8 Characterization of the bias of Fisher's exact test and the binomial test when no enrichment exists using simulated data

We simulated ChIP-seq peaks under the null hypothesis of no association with any GO term. Instead of simulating peak locations, we randomly sampled genes to represent the count of peaks occurring within the locus of a gene. Genes were sampled by two methods: 1) randomly sampling genes with replacement, and 2) randomly sampling genes in proportion to their locus length (as defined in Methods 5.2.1), also with replacement. The first method simulates peaks occurring within genes with no dependence on their locus length (Fisher's exact test assumption). The second method simulates peaks being assigned to genes with probability in proportion to locus length (binomial test assumption). We sample a total of 10,000 genes, and vary the percentage (0,10,25,50,75,100%) of the total 10,000 genes that are sampled by locus length. We tested a subset of our simulations for gene set enrichment (0, 25, 50 and 75% peaks sampled by locus length). For the binomial test we use the number of times a gene is sampled as the count of peaks in the gene. For Fisher's exact test and ChIP-Enrich, a gene is labeled as having a peak if the count of peaks is ≥ 1. Each GO term is then tested for enrichment using either Fisher's exact test, the binomial test, or ChIP-Enrich. We repeat this process 1000 times for each test and percentage of peaks sampled by locus length, and calculate the median of the 1000 simulation p-values at each quantile of the 2565 GO term p-values.

5.2.9 Permutations of experimental data and randomly generated GO terms for the comparison of ChIP-Enrich, Fisher's exact test, and the binomial test

We performed simulations to assess the behavior of each enrichment test under two null scenarios of no true enrichment. For both scenarios, we used experimental ChIP-seq datasets for E2F4 (a member of the E2F transcription factor family) and NR3C1 (GR; glucocorticoid receptor). In the first scenario, we permute the count of peaks per gene among genes with similar locus length to simulate no true association between the number of peaks assigned to a gene, and whether the gene belongs to the GO term being tested. We sorted genes by their locus length and then divided them into bins of 100. We then randomly permute the count of peaks for the genes within each bin. This permutation maintains the relationship between the locus length of a gene and its likelihood of receiving a peak, while removing any association between membership of a gene in a GO term and its count of peaks. After permuting the peaks within each bin of genes, we test GO terms for enrichment with each of the three tests. For the second scenario, we created random gene sets of the same number and size distribution of GO, i.e., for each GO term we randomly sampled the same number of genes from all genes annotated in GO. We then tested the randomly generated GO terms for enrichment with each of three tests (Fisher's exact test, ChIP-Enrich, and the binomial test), using the peaks from the E2F4 and GR experimental datasets.

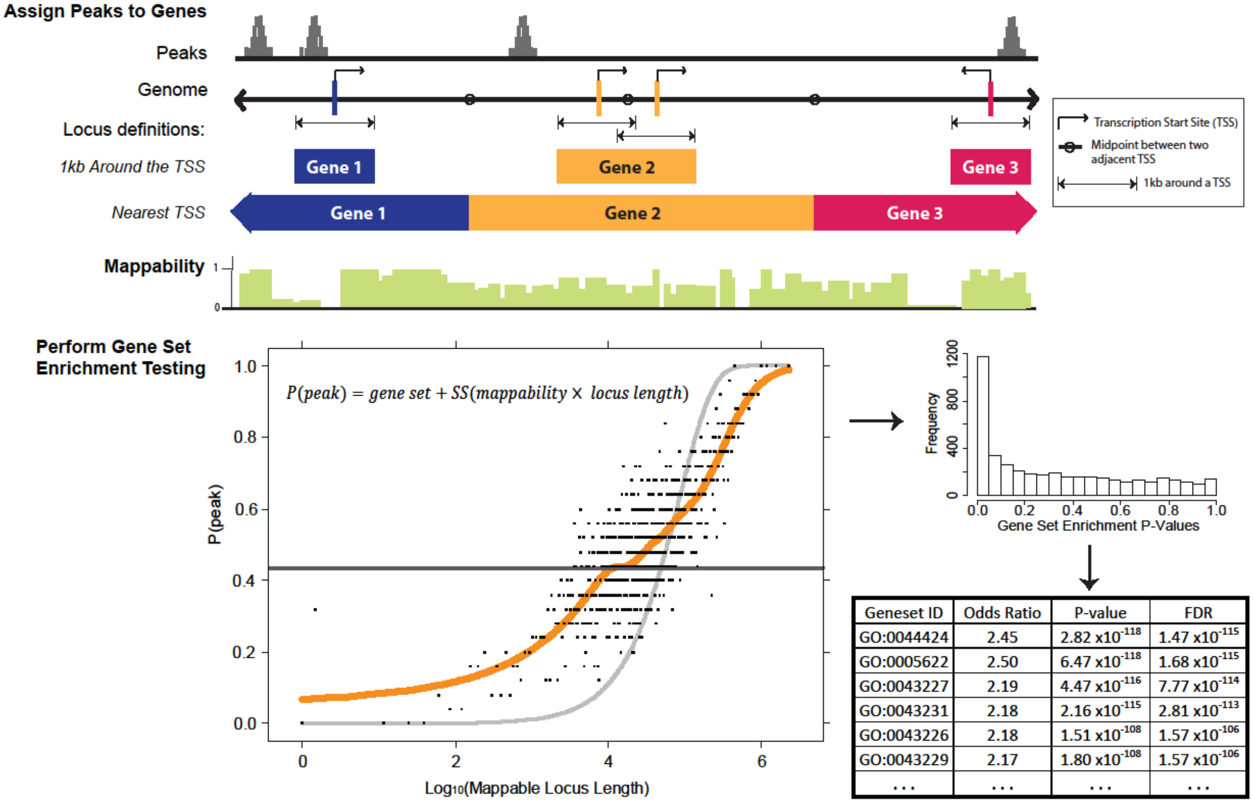
5.2.10 R package

We implemented our method in the *chipenrich* package for the R statistical software environment and will be made available through Bioconductor (Gentleman et al. 2004). We provide our ChIP-Enrich test for gene set enrichment along with the two existing approaches: Fisher's exact test, and the binomial test for count of peaks. In addition to Gene Ontology, we provide 15 additional annotation sources containing over 20,000 total sets of genes. These sets of genes were previously collected into a database for the LRpath gene set enrichment testing tool (http://lrpath.ncibi.org) (Kim et al. 2012) and are now available as part of the *chipenrich* R package. We currently support both the human genome build hg19 and the mouse genome build mm9. For ease of use, users may either supply an R data frame or the path to a BED format file containing the peak locations as input. Runtime is typically 30 minutes to 2 hours (for testing all GO terms) depending on the dataset and hardware, and can be as low as 5-10 minutes for smaller sets of genes such as KEGG or Biocarta pathways. We offer a number of alternative locus definitions in ChIP-Enrich in addition to the nearest TSS locus definition: (1) gene midpoints - a locus is bound by the midpoints between the adjacent gene loci, using the smallest and greatest base pair position of the gene transcript's TSS and TES to define loci boundaries; (2) exons – peaks are assigned to gene exons; (3) 1kb – peaks within 1kb up or downstream of a TSS are assigned to the gene; (4) 5kb – peaks within 5kb up or downstream of a TSS are assigned to the gene. Note that only peaks falling within the defined regions are used in locus definitions 2-4. Users may also generate diagnostic plots for their peak datasets (see the *chipenrich* vignette.)

5.3 Results

5.3.1 Overview of the ChIP-Enrich approach

Figure 5.1 illustrates the steps in ChIP-Enrich. We created gene locus definitions using boundaries (*see methods*) defined as 1) the midpoint between the closest TSSs of two adjacent genes, or 2) the region within 1 KB of the TSS. We assign peaks to genes based on the chosen locus definition, and use a logistic regression model to test for GO term enrichment. A binomial cubic smoothing spline is used to model the probability of a peak as a function of gene locus length and mappability), and to adjust for this relationship in our analysis of GO term enrichment. ChIP-Enrich calculates p-values and false discovery rates for enrichment (or depletion) of pre-defined biological gene sets from 15 different annotation databases previously described in the LRpath gene set enrichment testing software (Kim et al. 2012).


Assign Peaks to Genes
Peaks
Genome
Locus definitions:
1kb Around the TSS
Nearest TSS
Gene 1
Gene 2
Gene 3
Transcription Start Site (TSS)
Midpoint between two adjacent TSS
1kb around a TSS
Mappability
1
0
Perform Gene Set Enrichment Testing
P(peak) = gene set + SS(mappability × locus length)
P(peak)
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
Log10(Mappable Locus Length)
Frequency
0
400
800
1200
0.0
0.2
0.4
0.6
0.8
1.0
Gene Set Enrichment P-Values

Geneset ID	Odds Ratio	P-value	FDR
GO:0044424	2.45	2.82×10^{-118}	1.47×10^{-115}
GO:0005622	2.50	6.47×10^{-118}	1.68×10^{-115}
GO:0043227	2.19	4.47×10^{-116}	7.77×10^{-114}
GO:0043231	2.18	2.16×10^{-115}	2.81×10^{-113}
GO:0043226	2.18	1.51×10^{-108}	1.57×10^{-106}
GO:0043229	2.17	1.80×10^{-108}	1.57×10^{-106}
…	…	…	…

Figure 5.1. Overview of the ChIP-Enrich method. ChIP-seq peaks are first assigned to genes using a chosen gene locus definition. Peaks can be assigned to the gene with the nearest TSS, the 1kb region on either side of a TSS, or a number of other options (top panel.) If the user wishes to adjust for sequence mappability, we multiply the length of each gene locus by the average mappability across that locus. We then test each gene for enrichment using a logistic regression model, and adjust for the locus length of each gene using a binomial cubic smoothing spline term. The bottom left panel represents a visualization of the spline fit (orange) to experimental data after assigning peaks to genes. For visualization only, each point is a bin of 25 genes, plotted as the average proportion of genes assigned a peak within the bin against the average $\log_{10}$ locus length. The dark gray line represents the fit if Fisher's exact test assumption held (in the absence of enrichment, all genes are equally likely to be assigned ≥1 peak). The light gray logistic-shaped curve represents the fit if the assumption of the binomial test held (expected number of peaks assigned to a gene is proportional to locus length.) P-values and q-values for each gene set tested are calculated and reported to the user.

5.3.2 Potential confounding from locus length and mappability

Taher and Ovcharenko (2009) showed that locus length is related to GO term membership for loci defined as the gene and half the intergenic region between adjacent genes (Taher and Ovcharenko 2009). Using our locus definition of the midpoint between TSS (locus definition for all presented analysis (Methods 5.2.1), we similarly found that genes in GO terms related to nucleosome, protein-DNA complexes, and translation have short locus lengths (Table 5.5). In contrast, genes in GO terms representing nervous system development, cell adhesion, and transcription tend to have long locus lengths (Table 5.6).

The probability that a peak will be observed in a gene locus is also affected by the mappability of the DNA within the locus. If mappability varies randomly throughout the genome, then it is unlikely to affect enrichment testing results. However, if the mappability of loci varies by GO term membership, then mappability has the potential to bias enrichment testing results. To assess the association between mappability and GO term membership, we calculated mappability at each base pair across the hg19 version human genome for different sequence read lengths (see Methods 5.2.3.) Next, we calculated a mappability score for each gene defined as the average mappability across the gene's locus length. As expected, the gene mappability scores show wide variability at lower read lengths and mappability increases as read lengths increase from 24 to 100 bp (Figure 5.9. A). We found that genes with low mappability are significantly more likely to be present in sensory and xenobiotic response and oxygen related terms, whereas genes with high mappability are significantly more likely to be involved in the nervous system or development terms (at the 50-mer read length, q-value $< 3.0 \times 10^{-16}$) (Figure 5.9. B,C.). Several GO terms (e.g., central nervous system development) had longer locus lengths and higher mappability, increasing the possibility of confounding in gene enrichment tests.

5.3.3 Comparison of ChIP-Enrich, Fisher's exact test, and the binomial test under the null hypothesis of no enrichment using simulated data

In the absence of GO term enrichment, Fisher's exact test assumes that the probability of a peak is the same for every gene, which is satisfied when the probability of a gene having a peak is independent of its locus length. In contrast, the binomial test assumes that the probability of a peak occurring in a locus is proportional to the locus length. To examine the sensitivity of Fisher's exact test, the binomial test, and ChIP-Enrich to these assumptions, we simulated datasets of peaks containing a mixture of peaks sampled independently of locus length (Fisher's exact test assumption) and peaks sampled in proportion to locus length (binomial test assumption.) Rather than simulate peak locations, we simulated the number of peaks per gene by sampling a total of

10,000 genes as a mixture of: 1) genes sampled at random with replacement (representing peaks assigned with equal probability to each gene), and 2) genes sampled randomly in proportion to their locus length with replacement (probability of peak assignment to a gene is proportional to locus length.) As the percentage of peaks sampled by locus length increases, the relationship between the probability of a gene having a peak and locus length changes from relatively flat (Figure 5.2, top left) to increasingly logistic shaped (Figure 5.2, bottom right).

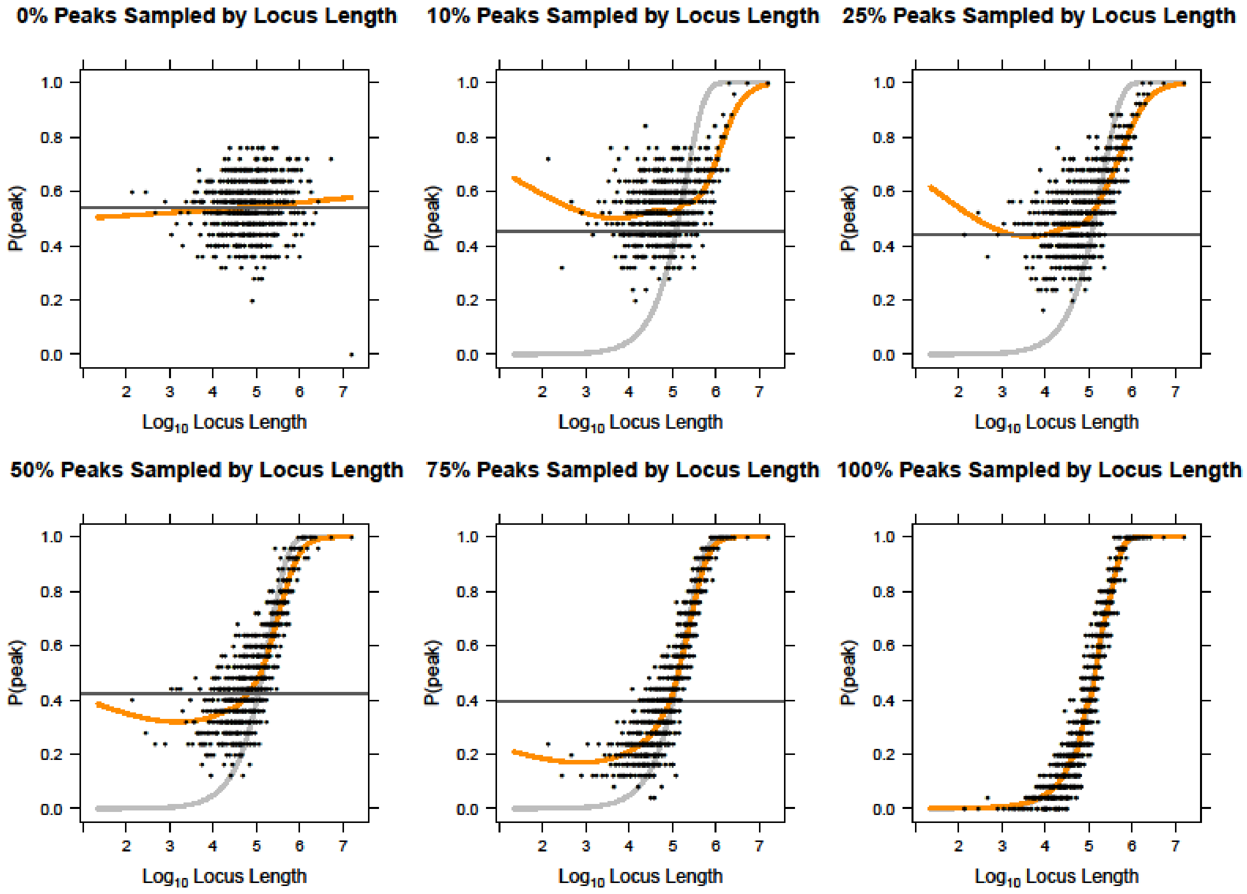


Figure 5.2. Probability of a gene having a peak given its locus length, as the percentage of peaks sampled by locus length increases. For visualization, each point is a bin of 25 genes, plotted as the average proportion of genes having a peak within the bin against the average $\log_{10}$ locus length. The dark grey horizontal line represents the model where peaks occur within genes with no relationship to their locus length. The light grey line represents the theoretical probability of a locus having ≥1 peak given its length and the total length of the genome. The orange line is a binomial smoothing spline fit to the underlying data (the 0/1 vector denoting whether a peak was assigned to a gene vs. the $\log_{10}$ locus length of each gene.)

We tested each of our simulated datasets (genes sampled with replacement to represent peaks) for enriched GO terms using Fisher's exact test, the binomial test, and

ChIP-Enrich (Methods 5.2.8) and plotted the sorted observed $–\log_{10}$(p-values) versus the expected $–\log_{10}$ p-values (Figure 5.3.) In the presence of no enrichment, every plotted line should follow the y=x line. To increase the stability of our results, we performed each simulation and enrichment test 1000 times and plotted the median p-value for each quantile, i.e. the median of the most significant p-values, the median of the next most significant p-values, and so on.

When the peaks have equal probability of occurring in each gene (random 0%) Fisher's exact test shows a slight deflation of the most significant p-values as expected due to the discrete nature of the data (Upton 1992) (see also Supplementary Results 5.6.3.) With increasing proportions of peaks in proportion to locus length, Fisher's exact test becomes increasingly anti-conservative: at 100% peaks sampled by locus length, the median p-values observed for the 0.5, 0.05 and 0.0001 quantiles (equivalent to the expected p-values at these values) were 0.21, 5.39×10^{-5}, and 5.44×10^{-17}. In contrast, the binomial test becomes increasingly anti-conservative for increasing proportions of peaks with equal probability in each gene, with the highest level of inflation observed when the probability of having a peak is not related to gene length (0% peaks sampled by locus length): the median p-values observed for the 0.5, 0.05 and 0.0001 quantiles were 0.68, 1.1×10^{-5}, and 1.55×10^{-24}. ChIP-Enrich shows no inflation from the expected distribution in any scenario but instead shows a similar slight deflation of p-values as observed for Fisher's exact test with 0% peaks sampled by locus length.

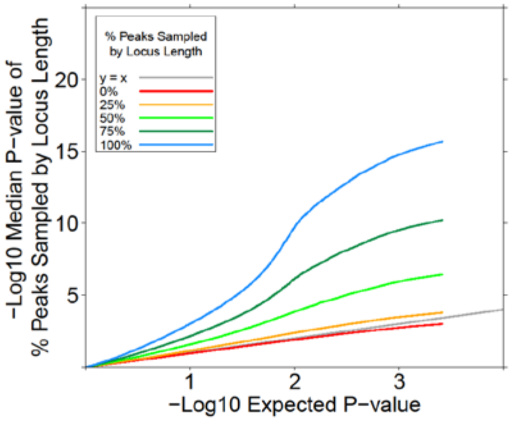
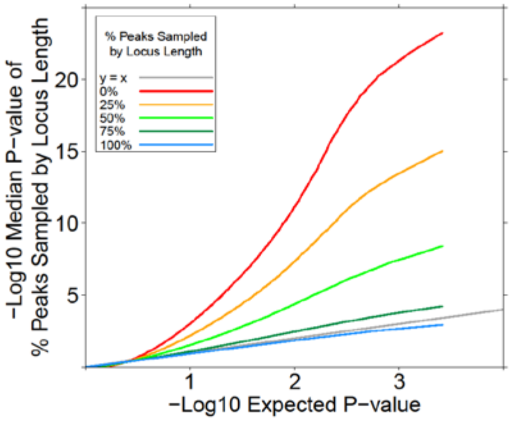
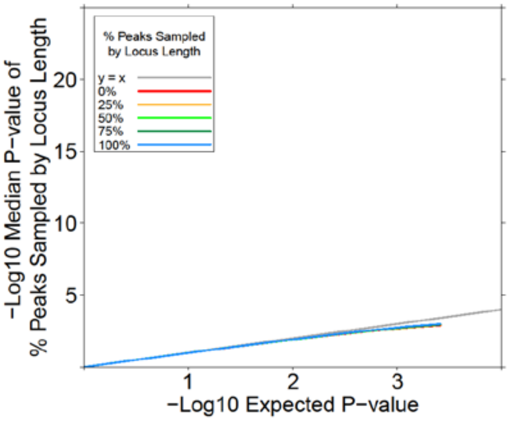
Test	Assumptions	QQ Plot
Fisher's Exact Test	Assumes each gene has an equal probability of having a peak	
Binomial	Assumes probability of a gene having a peak is proportional to locus length	
ChIP-Enrich	Empirically estimates probability of having a peak based on locus length	

Figure 5.3. Fisher's Exact and the binomial test represent extreme assumptions for enrichment testing for ChIP-Seq data, while ChIP-Enrich empirically estimates the correct balance between these two extremes. Incorrect assumptions at either end leads to biased significance levels. Median p-values are shown for 1000 simulations of Fisher's exact test, ChIP-Enrich, and the binomial test with increasing percentages (0-100%, red to blue) of peaks sampled by locus length from a total of 10,000 peaks.

5.3.4 Observed relationship between probability of a peak and locus length in experimental data

To explore the behavior of ChIP-Enrich over experiments covering a broad range of binding distributions, we analyzed eight publicly available ChIP-seq datasets from the literature to analyze (Reddy et al. 2009; Heikkinen et al. 2011; Lee et al. 2011; Mokry et al. 2012) (Figure 5.21). Using four of these transcription factors that have a mid-range number of total peaks (GR, beta catenin, E2F4, TBP), we see a trend in the relationship between probability of a peak and locus length. This trend ranges from GR, which tends to bind distally to genes (77.9% of peaks occur > 50 kb from the TSS), to TBP, which tends to bind proximal to genes (63.3% of peak occur within 1kb of the TSS) (Figure 5.4; top left and bottom right). Beta catenin and E2F4 represent binding distributions intermediate between these two more extreme datasets (Figure 5.4 top right and bottom left.) These binding distributions, along with the total number of peaks and other possible factors, lead to different relationships between the presence of a peak within a gene locus, and its observable locus length, and all four datasets exhibit a departure from both the assumption that all genes are equally likely to have a peak, and that the probability that a peak being assigned to a gene is directly proportional to its locus length. The relationship for GR closely resembles that of random peaks occurring throughout the genome for a wide range of locus lengths ($\log_{10}$ locus length between 4 and 5, or $\approx$ 56% of all genes.) In contrast for TBP, we see the relationship is flatter, more closely resembling the assumption of Fisher's exact test (Figure 5.4, horizontal line). The spline fit tends to flatten for datasets with small numbers of peaks and rises for datasets with larger numbers of peaks (Figure 5.21). From these transcription factors, we selected two datasets (E2F4, GR) that exhibit differences in binding relative to TSSs to illustrate our method and to examine how well ChIP-Enrich identifies the known functions of well-studied transcription factors.

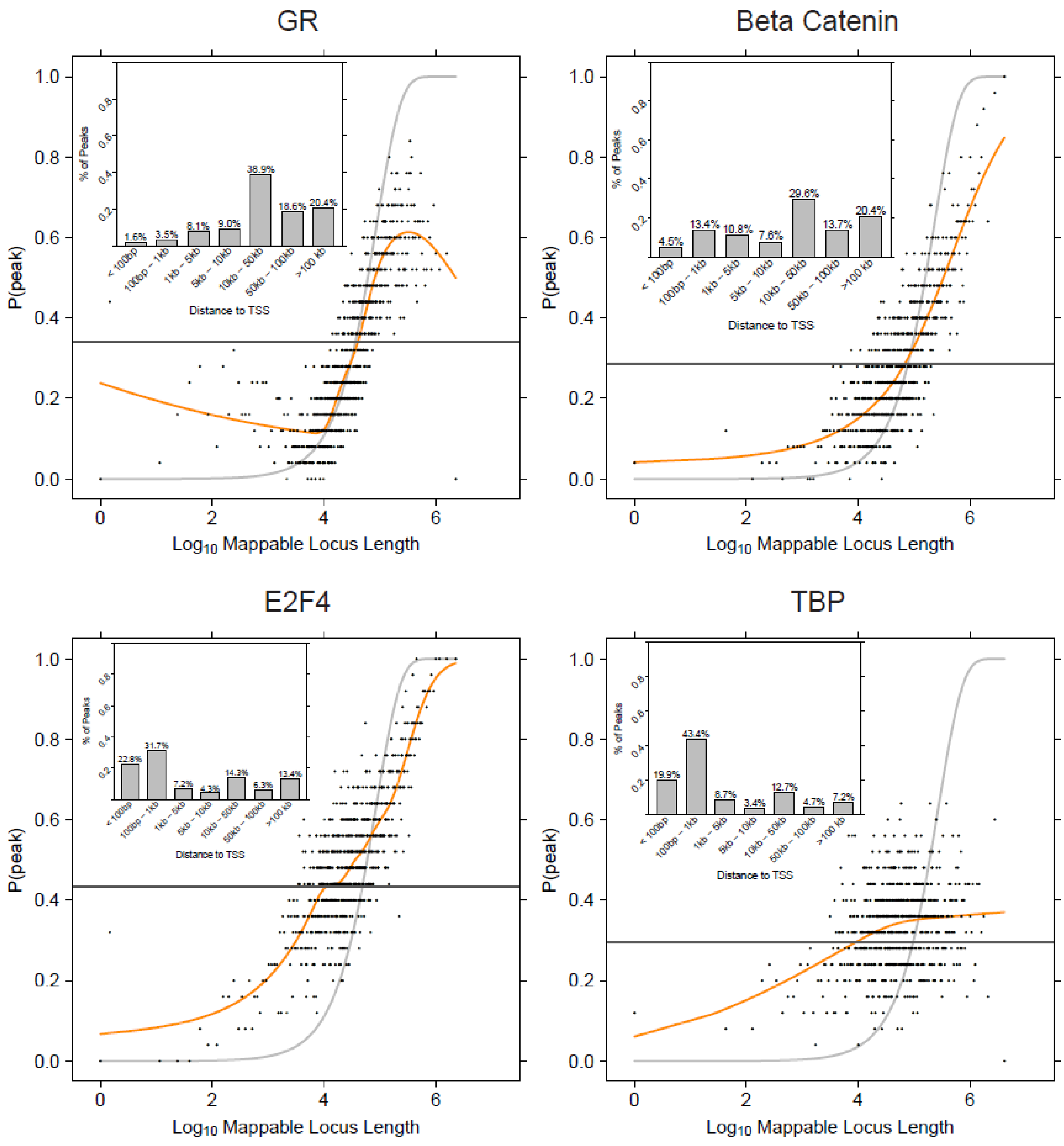


Figure 5.4. Spline fit trend for four experimental datasets ordered by increasing proportion of peaks within 1KB of the TSS (GR, Beta Catenin, E2F4, TBP.) As the proportion of peaks near the TSS increases, the spline fit begins to flatten, representing the scenario where genes are being assigned a peak with no relation to their locus length. Each of the datasets here demonstrate that the probability of a gene being assigned a peak is not a direct function of the locus length, nor is it equal over all genes. The E2F4 dataset is from (Lee et al. 2011), the GR dataset is from (Reddy et al. 2009), and both Beta Catenin and TBP are from (Mokry et al. 2012).

5.3.5 Comparison of ChIP-Enrich, Fishers exact test and the binomial test for experimental and permuted E2F4 and GR datasets

Given our initial simulations (Results 5.3.3) and the observed patterns of the probability of a peak and locus length in the experimental datasets (Results 5.3.4), we hypothesized that, in the absence of gene set enrichment, Fisher's exact test and the binomial test might also have anti-conservative results in these experimental datasets due to violations of the underlying assumptions of these tests. Using permutations of experimental datasets (E2F4, GR) and by simulating GO terms with random sets of genes drawn from all genes annotated in GO, we created two scenarios under which no true enrichment exists, and therefore no evidence for enrichment should be detected.

In the first scenario, we binned genes according to their locus length, and then randomly shuffled the count of peaks assigned to genes within each bin. This procedure preserves the overall relationship in the dataset between peak count and locus length, as well as between locus length and GO term membership. In this scenario, there is still potential for confounding of GO term membership and peak count due to locus length, and also potential for violation of the assumption that the peak count is proportional to locus length.

For the second scenario, we created gene sets of the same number of genes as each GO term by randomly sampling genes annotated in GO without replacement and tested the original E2F4 and GR datasets for GO enrichment (where count of peaks per gene, and the locus length of each gene, from the original dataset are preserved.) This procedure removes the relationship between the GO term and locus length, but preserves the relationship between a given gene and its peak count and locus length. In this scenario there is no potential for confounding due to locus length, but there is potential for violation of the assumption that the peak count is proportional to locus length.

Figure 5.5 shows the expected $–\log_{10}$ p-values versus the $–\log_{10}$ p-values for the two scenarios above. For GR and E2F4, overall ChIP-Enrich shows the expected distribution of p-values, while being slightly conservative for the most extreme p-values in scenario 1 (Figure 5.5, center column). In contrast, for GR and to a lesser extent for E2F2, Fisher's exact test shows inflated $-\log_{10}$ p-values when randomly permuting peaks within bins of gene locus length (scenario 1) (Figure 5.5, left column), because of the confounding effect of locus length. For randomly generated gene sets (scenario 2), however, Fisher's exact test shows the expected distribution of p-values (with a slight deflation relative to the x=y line). For E2F4 and GR, the binomial test shows inflation of $-\log_{10}$(p-values) under both scenarios, with less inflation for E2F4 in scenario 2 when the relationship between GO term and locus length is not preserved (Figure 5.6, right column). Using additional simulations (Supplementary Results 5.6.4), we observe that in both the E2F4 and GR datasets, random sets of genes with shorter than average locus length have an excess of peaks (Figure 5.16.). For E2F4 in particular, the binomial test detects these short sets of random genes as highly significant, and conversely all the longest sets of random genes have p-values = 1 (Figure 5.17.).

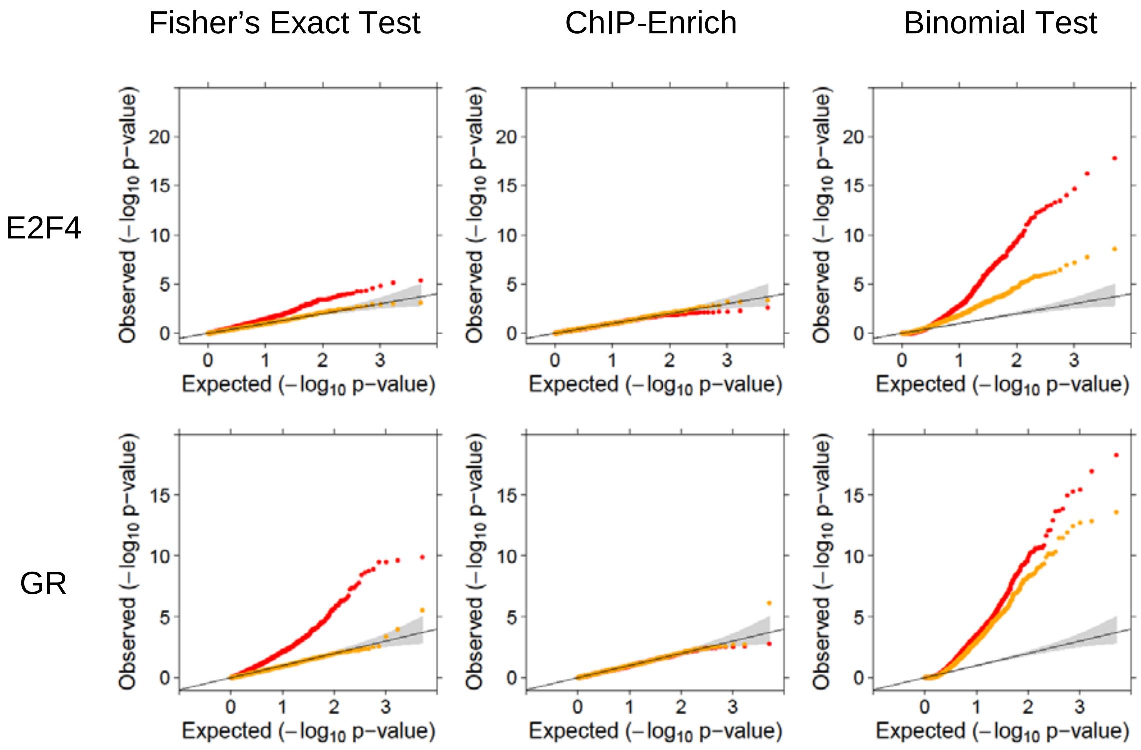


Figure 5.5. Fisher's exact test and the binomial test show inflated p-value distributions while ChIP-Enrich shows the correct p-value distribution under two scenarios of no enrichment: permuting counts of peaks within bins of gene locus length, and randomly generated sets of genes. QQ plots are shown from permuting peak counts within locus length bins (red), and randomly generated sets of genes (orange.) Gray shading represent the 95% confidence interval. Both Fisher's exact test and the binomial test show an inflated p-value distribution when permuting counts of peaks within gene locus length bins. The binomial test also shows an inflated p-value distribution when using randomly generated sets of genes instead of GO terms. ChIP-Enrich shows the expected distribution of

p-values, while possibly being slightly conservative for the most extreme p-values when permuting counts of peaks within locus length bins.

We subsequently compared results from the two scenarios above to the results from the original datasets (Figure 5.6). For each test, the original dataset results were stronger (Figure 5.6, blue) than those observed in the permuted and simulated datasets (Figure 5.6, red and orange) indicating that true biological enrichment is likely being detected by each test. In the original datasets, the binomial test showed substantially stronger signals than ChIP-Enrich or Fisher's exact test in both the GR and E2F4 datasets, and Fisher's exact test showed stronger signals than ChIP-Enrich in the GR datasets (Figure 5.6). However, since the significance of the results of Fisher's exact and the binomial test are inflated, the results cannot be interpreted directly.

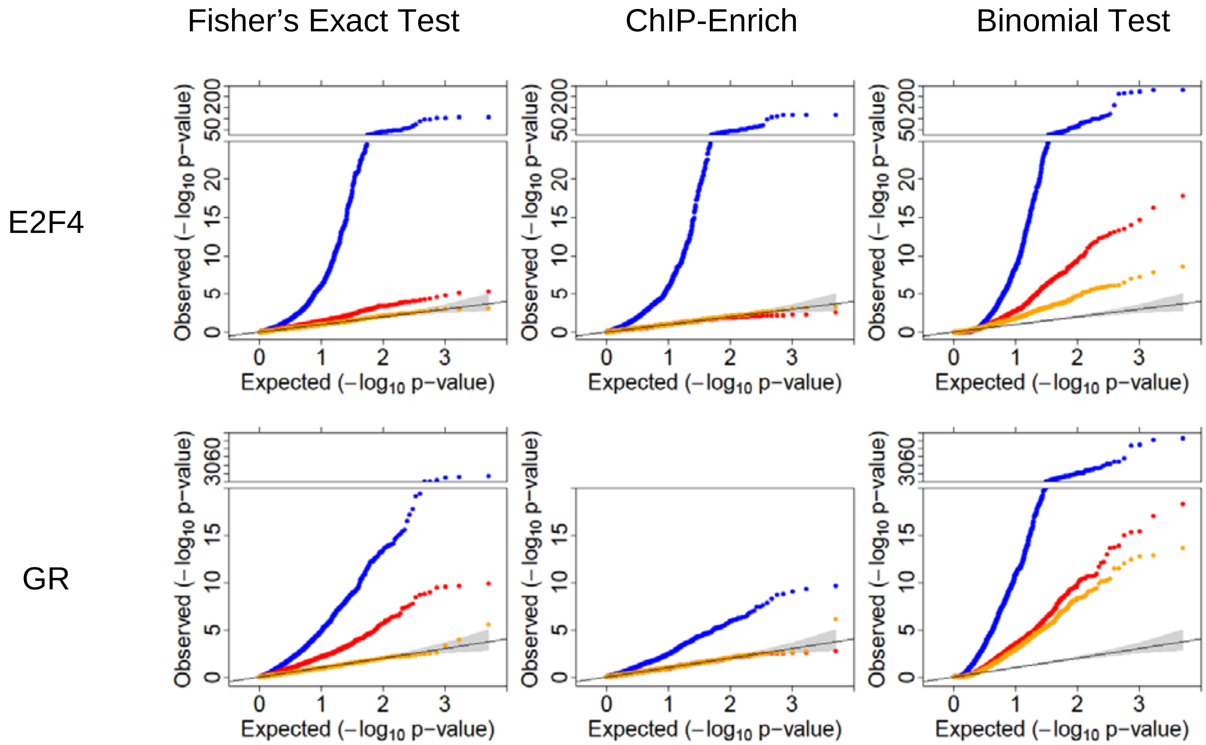


Figure 5.6. Comparison of p-values from two null scenarios with experimental data p-values. Observed $-\log_{10}$ p-values are from testing the actual dataset (E2F4 or GR) with no permutations and using true GO terms (blue), permuting peaks within locus length bins (red), and randomly generated sets of genes (orange.) Figure 5.5 shows a zoomed in figure of the red and orange data points. Gray shading represents the 95% confidence interval. Fisher's exact test and the binomial test both exhibit an enrichment (blue) over either null scenario (orange, red) that likely represents true biological signal, though the observed $-\log_{10}$ p-values are inflated. ChIP-Enrich shows the expected p-value distribution under both null scenarios, albeit with a slight deflation for highly significant p-values.

5.3.6 Analysis of two ChIP-seq datasets (transcription factors E2F4 and GR) with ChIP-Enrich

We investigated whether ChIP-Enrich was able to identify the known functional roles of E2F4 and GR. For E2F4, we found enrichment of multiple cell cycle terms, including "M phase of mitotic cell cycle," "cell division," and "cell cycle arrest" (Table 5.1.) E2F4 and other E2F family of transcription factors are primarily known to be involved in cell cycle progression. E2F4 is known to act as both an activator and repressor of transcription, and binds to the promoter regions of genes known to function in a diverse range of functions, including DNA damage repair, cell cycle progression, and apoptosis (Gaubatz et al. 2000; Plesca et al. 2007).

Rank	GO ID	GO Description	# Genes with Peak	P-value	FDR
1	GO:0044427	chromosomal part	324	1.07E-22	9.31E-22
2	GO:0000279	M phase	335	3.32E-21	2.02E-19
3	GO:0000087	M phase of mitotic cell cycle	256	1.33E-20	7.13E-19
4	GO:0048285	organelle fission	256	1.70E-20	8.84E-19
5	GO:0000280	nuclear division	249	3.25E-20	1.64E-18
6	GO:0006974	response to DNA damage stimulus	342	2.77E-19	1.18E-17
7	GO:0006281	DNA repair	243	4.31E-17	1.61E-15
8	GO:0051325	Interphase	245	1.16E-16	4.17E-15
9	GO:0051329	interphase of mitotic cell cycle	239	1.98E-16	7.00E-15
10	GO:0016071	mRNA metabolic process	334	1.06E-15	3.60E-14
11	GO:0051301	cell division	274	1.84E-15	6.13E-14
12	GO:0005819	Spindle	147	2.02E-15	1.58E-14
13	GO:0045786	negative regulation of cell cycle	271	1.52E-14	4.81E-13
14	GO:0006260	DNA replication	183	3.54E-14	1.10E-12
15	GO:0031012	extracellular matrix	132	4.65E-14	3.45E-13
16	GO:0000228	nuclear chromosome	162	8.72E-14	6.16E-13
17	GO:0007606	sensory perception of chemical stimulus	34	1.58E-13	4.76E-12
18	GO:0006397	mRNA processing	252	2.61E-13	7.73E-12
19	GO:0007050	cell cycle arrest	227	7.45E-13	2.11E-11
20	GO:0000775	chromosome, centromeric region	119	1.20E-12	7.94E-12

Table 5.1. Top 20 GO terms identified from applying ChIP-Enrich to the E2F4 dataset. Only GO terms with less than 500 genes are shown, as this removes large non-specific GO terms.

For the glucocorticoid receptor-alpha (GR), among the top 25 significant GO terms we found a cluster of terms related to glucocorticoid response, such as “response to corticosteroid stimulus,” and “response to glucocorticoid stimulus,” and another cluster relating to the regulation of lipids (Table 5.2). GR is known to be involved in the response to glucose, steroids, and the activation of lipolysis in adipose tissue (Xu et al. 2009; Yu et al. 2010). A third cluster of GO terms represents functions relating to angiogenesis, such as “angiogenesis”, “blood vessel morphogenesis,” “vasculature development,” and “blood vessel development” (Table 5.2). Because GR is not annotated to angiogenesis or a closely related GO term, to examine this finding, we used the MEME software (Bailey et al. 2009) to *de novo* identify motifs overrepresented in the 460 peaks assigned to genes in the angiogenesis GO term relative to reference genome (detected at the Bonferroni-adjusted $p < 0.05$ level). The highest ranked motif from the MEME analysis had the consensus sequence 5’-AGAACAnnnTGTnCT-3’ (Figure 5.22), and was identified in 262 of 460 peaks in the angiogenesis GO term (Table 5.9). Results from the motif comparison tool Tomtom indicated that the best matched transcription factor binding site to this motif was the glucocorticoid receptor element (GRE) ($P=4.43x10^{-07}$), to which GR is known to bind. The GRE motif also tends to occur near the middle of these peaks, whereas the second best matched motif occurs at random throughout the peaks as expected for a false positive result (Figure 5.23, Figure 5.24). Eight of the peaks assigned to genes annotated to angiogenesis were assigned to either *VEGF-A* or *VEGF-C,* which are known to be key regulators of angiogenesis (Table 5.10). Five of these peaks contained a GRE and all five were >25kb from the TSS of the assigned gene.

Rank	GO ID	GO Description	# Genes with Peak	P-value	FDR
1	GO:0001525	angiogenesis	132	3.86E-08	1.13E-05
2	GO:0048514	blood vessel morphogenesis	154	1.61E-07	3.35E-05
3	GO:0006112	energy reserve metabolic process	88	2.86E-07	5.35E-05
4	GO:0070161	anchoring junction	100	5.26E-07	3.85E-05
5	GO:0001944	vasculature development	183	5.68E-07	8.84E-05
6	GO:0042598	vesicular fraction	109	6.48E-07	3.85E-05
7	GO:0001568	blood vessel development	173	6.55E-07	9.42E-05
8	GO:0005792	microsome	105	8.27E-07	4.09E-05
9	GO:0019904	protein domain specific binding	202	1.04E-06	1.37E-04
10	GO:0005912	adherens junction	91	1.12E-06	4.73E-05
11	GO:0032787	monocarboxylic acid metabolic process	153	2.32E-06	2.41E-04
12	GO:0042493	response to drug	138	3.77E-06	3.29E-04
13	GO:0051056	regulation of small GTPase mediated signal transduction	177	3.87E-06	3.29E-04
14	GO:0016323	basolateral plasma membrane	122	4.98E-06	1.48E-04
15	GO:0071702	organic substance transport	209	5.65E-06	4.01E-04
16	GO:0010876	lipid localization	98	5.83E-06	4.01E-04
17	GO:0019216	regulation of lipid metabolic process	72	5.86E-06	4.01E-04
18	GO:0051272	positive regulation of cellular component movement	88	6.00E-06	4.01E-04
19	GO:0008610	lipid biosynthetic process	177	6.95E-06	4.28E-04
20	GO:0007264	small GTPase mediated signal transduction	220	7.09E-06	4.28E-04
21	GO:0007264	actin cytoskeleton	139	8.39E-06	2.08E-04
22	GO:0031960	response to corticosteroid stimulus	67	1.49E-05	7.51E-04
23	GO:0006690	icosanoid metabolic process	31	2.12E-05	9.84E-04
24	GO:0051384	response to glucocorticoid stimulus	62	2.27E-05	9.86E-04
25	GO:0043434	response to peptide hormone stimulus	154	2.43E-05	1.01E-03

Table 5.2. Top 25 GO terms identified from applying ChIP-Enrich to the GR dataset. Only GO terms with less than 500 genes are shown, as this removes large non-specific GO terms.

5.3.7 Distance of peak to TSS and GO term enrichment

We investigated whether the same GO terms appeared enriched when we restricted our analysis to regions within 1 kb of the TSS (compared to analysis of all peaks) (Figure 5.7). In the GR dataset, the majority of peaks are >10 kb from a TSS. Of the 188 GO terms that showed significant enrichment in an analysis of all GR peaks, 177 were not detected as significantly enriched based on analysis of GR peaks within 1 kb of a TSS (using an FDR <.05 to detect GO terms that are significantly enriched.) Table 5.3 lists the top 20 most significant GO terms detected only with analysis of all peaks. All of these terms overlap with the top 25 GO terms overall, including “angiogenesis.” In contrast, only a single GO term, “basement membrane”, was significant for the 2 kb analysis but not for the all peaks analysis. Using only peaks near the TSS would miss GO terms of interest for GR, and likely for other transcription factors with similar patterns of binding. In contrast to GR, the majority of E2F4 peaks occur <1 kb of a TSS. Performing analysis with either all peaks, or only those peaks within 1 kb of a TSS, for E2F4 yields very similar results (Figure 5.8).

GO ID	GO Description	P-value All	FDR All	P-value 1KB	FDR 1KB
GO:0001525	angiogenesis	3.86E-08	1.13E-05	4.19E-02	5.43E-01
GO:0048514	blood vessel morphogenesis	1.61E-07	3.35E-05	5.39E-03	1.89E-01
GO:0006112	energy reserve metabolic process	2.86E-07	5.35E-05	5.55E-01	9.93E-01
GO:0001944	vasculature development	5.68E-07	8.84E-05	3.10E-03	1.56E-01
GO:0042598	vesicular fraction	6.48E-07	3.85E-05	2.24E-03	8.32E-02
GO:0001568	blood vessel development	6.55E-07	9.42E-05	7.50E-03	2.34E-01
GO:0005792	Microsome	8.27E-07	4.09E-05	3.45E-03	1.03E-01
GO:0019904	protein domain specific binding	1.04E-06	1.37E-04	1.62E-01	9.64E-01
GO:0042493	response to drug	3.77E-06	3.29E-04	7.07E-04	9.44E-02
GO:0051056	regulation of small GTPase mediated signal transduction	3.87E-06	3.29E-04	9.71E-01	1.00E+00
GO:0071702	organic substance transport	5.65E-06	4.01E-04	1.12E-01	7.50E-01
GO:0010876	lipid localization	5.83E-06	4.01E-04	6.04E-02	6.07E-01
GO:0019216	regulation of lipid metabolic process	5.86E-06	4.01E-04	5.38E-01	9.93E-01
GO:0051272	positive regulation of cellular component movement	6.00E-06	4.01E-04	2.60E-02	4.64E-01
GO:0008610	lipid biosynthetic process	6.95E-06	4.28E-04	7.04E-01	1.00E+00
GO:0007264	small GTPase mediated signal transduction	7.09E-06	4.28E-04	2.47E-01	8.86E-01
GO:0015629	actin cytoskeleton	8.39E-06	2.08E-04	6.35E-01	9.81E-01
GO:0031960	response to corticosteroid stimulus	1.49E-05	7.51E-04	4.21E-04	7.87E-02
GO:0006690	icosanoid metabolic process	2.12E-05	9.84E-04	5.71E-02	6.04E-01
GO:0051384	response to glucocorticoid stimulus	2.27E-05	9.86E-04	3.45E-03	1.61E-01

Table 5.3. Significant GO terms when using all GR peaks that lose significance after restricting to peaks within 1 kb of the TSS. The top 20 GO terms with FDR < 0.05 using all peaks and FDR > 0.05 when using peaks near the TSS are listed.

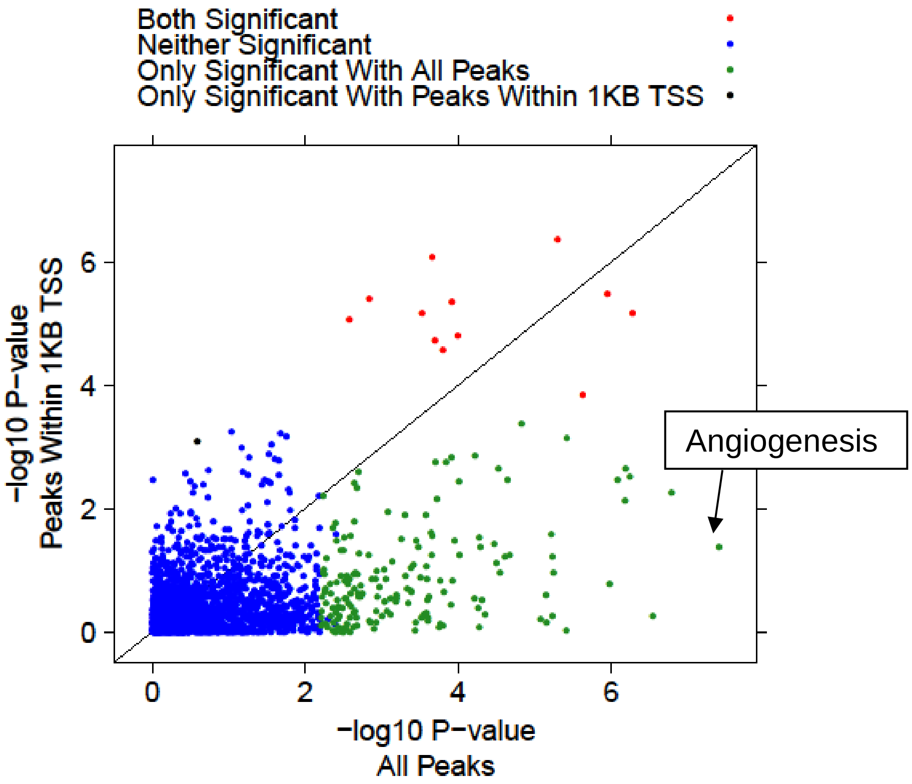


Figure 5.7. Comparison of p-values for ChIP-Enrich using all peaks in the GR dataset and only those peaks within 1 kb of a TSS. Each point represents a GO term, colored by whether it was significant at an FDR of 0.05 for either the case of using all peaks in the dataset (green), using only those peaks near the TSS (black), neither (blue), or both (red.) Only those GO terms with fewer than 500 genes are shown. FDR adjustment is performed separately for each GO branch.

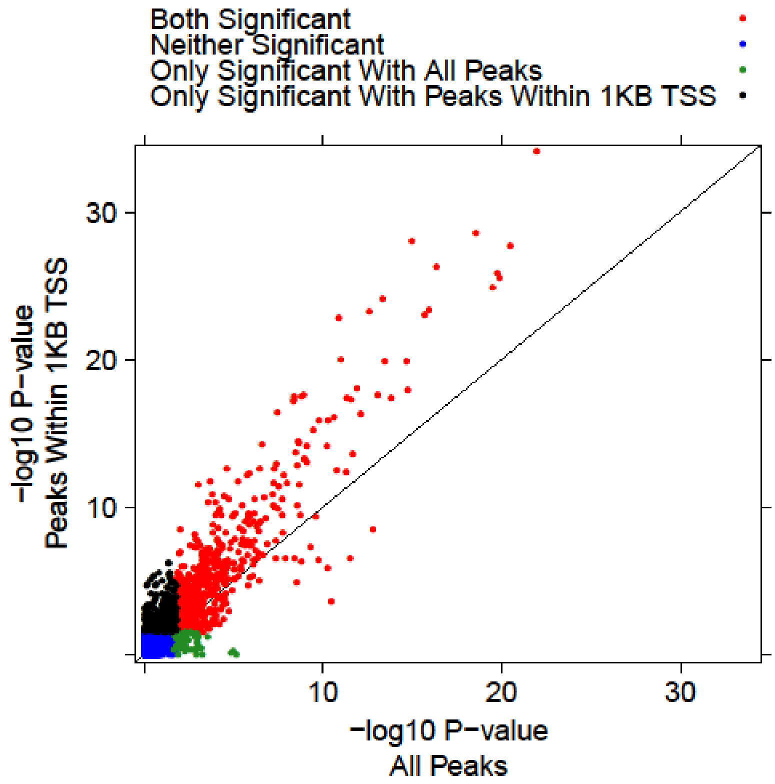


Figure 5.8. Comparison of p-values for ChIP-Enrich using all peaks in the E2F4 dataset and only those peaks within 1 kb of a TSS. Each point represents a GO term, colored by whether it was significant at an FDR of 0.05 for either the case of using all peaks in the dataset (green), using only those peaks near the TSS (black), neither (blue), or both (red.) Only those GO terms with fewer than 500 genes are shown. FDR adjustment is performed separately for each GO branch.

5.4 Discussion

Biological interpretation of genome wide deep sequencing applications such as ChIP-Seq, MeDIP-Seq, and others that result in a list of identified genomic regions, is complicated by technical false positives and false negatives, the high number of peaks often detected, uncertainties in association between regulatory regions and transcribed elements, and the possibility that a large percentage of bindings do not directly regulate any gene. We developed ChIP-Enrich, a gene set enrichment testing method for ChIP-seq and other sequencing experiments, that eliminates the need to assume a specific relationship between the probability of a peak and gene locus length by modeling it based on the data. We show that two existing approaches, Fisher's exact test and the binomial test, make assumptions regarding the relationship between the probability of observing a peak and the estimated locus (regulatory) length of a gene that are not consistent with the patterns observed in experimental data. These assumptions can lead to inflated significance levels and an excess of false positive enrichment results. While both Fisher's exact test and the binomial test result in a true enrichment signal beyond the level of bias for both GR and E2F4, this may not always hold true. For transcription factors with weaker signal, or for differential histone or DNA methylation studies, there may be a greater level of noise, and the bias in the Fisher's exact and binomial tests may outweigh the true signal, resulting in the majority of significant findings being false positives. In contrast with these previous tests for enrichment testing with ChIP-Seq data, ChIP-Enrich uses a data-dependent approach to model the relationship between the probability of observing a peak and the estimated regulatory (locus) length of a gene. This approach to empirically modeling the relationship, rather than making incorrect assumptions regarding its form, should result in a test that has the expected type 1 error rate for other experimentally observed data.

In addition to locus length, we also show that mappability plays a role as a potential confounder, by modifying the observable locus length in deep sequencing experiments, and like locus length, the mappability of genes is correlated with GO term membership. Our estimates of mappability represent a first order approximation of genomic regions that are capable of producing a peak when binding is present. Improved estimation of which genomic regions are mappable and able to be sequenced by commonly used technologies such as Illumina HiSeq-2000 will improve the estimation of the mappable locus length.

The binomial test applied in GREAT (McLean et al. 2010) uses the total number of binding sites associated with a gene set rather than the number of genes with at least 1 peak designated to that gene set. This assumes that a greater number of peaks indicates that a gene is more likely to be regulated. However, the strength of gene regulation is not necessarily proportional to the number of peaks binding near it (Rye et al. 2011). As demonstrated in our simulations (Supplementary Results 5.6.4) using randomly generated GO terms, a single gene with an unexpectedly high number of peaks can drive the enrichment results for multiple GO terms. Future research is required to determine the best balance between use of the number of binding sites and the number of genes with binding sites, which is likely to be experiment-dependent.

Using experimental ChIP-seq datasets, we found that ChIP-Enrich is able to identify GO terms related to the known biology of the transcription factors GR and E2F4. The glucocorticoid receptor-alpha (GR) acts as a ligand receptor for a class of steroid hormones known as glucocorticoids. Glucocorticoids are known to be involved in numerous biological processes, including the regulation of glucose synthesis in response to circadian rhythms, suppression of inflammation in response to stress, inhibition of glucose uptake, activation of lipolysis in adipose tissue, fetal development, angiogenesis, and many other processes (Geley et al. 1996; Dostert and Heinzel 2004). GR-alpha binds glucocorticoids in the cytoplasm, and then translocates to the nucleus to subsequently bind to glucocorticoid response elements (GREs). Binding of the GR complex can serve to either inhibit or activate the expression of genes. Reddy et al (2009) identified differentially expressed genes responsive to treatment by dexamethasone, which acts as a synthetic glucocorticoid that stimulates expression of GR-regulated genes. Through gene set enrichment analysis using DAVID (Dennis et al. 2003), they uncovered many possible underlying pathways for response to glucocorticoids, including stress response, organ development, cell differentiation, hormone secretion, and apoptosis. Our gene set enrichment analysis using ChIP-Enrich on the GR dataset matched many of these known functions, showing that ChIP-Enrich is capable of capturing known biology with a ChIP-seq dataset exhibiting binding distal to TSSs.

As demonstrated by our analysis of GR, functional binding may occur distally to genes. By including all peaks, we identified angiogenesis as a significant target pathway of GR. GR has been reported to play a role in angiogenesis (Dostert and Heinzel 2004; Small et al. 2005; Yano et al. 2006; Logie et al. 2010), even though it is not annotated to this function in Gene Ontology or other pathway databases. GR has also been linked to inhibition of angiogenesis through non-transcription factor activities (Leung et al. 2006), although the extent to which it directly regulates angiogenesis-related genes as a group is unknown. GR has recently been shown to mediate the inhibition of angiogenesis by corticosterone and 11-dehydrocorticosterone in wild-type mice (Small et al. 2005). It was also reported to repress proliferin gene expression in mice as well; proliferin is an angiogenic hormone normally present during development of the placenta, but is also expressed in the skin, and in immortalized cell lines (Dostert and Heinzel 2004). Another study in ARPE-19 cells showed that fluocinolone (a corticosteroid) inhibits VEGF secretion by acting through GR, thereby inhibiting angiogenesis (Ayalasomayajula et al. 2009). Logie et al. (2010) reported that glucocorticoids inhibit tube-like structure formation in human endothelial cells, and that this inhibition is mediated by GR. Here, we identified angiogenesis as a major target of regulation for GR, and identified the GRE motif *de novo* using only peaks assigned to angiogenesis-related genes, and found that this motif occurred in the majority of peaks. Because binding associated with angiogenesis occurred distal to genes, this finding would have been missed if only looking at peaks proximal to transcription start sites.

One way to use Fisher's exact test for gene set enrichment testing on ChIP-seq data is to only consider peaks within a few kb of TSSs to remove the confounding effect of locus length. We showed through our analysis of the GR dataset that using only those peaks near the TSS would result in missing the enrichment of biologically relevant sets of genes. Limiting analyses to sites proximal to the TSS is also prone to error when the start site(s) is not definitively known, such as in cases where a gene has a different TSS across multiple tissue types, or alternative first exons.

One primary limitation of existing methods, including our own, is the assumption that true binding is independent of gene locus length. In the event that a transcription factor specifically targeted long genes, existing methods including ChIP-Enrich would likely suffer reduced power to detect enrichment. Similar methods for gene set enrichment in RNA-seq data have the same limitation (Young et al. 2010). To our knowledge, no such evidence for transcription factors binding preferentially to short or long genes has been described. We also note that existing approaches are limited by the current state of the art for gene and transcript definitions, and by the functional annotation of genes and their membership in GO terms (or other sets of genes.) As these definitions improve, inference regarding the true biological function of a protein will also improve.

5.5 Supplementary Methods

5.5.1 Additional simulations of the ChIP-Enrich type 1 error rate

We performed additional simulations with ChIP-Enrich to demonstrate the discrete nature of the deflation in type 1 error, which is more readily apparent with a greater number of simulation iterations. Increasing the number of iterations also allows us to examine the type 1 error rate at smaller nominal levels of significance. Rather than simulate the type 1 error of ChIP-Enrich by performing enrichment testing of all GO terms, we instead chose a number of terms that are representative of the possible ranges for median locus length and number of genes in the GO term. In Figure 5.11 we plot the $\log_{10}$ number of genes against the $\log_{10}$ median length of genes in each GO term. Points circled in red represent the chosen GO terms for our simulations, selected to represent the extreme edges of the distribution. Category GO:0008610 was chosen because it lies at the median of both marginal distributions. Table 5.7 lists each chosen GO category.

To assess the type 1 error rate of the ChIP-Enrich, we simulated data under the null assumption of peaks occurring randomly throughout the genome. We sampled genes according to their locus length with replacement to simulate peaks occurring randomly within genes across the genome. The simulation was performed by 1) randomly sampling 7500 genes (by their locus length, with replacement), where sampled genes are considered to have a peak, and 2) calculate the p-value using ChIP-Enrich to test for enrichment over 5 selected GO terms (Table 5.7, Figure 5.11.) We performed 1 million iterations of this simulation and calculated the type 1 error rate for nominal significance levels 10^{-1} to 10^{-4} (Figure 5.12.)

5.5.2 Testing for enriched GO terms with genes of longer (or shorter) than average locus length

We used DAVID (Dennis et al. 2003) to test for enrichment of genes with longer and shorter locus lengths. The 500 genes with the longest locus lengths were input into DAVID as the test set gene list, and the background gene list was set to all of the genes in our gene list. The same process was repeated with the 500 genes that had the shortest locus lengths. Results were limited to GO terms with <2,000 genes and FDR<0.05 in order to report more specific categories.

5.5.3 Testing for enriched GO terms with genes having higher or lower than average mappability

We used LRpath (Kim et al. 2012) to test for GO terms that were enriched with genes that had higher and lower than expected mappability values. LRpath normally takes a list of p-values as input, which are then log-transformed in the logistic regression model. In order to preserve mappability, values were exponentiated prior to input in the LRpath web application (lrpath.ncibi.org). Results were limited to GO terms with <2,000 genes and FDR<0.05 in order to report more specific categories.

5.5.4 Comparing two alternative locus definitions for assigning peaks to genes: nearest TSS and nearest gene definitions

An alternative locus definition for assigning peaks to genes is to assign peaks to the nearest gene, rather than the nearest TSS. We might consider this locus definition for proteins known to bind distal to TSSs, or for other types of data, such as regions of histone modifications or methylation, as these types of regions could be more likely to regulate the nearby gene, and not necessarily the gene with the closest TSS.

To assign peaks to the nearest gene, we define a locus as the genomic region bounded on the upstream side by the midpoint between the gene’s most upstream TSS or TES and the closest TSS or TES of the next upstream gene, regardless of which strand the adjacent upstream gene is on. Likewise, the gene is bounded on the downstream side by the midpoint between the gene’s most downstream TES or TSS and the closest TSS or TES of the next downstream gene, regardless of which strand the adjacent downstream gene is on. When genes physically overlap (excluding genes completely contained within another gene), the overlapping region is divided at the midpoint of the overlap. If a gene is completely overlapped by another, the nested gene is not assigned a locus, except when the nested gene is entirely within an intron of the outermost gene.

To compare how these two locus definitions could affect the enrichment results, we used ChIP-Enrich on all 8 ChIP-seq peak datasets (Methods 5.2.7) to test for enrichment of GO terms using both locus definitions of peak assignment, and plotted the $–\log_{10}$ p-values from the two locus definitions against each other (Figure 5.13).

5.5.5 Comparing alternative models for testing gene set enrichment

We tested each of the 8 ChIP-seq peak datasets (Methods 5.2.7) for GO term enrichment using a number of potential models. Our default choice of model was a binomial smoothing spline, fit with 10 knots, on the $\log_{10}$ mappable locus length. Using this model, we identified the top 15 GO terms from each dataset, and compared the $–\log_{10}$ GO term enrichment p-values from alternative choices of models. We first considered binomial smoothing spline models, and varied the number of knots used in estimating the spline (2, 5, 10, and 25 knots.) Each of these models were fit on the $\log_{10}$ mappable locus length. We also included a binomial smoothing spline model of 10 knots fitted on mappable locus length without $\log_{10}$ transformation. In addition to smoothing spline models, we considered three logistic regression models, which included all 1st, 2nd, and 3rd order $\log_{10}$ mappable locus length terms, respectively.

5.5.6 Tests for enrichment of peaks within Gene Ontology genes

We wished to examine whether genes annotated in GO were enriched for peaks relative to those genes not annotated in GO. We used a two-sided Fisher's exact test to test for association between whether a gene was annotated in GO, and whether it was assigned a peak. We also used a one-sided binomial test on whether the proportion of peaks in GO was greater than expected, where the probability parameter (the likelihood of a peak occurring in GO) was the sum of the locus lengths of genes in GO, divided by the total mappable genome length ($\approx 2.7 \times 10^9$ bp) (Rozowsky et al. 2009). We applied both tests to all 8 ChIP-seq datasets separately across each branch of GO (cellular component, molecular function, and biological process) (Table 5.8.)

5.5.7 Effect of varying genome length on the binomial test

We used the binomial test as originally described in (McLean et al. 2010) on the GR ChIP-Seq peak dataset and varied the genome length parameter across a range of values, starting with the approximate mappable genome length of 2.7 Gbp for hg19, and ending with the non-gapped genome length of 3.0 Gbp. The length of each gene is the length of the gene locus region defined in Methods 5.2.1.

5.5.8 Simulations of random sets of genes of varying sizes

We simulated 2500 sets of 50, 250, and 500 genes selected at random from the total set of Entrez genes for which we previously generated a gene locus region (see Methods 5.2.1). For each overall set of random gene sets, we tested two experimental ChIP-Seq datasets (E2F4, GR) with Fisher's exact test, the binomial test, and ChIP-Enrich. We generated quantile-quantile (QQ) plots for each test to determine if any bias was present in the p-value distribution.

5.5.9 Motif search within peaks assigned to angiogenesis genes

We used tools from the MEME Suite (Bailey et al. 2009) to identify known transcription factor motifs. Multiple Em for Motif Elicitation (MEME) conducted an unsupervised search for a common motif in the 460 peak regions from the GR dataset that were assigned to genes in the angiogenesis gene set. We set MEME to search for the three most prevalent motifs with a maximum 15bp width anywhere in the input DNA regions and the reverse complement sequences of those regions, with all other settings on default. We used Tomtom (Gupta et al. 2007), a motif comparison tool that is part of the MEME suite, with the results from MEME to find the best matched motifs in JASPAR, a database that contains a curated set of transcription factor binding sites for eukaryotes that have been published and experimentally defined.

5.6 Supplementary Results

5.6.1 Comparing binomial smoothing spline and logistic regression models

We investigated logistic regression models including up to 3rd order $\log_{10}$ locus length terms as an alternative to the binomial smoothing spline approach by comparing model fits (Figure 5.18) and GO term enrichment p-values (Figure 5.20) for the top 15 GO terms using our default model (binomial smoothing spline, 10 knots, fit on $\log_{10}$ mappable locus length) from each dataset. The model fits in many cases are similar, with deviations from each other occurring only within the tail regions (with the exception of CMYC, which shows a strong difference for the 1st order logistic regression.) GO term enrichment p-values were similar when compared to binomial smoothing spline models (Figure 5.20) for most datasets, though in some cases the logistic regression models achieved higher $-\log_{10}$ p-values (1st order logistic regression for GR, TCF4, 2nd order logistic regression for GR and TCF4, and 3rd order logistic regression for VDR), and lower $-\log_{10}$ p-values in others (1st order logistic regression for VDR, E2F4.) Increasing the number of knots above 10 for the binomial smoothing spline models did not affect model fit (Figure 5.19) or inference (Figure 5.20), and resulted in high running times. Decreasing the number of knots below 5 resulted in more significant p-values for a small minority of GO terms (Figure 5.20). We compared the fit of the binomial smoothing spline model using mappable locus length and $\log_{10}$ mappable locus length, and while we observed some variation in spline fit compared to models fit on $\log_{10}$ length (Figure 5.19), we did not see any change in $-\log_{10}$ p-values (Figure 5.20). Overall, we observed little effect on $-\log_{10}$ p-values from choice of smoothing spline or logistic regression models, number of knots for fitting the smoothing spline models, or $\log_{10}$ transformation of mappable locus length. Further testing on additional experimental or simulated datasets could provide a better understanding of the differences between these choices of models and fitting methods.

5.6.2 Overestimating genome length causes inflation of the binomial test p-values

The original binomial test (McLean et al. 2010) uses a genome length parameter when computing the probability of a peak occurring within a GO term. This genome length is typically taken to be the total length of the non-gapped genome in GREAT (McLean et al. 2010). However, this total length is not accessible to next-generation sequencing technologies, but rather the subset that is mappable (reads may uniquely map to a site in the genome) and sequenceable (reads are capable of being generated from a site in the genome). Both of these properties will reduce the total length of the genome in which a ChIP-seq peak could be called. Therefore, we wished to examine the effect of overestimating the genome length on the binomial test. We applied the binomial test to test for enrichment amongst GO terms using the GR dataset with three genome length parameters: 2.7E9 (representing an estimate of the mappable genome (Rozowsky et al. 2009)), 2.8E9 (an intermediate value), and 3.0E9 (representing the approximate length of the non-gapped hg19 genome.) From this, we can observe two important points: 1) in general, overestimating the genome length causes an inflation of the binomial test, and 2) this effect is dependent on the total length of genes in a GO term, such that longer GO terms show the greatest shift in their p-values (Figure 5.14..) Our corrected binomial test (Methods 5.2.5) computes the genome length directly from the locus definitions, and therefore can account for the length of the mappable genome.

5.6.3 Additional simulations of ChIP-Enrich type 1 error rate

We performed additional simulations with ChIP-Enrich to demonstrate the discrete nature of the deflation, which is more readily apparent with a greater number of simulation iterations. We performed 1 million iterations of applying ChIP-Enrich to 5,000 peaks simulated by randomly sampling genes in proportion to their length with replacement. We selected five GO terms to represent a range of total gene lengths and number of genes. ChIP-Enrich appears slightly conservative in general (Figure 5.12.), however since both Fisher's exact test and ChIP-Enrich are tests on discrete data, the nominal type 1 error rate cannot always be achieved (Upton 1992).

5.6.4 Comparison of ChIP-Enrich, Fisher's exact test, and the binomial test using randomly generated gene sets with a constant number of genes

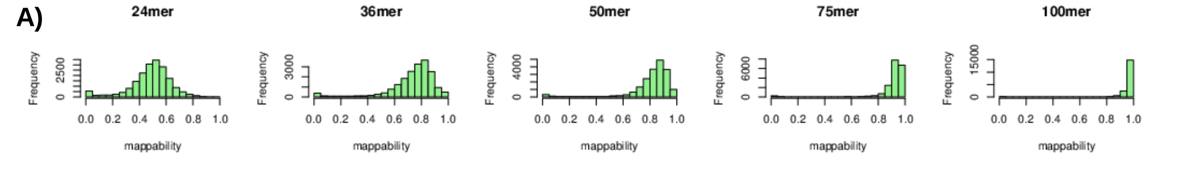
A gene set enrichment test should not detect enrichment when the sets of genes being tested are selected at random. We performed simulations in which we generated 2500 randomly sampled sets of genes of size 50, 200, and 500, and tested these using Fisher's exact test, the binomial test, and ChIP-Enrich on experimental ChIP-seq peak datasets. We generated QQ plots (Figure 5.15) for each combination of enrichment test and random gene set size, on two experimental datasets (E2F4 and GR). We see that with random gene sets, the binomial test detects an excess of significant gene sets. As we have shown, the binomial test makes a strong assumption regarding the distribution of the number of peaks that should occur within a set of genes, given the total length of all genes in the set. In experimental data with random sets of genes, this assumption does not always hold, as seen for the E2F4 and GR datasets where both datasets exhibit a larger number of peaks within gene sets of short total length than expected by chance (Figure 5.16.) In particular, the binomial test is biased towards detecting gene sets with shorter than average locus lengths as significant (Figure 5.17.) This can be caused by a single or few genes with an excess of peaks within the gene set. For example, we examined the random gene sets of size 200 that contained the gene with the most significant excess of peaks given its length, and noted that ≈48% (16/33) of these random gene sets were significant at an FDR < 0.05. It is also important to note that the binomial test does not take into account whether peaks in a gene set are clustered in a few genes or spread evenly across many genes, the locus lengths of those genes, and therefore the resulting p-value does not account for any extra-variability in peaks among the genes in the gene set. In contrast to the binomial test, Fisher's exact test does not detect significant associations, owing to the fact that the confounding effect of gene locus length on both membership in gene sets and the existence of a peak within a gene locus has been removed. ChIP-Enrich, much like Fisher's exact test, performs correctly with random sets of genes and detects no excess of significantly associated gene sets.

5.6.5 Assigning peaks to the nearest gene vs. the nearest TSS

We explored how assigning peaks to the nearest gene might affect our enrichment results as compared to assigning peaks to the nearest TSS across each of the 8 ChIP-seq peak datasets. For each dataset, we used ChIP-Enrich to test for GO term enrichment after assigning peaks under the two locus definitions (Supplementary

Methods 5.5.4). For those datasets where binding occurs proximal to the TSS (E2F4, TBP), the two locus definitions produce highly correlated p-values (Figure 5.13). In contrast, for datasets where binding occurs distal to the TSS (GR – VDR), the two locus definitions still remain correlated, though many GO terms greatly increase in significance (up to ≈10 orders of magnitude) under the nearest gene locus definition, suggesting for these transcription factors that regulation of a set of genes is not dependent entirely upon binding near their TSSs. We also see that some GO terms are more significant under the nearest TSS locus definition (cMyc – E2F4 datasets), perhaps because the traditional mechanism of gene regulation (binding near the TSS) is also present for these transcription factors. An example of both mechanisms exists in the literature for GR, where binding near the TSS activates transcription of target genes, and binding occurring distally to TSSs occurs in enhancer regions and results in repression of transcription (Reddy et al. 2009). Both locus definitions should likely be analyzed when performing analysis and it is possible that other definitions not considered here could improve inference in identifying enriched gene sets.

5.7 Supplementary Figures and Tables



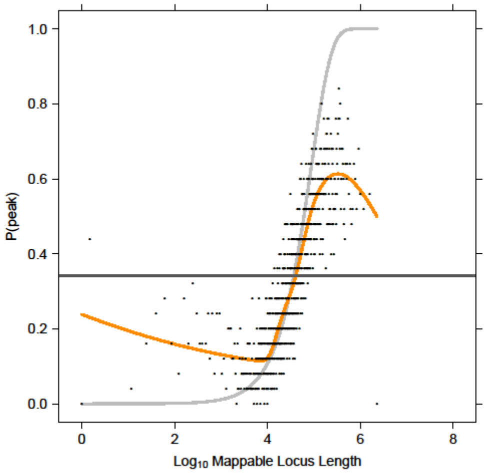
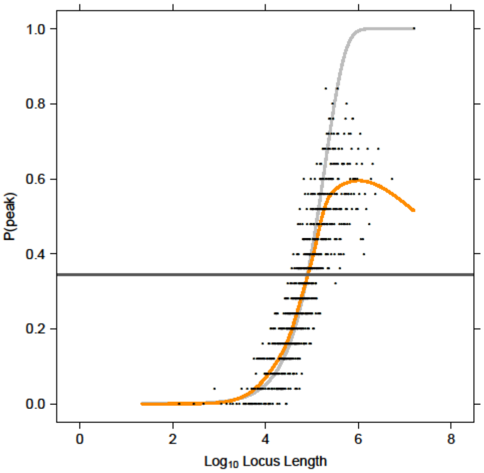
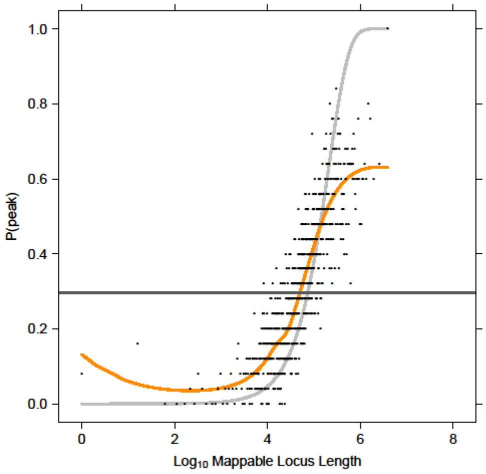
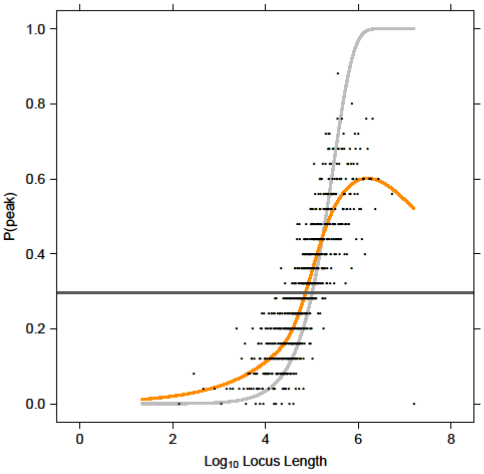
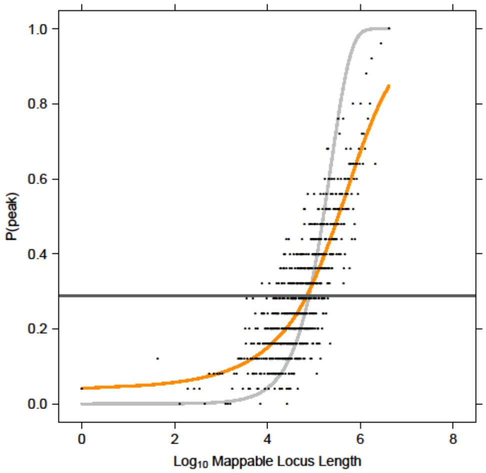
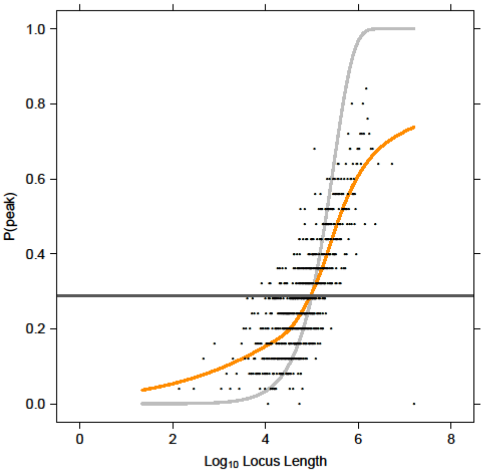
B) GO Terms whose Genes' Loci Have Higher Mappability

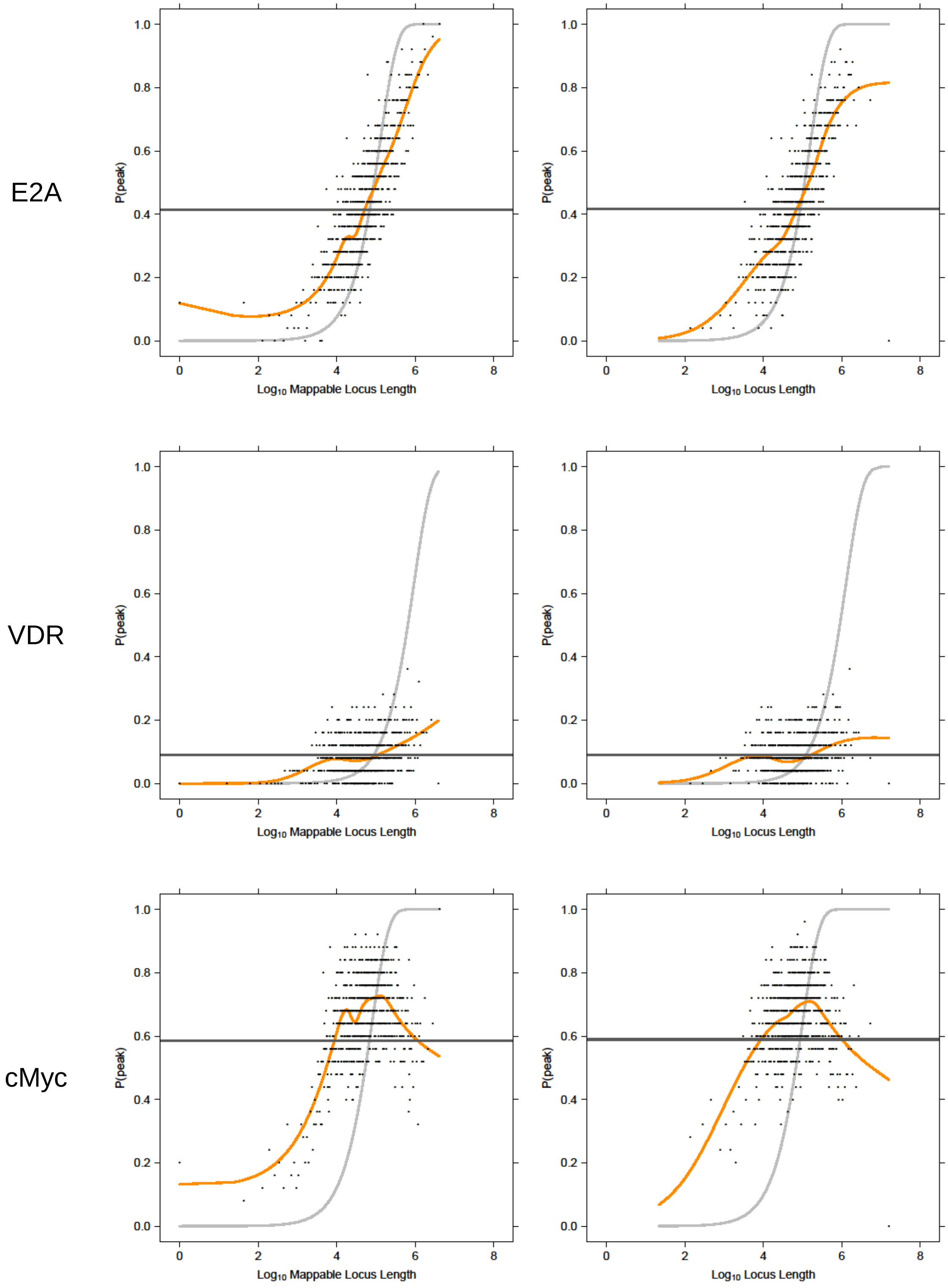
GO Term	# Genes	P-value	Q-value
Organ morphogenesis	642	2.6E-22	5.5E-19
Central nervous system development	454	2.9E-19	3.0E-16
Neurogenesis	634	1.4E-18	9.8E-16
Neuron differentiation	534	2.7E-18	1.4E-15
Cell development	786	5.5E-18	2.3E-15
Generation of neurons	589	1.6E-17	5.6E-15
Skeletal system development	272	2.6E-16	7.8E-14
Regionalization	217	1.9E-15	4.9E-13

C) GO Terms whose Genes' Loci Have Lower Mappability

GO Term	# Genes	P-value	Q-value
Olfactory receptor activity	114	1.6E-11	7.0E-09
Sensory perception of smell	131	1.3E-09	6.3E-08
Cellular defense response	60	3.0E-08	9.0E-07
Sensory perception of chemical stimulus	167	3.7E-08	1.1E-06
Oxygen binding	44	7.7E-08	8.7E-06
Cellular response to xenobiotic stimulus	35	2.2E-07	5.1E-06
Xenobiotic metabolic process	35	2.2E-07	5.1E-06
Electron carrier activity	156	4.9E-07	3.0E-05

Figure 5.9. Gene loci mappability distributions and correlation with Gene Ontology terms. Histograms show distribution of human (hg19) gene loci mappabilities for various sequencing read lengths (A). GO terms whose gene loci have significantly high (B) or low (C) mappability for 50-mer reads. GO biological processes and molecular functions were tested using the LRpath gene set enrichment program (Sartor et al. 2009).

Dataset	Spline w/ Mappability	Spline w/o Mappability
GR		
TCF4		
Beta Catenin		


E2A
VDR
cMyc
P(peak)
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
Log10 Mappable Locus Length
Log10 Locus Length

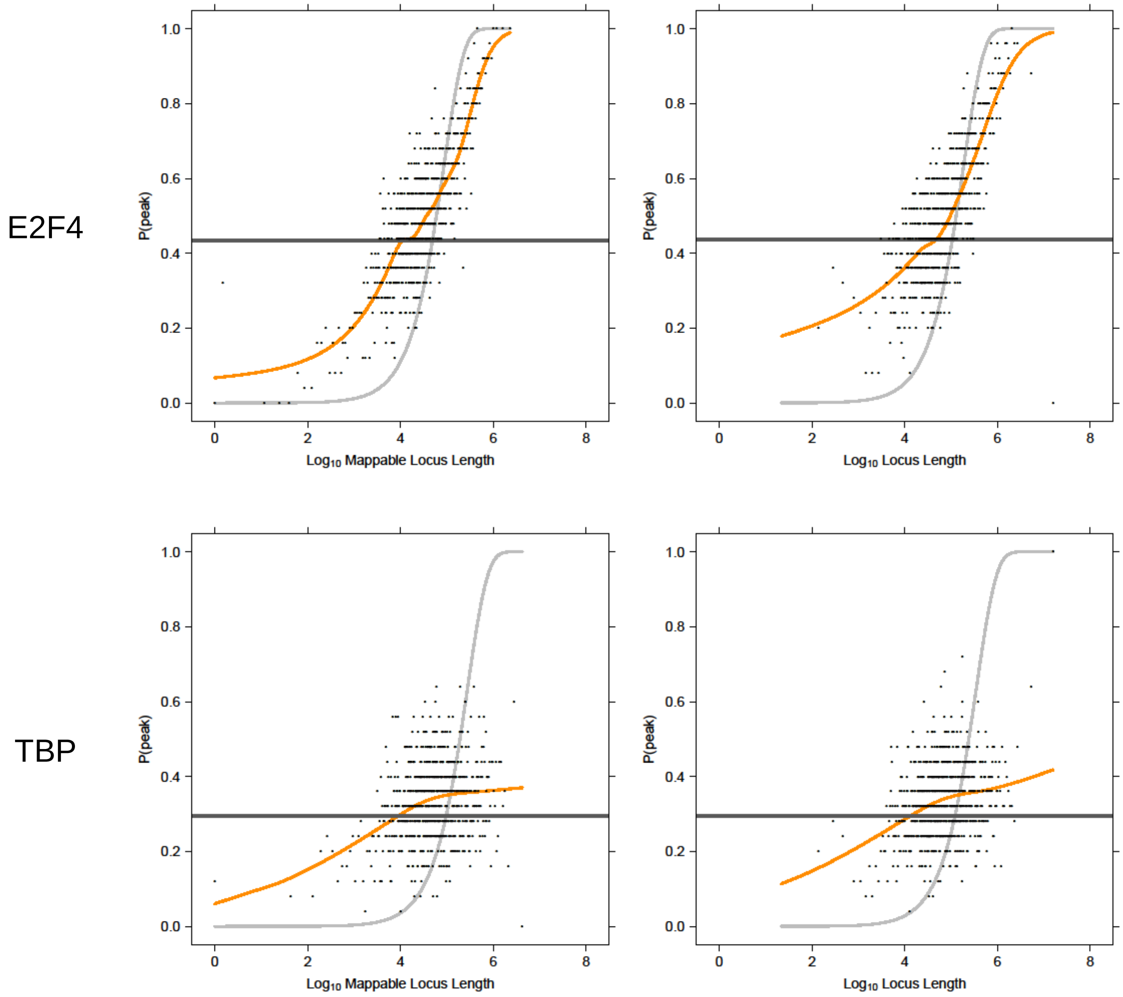


Figure 5.10. Comparison of spline fit with and without mappability adjustment. Each row corresponds to one of the eight ChIP-seq datasets selected for analysis. The plots show the fit of a binomial smoothing spline for the relationship between existence of a peak and either the log$_{10}$ gene locus length (right), or the log$_{10}$ mappable gene locus length (left.)

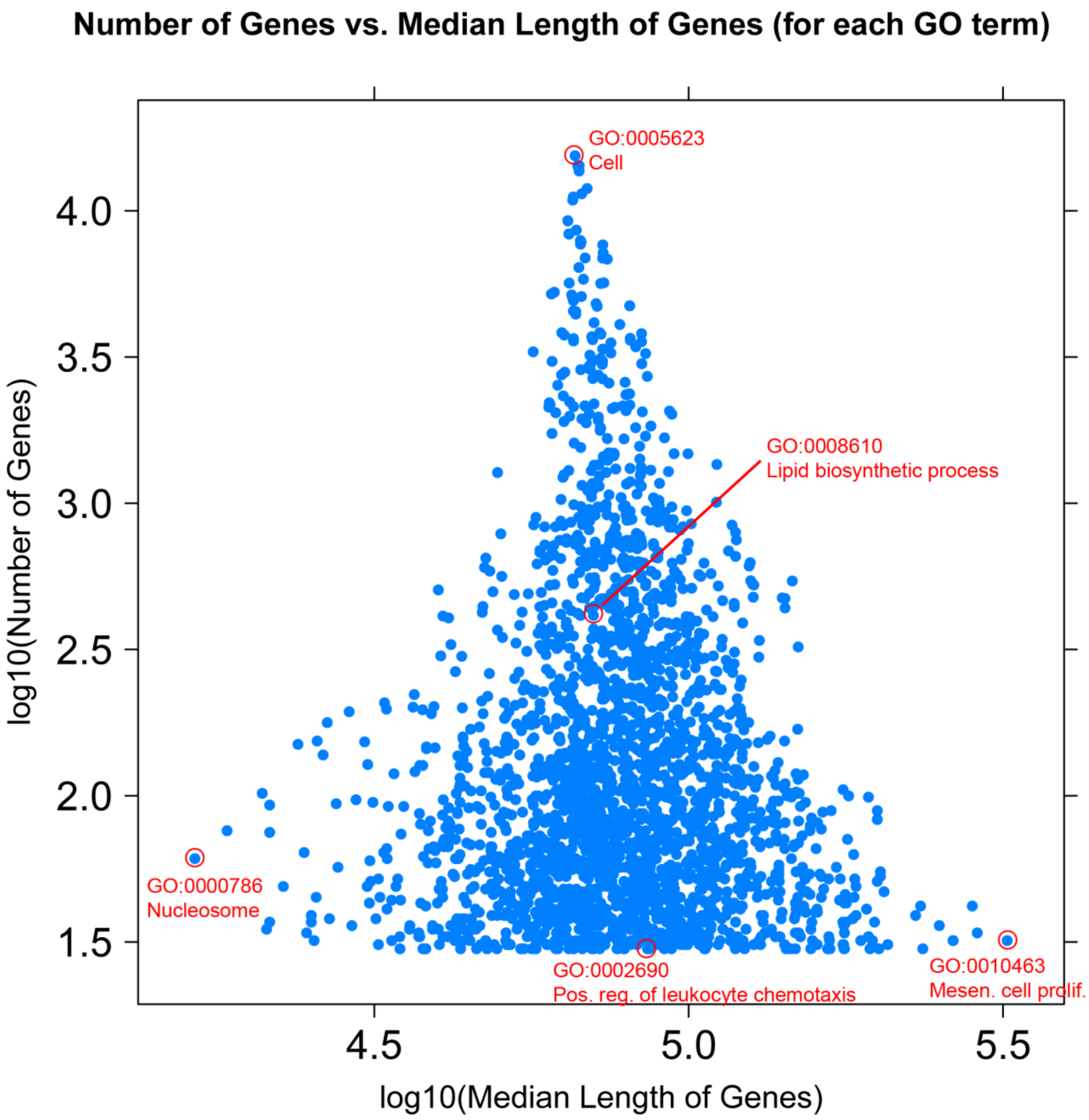
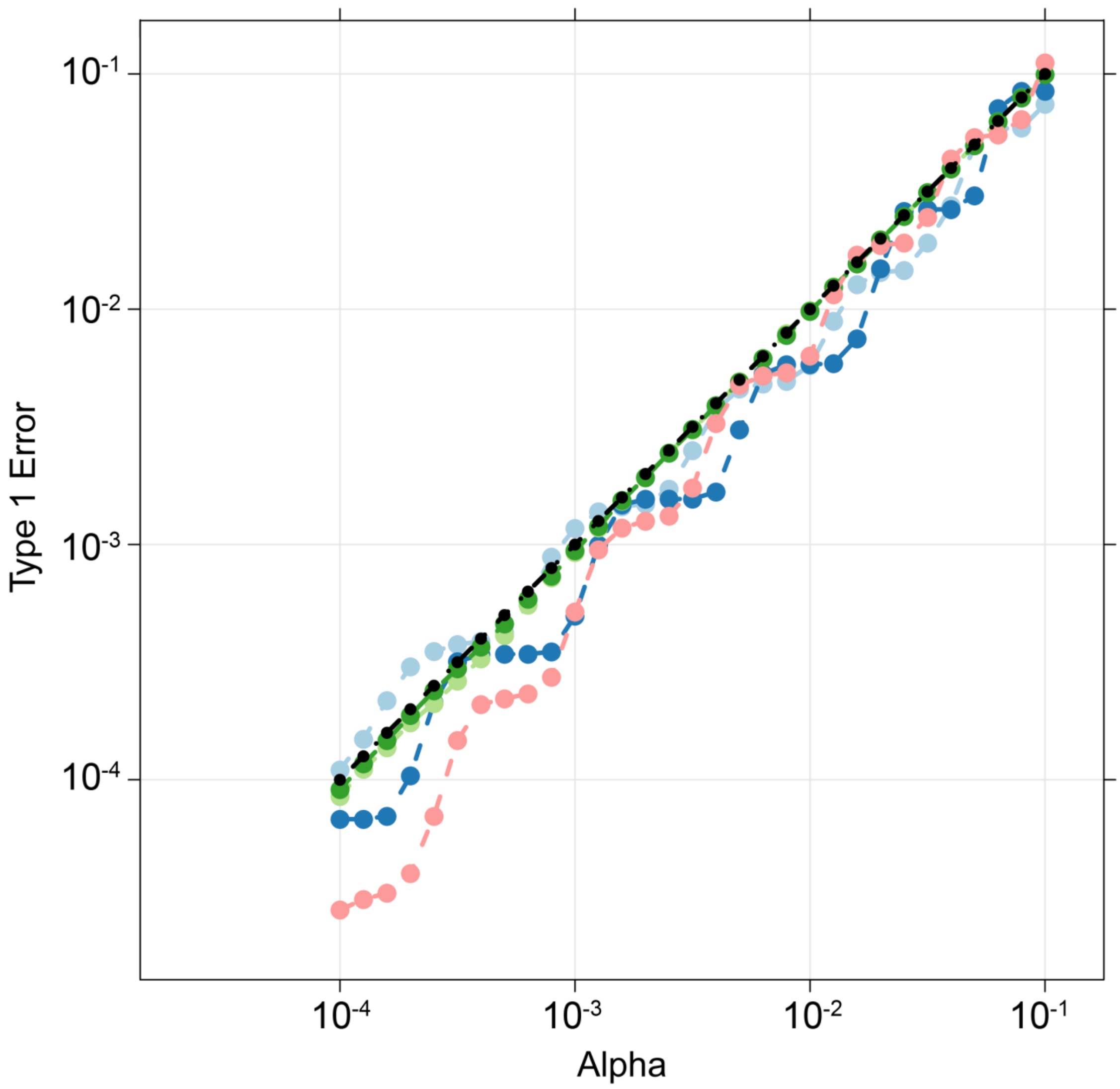


Figure 5.11. Gene Ontology terms selected for type 1 error simulations. We plot the log10 number of genes against the log10 median length of genes in each GO term. Terms selected for type 1 error simulations are circled in red.



GO Term	Number of Genes	Length of Genes	Key
GO:0000786	Few	Short	● – –
GO:0002690	Few	Median	● – –
GO:0005623	Many	Median	● – –
GO:0008610	Median	Median	● – –
GO:0010463	Few	Long	● – –

Figure 5.12. Type 1 error rate of ChIP-Enrich under the null model of random peaks occurring across the genome. The type 1 error rate of ChIP-Enrich is plotted against nominal levels of significance. We performed simulations of random peaks across the genome (random genes sampled by their locus length) independently for each of the 5 GO terms listed above. ChIP-Enrich appears to be conservative for GO terms with few genes of median locus length, and few genes of long locus length, though this is likely due to the fact that tests on discrete data cannot always meet the expected type 1 error rate and is further evidenced by the step-like nature of the curves shown above.

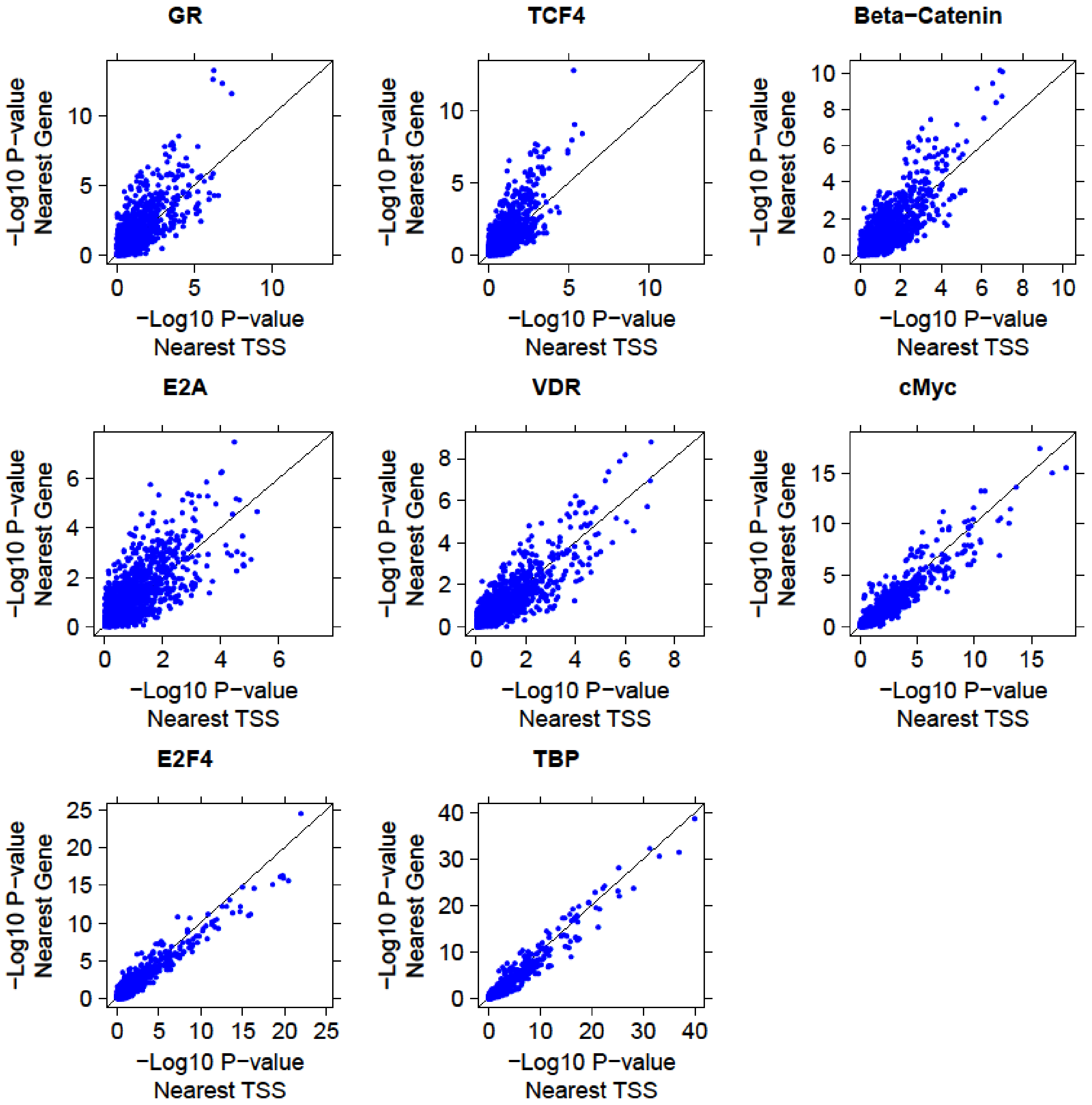


Figure 5.13. Comparison of alternative locus definitions for assigning peaks to genes: nearest gene and nearest TSS. For each of the 8 ChIP-seq peak datasets, we used ChIP-Enrich to test for GO term enrichment using both the nearest gene and nearest TSS locus definitions. We plot above the $-\log_{10}$ p-values from using the nearest gene definition against the $-\log_{10}$ p-values using the nearest TSS definition. While the p-values are correlated between the two definitions in each dataset, we can see a substantial difference between the two methods for certain GO terms. Examining the results from both definitions, and potentially others that we have not explored here, could be useful for identifying sets of genes regulated by alternative mechanisms.

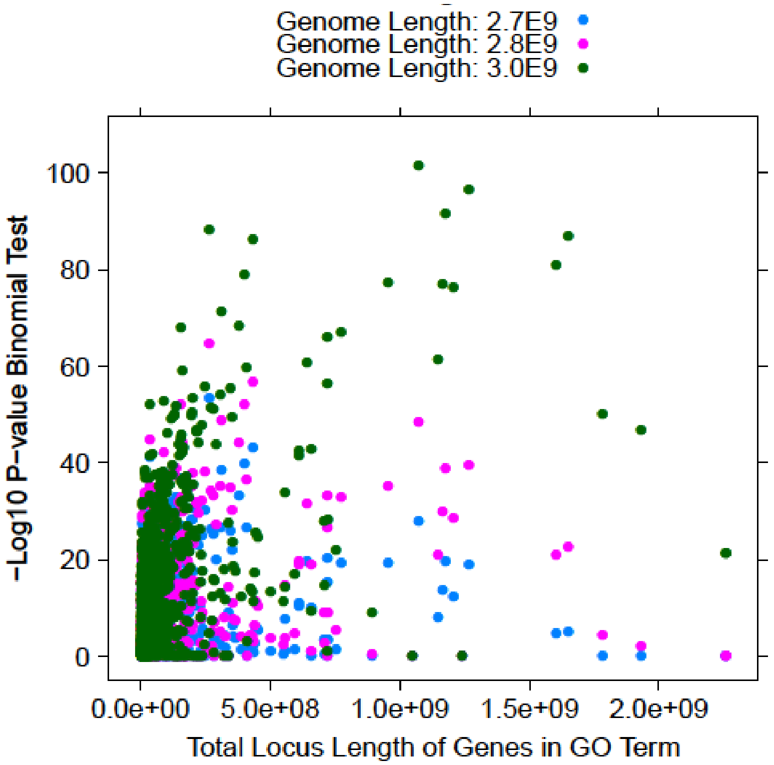


Figure 5.14. Binomial test p-values for Gene Ontology terms using a range of genome lengths. The –log10 p-value from the binomial test is plotted against the total length of the genes in each Gene Ontology term. Color represents the genome length value that was used for the binomial test. The binomial test here is the uncorrected version that uses a genome length constant when computing the probability of a peak occurring within a GO term.

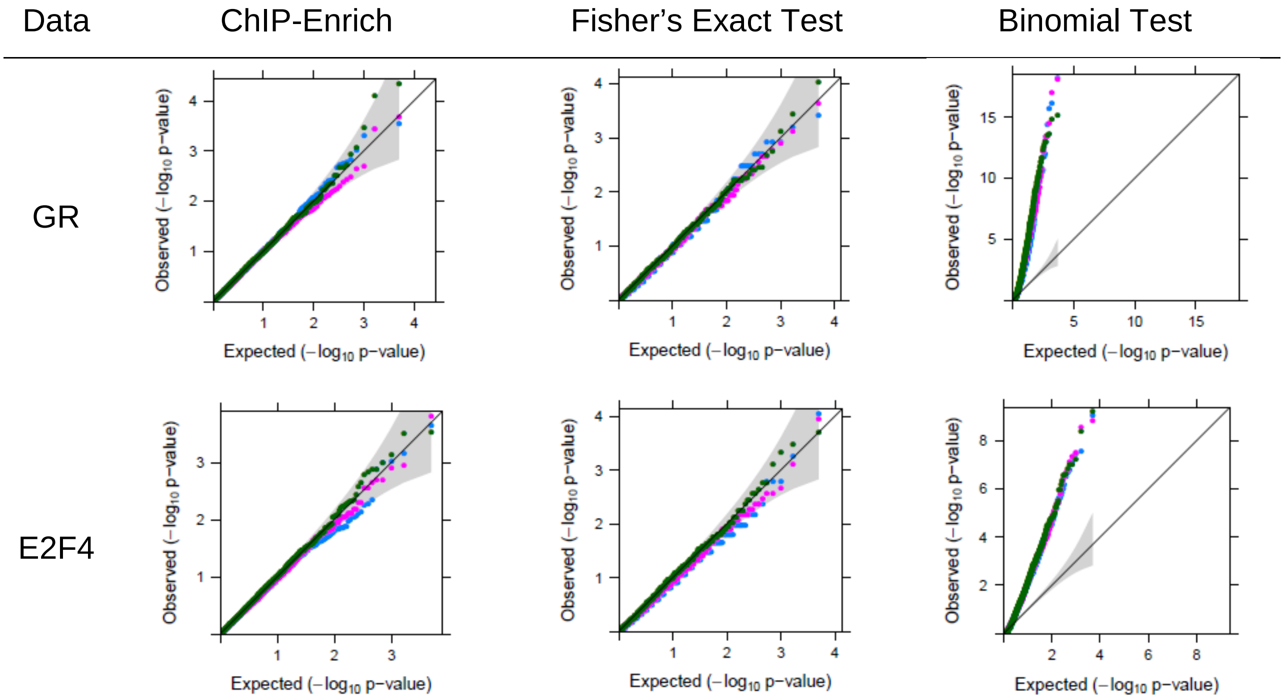


Figure 5.15. QQ plots for p-values from each enrichment test under random sets of genes and experimental data. Points represent p-values for 2500 random gene sets of size 50 (blue), 200 (pink), and 500 (green.) Gray regions represent the 95% confidence interval for the expected p-value quantiles. Each row corresponds to the GR and E2F4 datasets, and each column shows the results for either ChIP-Enrich, Fisher's exact test, or the binomial test.

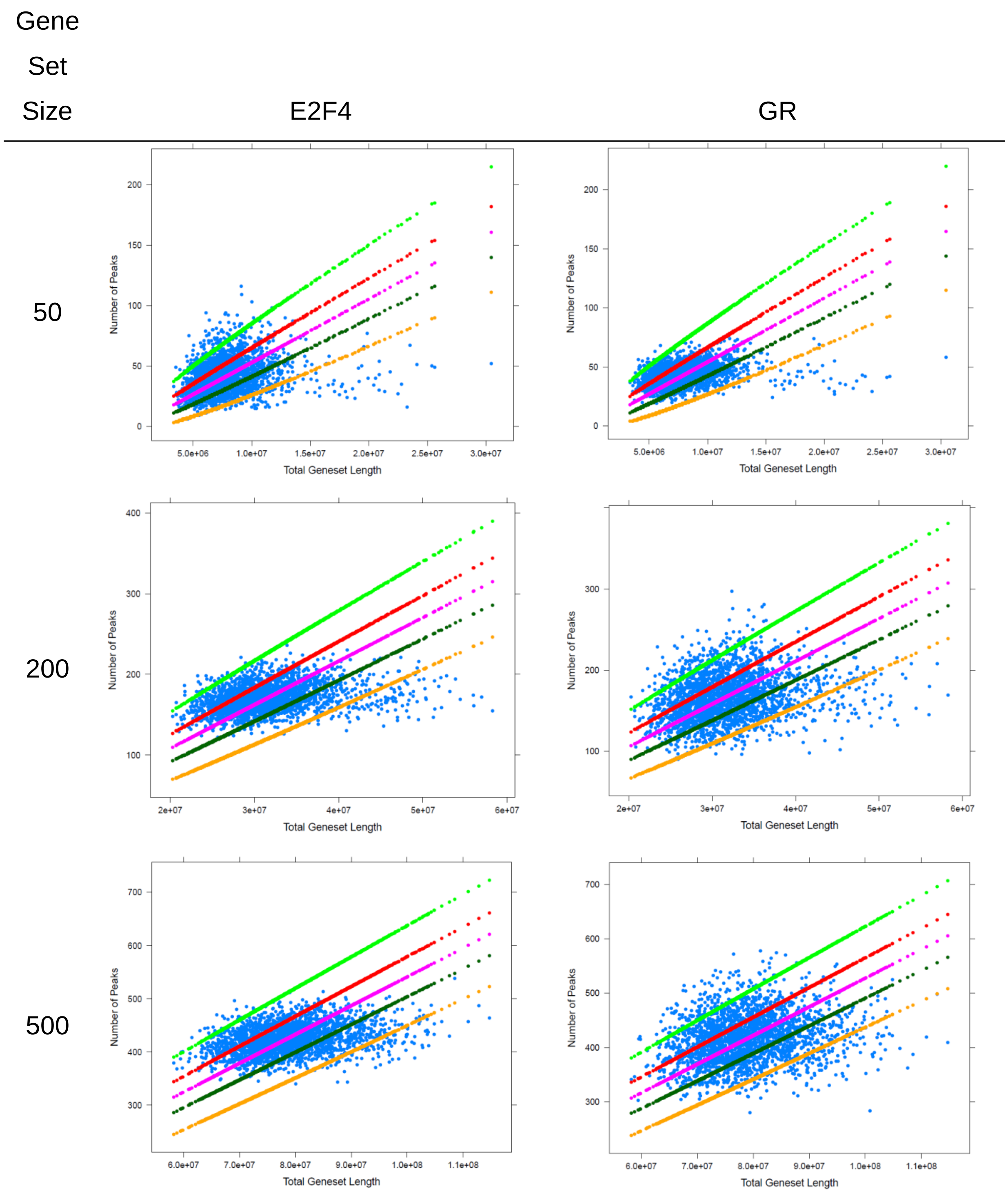


Figure 5.16. Number of peaks assigned to genes within random sets of genes plotted against the total length of genes in each set. Blue represents actual data. Pink represents the expected number of peaks, given a binomial distribution and probability of a gene set having a peak as the total length of genes in the set divided by the length of the non-gapped genome. Red and dark green represent the 5 and 95% quantiles for the count of peaks given the binomial distribution, and bright green/orange represent the Bonferroni adjusted 5 and 95% quantiles for the number of gene sets (2500).

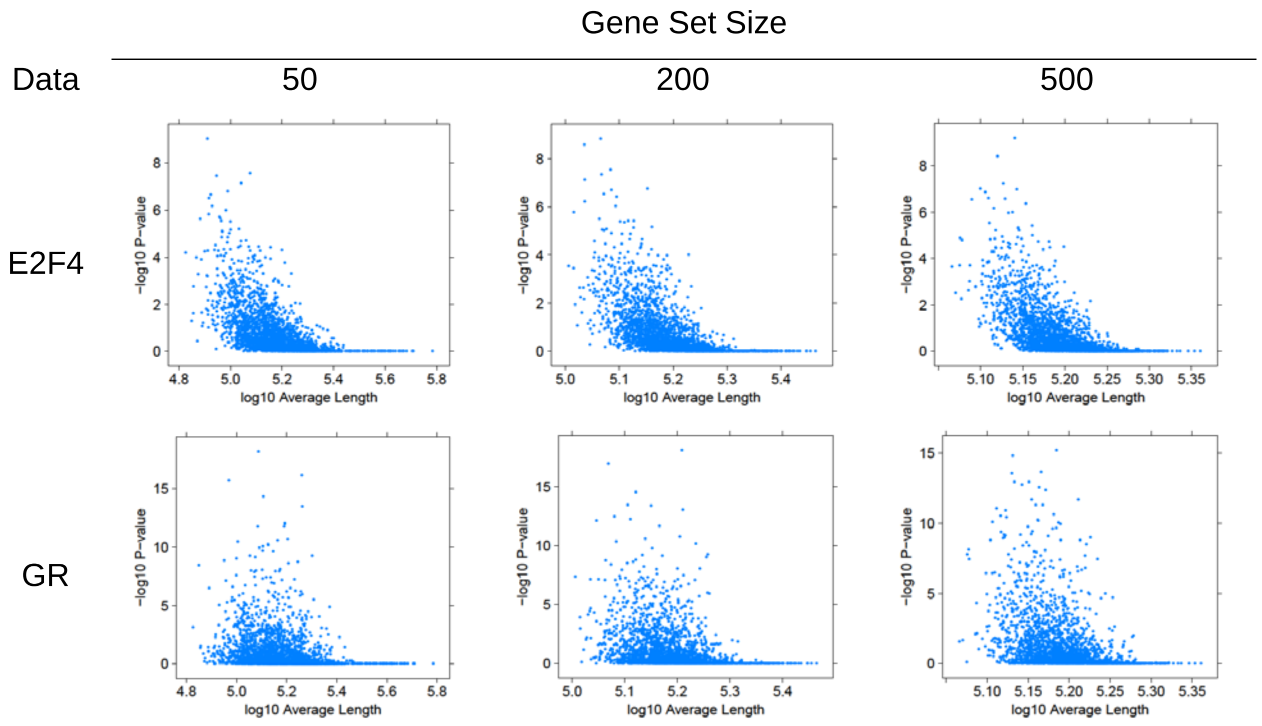


Figure 5.17. Binomial test p-values vs. log$_{10}$ average length of genes. The binomial test was applied to the E2F4 and GR datasets (rows), testing for enrichment with random sets of genes of size 50, 200, and 500 (columns.)

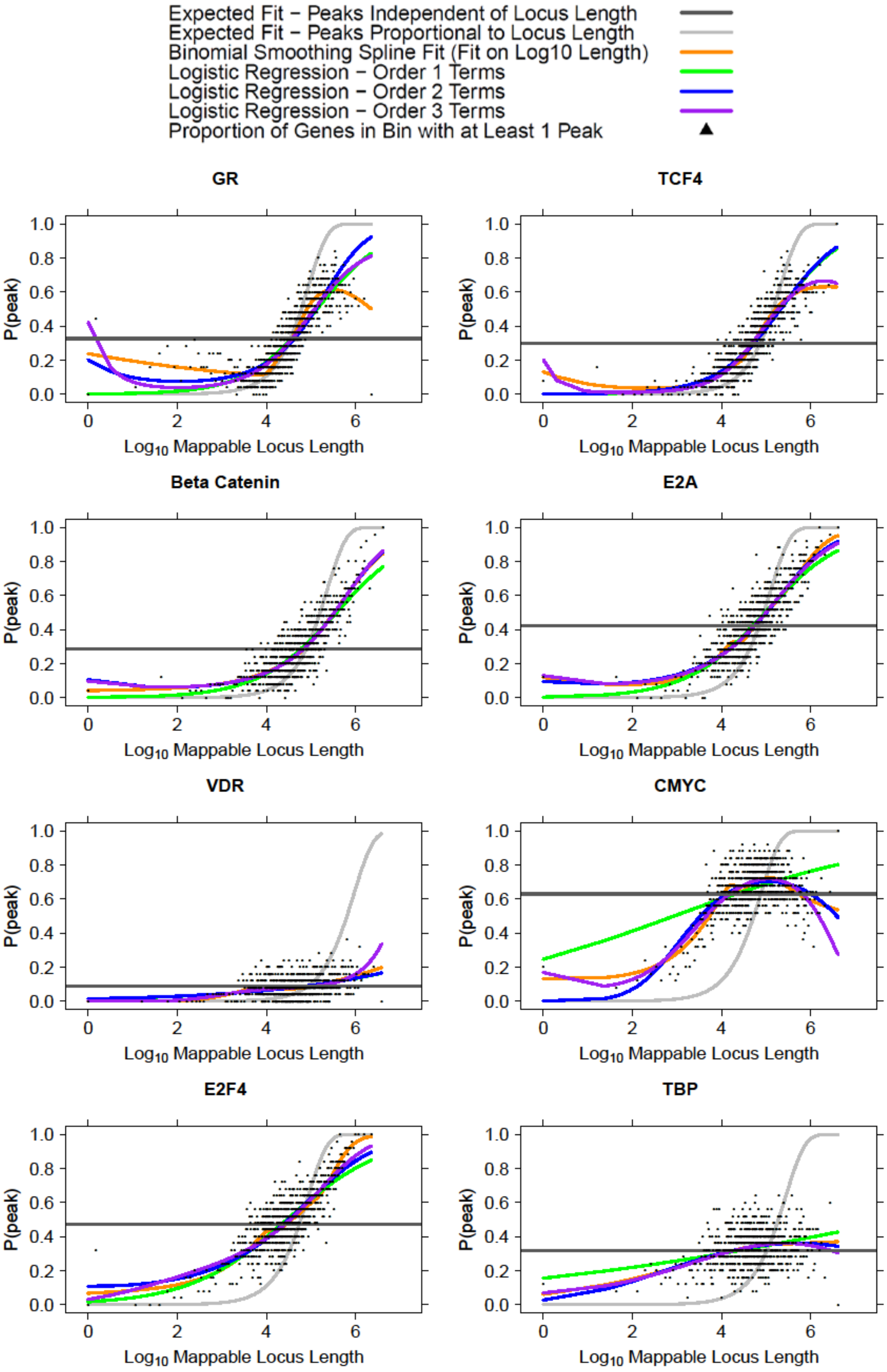


Figure 5.18. Comparing the binomial smoothing spline to logistic regression models. Each figure shows a spline fit plot for each of the 8 ChIP-seq peak datasets. The orange curve represents the fit from our binomial smoothing spline. The green, blue, and purple curves represent fits from using a logistic regression model with progressively higher order terms (green – up to first order terms, blue – up to second order terms, and purple – up to third order terms.) All models are fit on $\log_{10}$ mappable locus length.

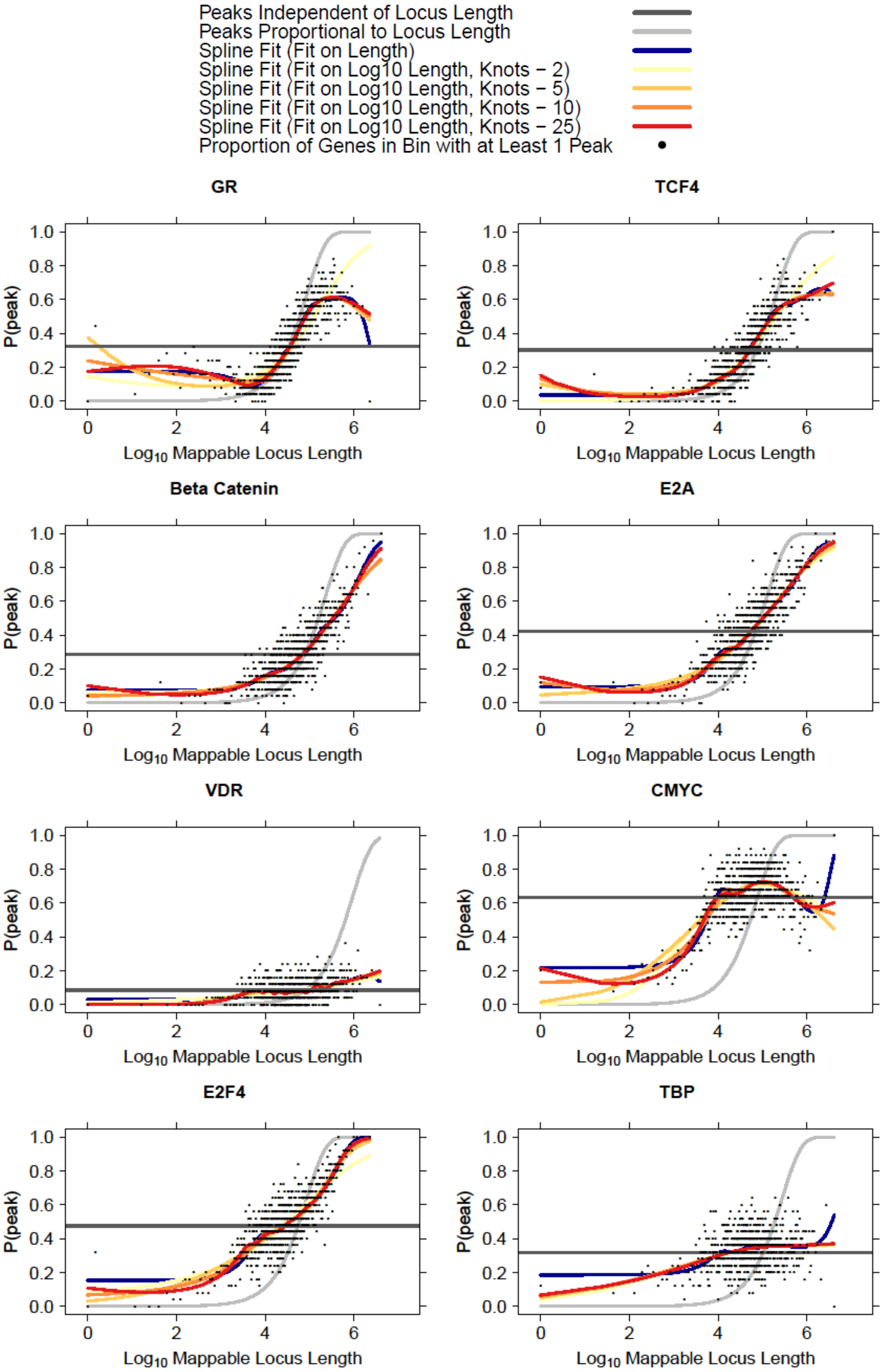


Figure 5.19. Binomial smoothing spline fits for potential models on experimental data. We plot a binomial smoothing spline fit to each of the 8 experimental datasets and vary the number of knots (2,5,10,25) used. We also include an additional alternative model where the spline is fit against linear scale mappable length (blue), rather than $\log_{10}$ mappable length.

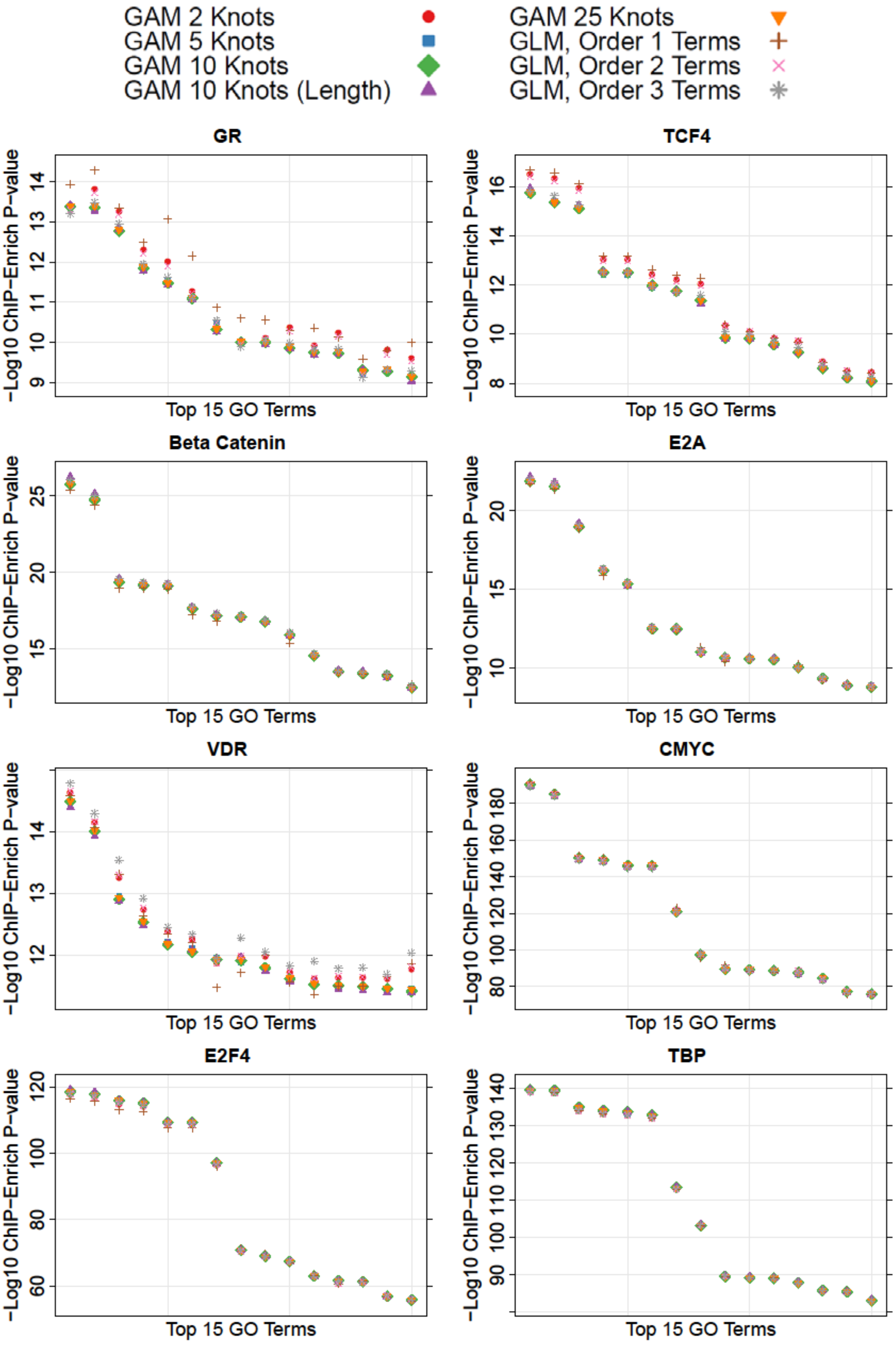
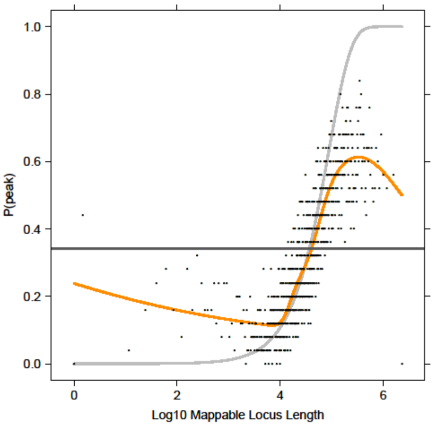
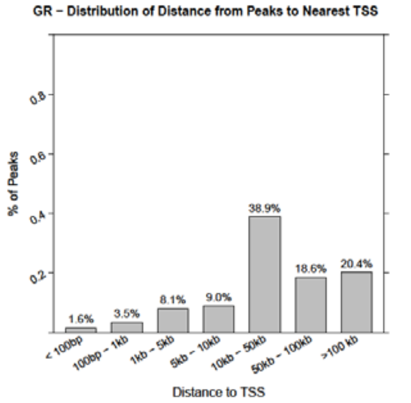
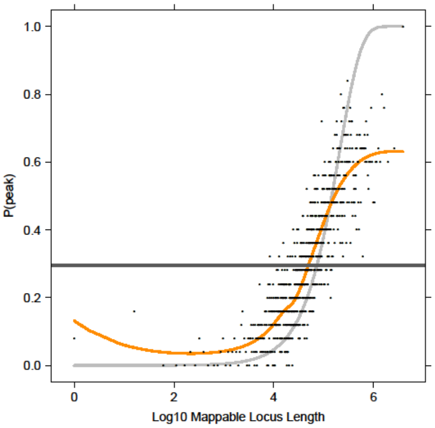
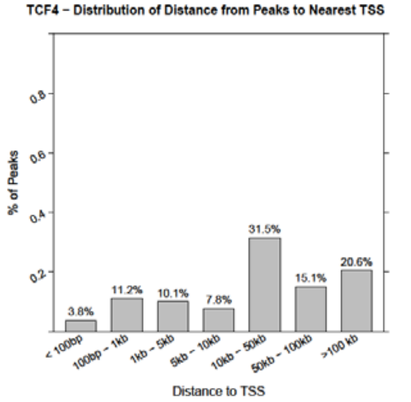
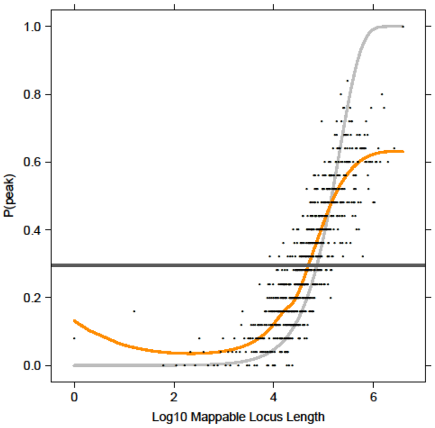
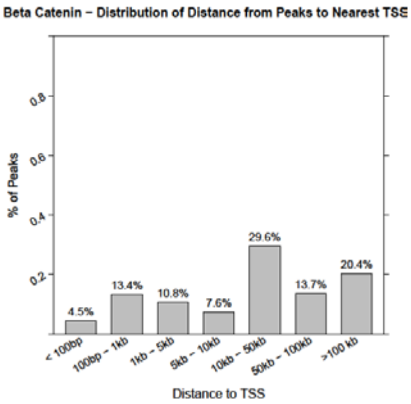
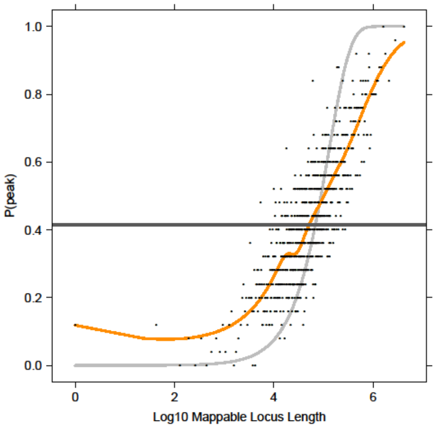
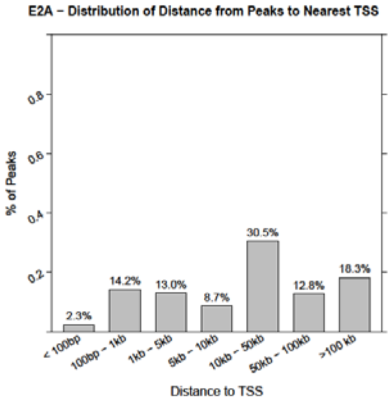
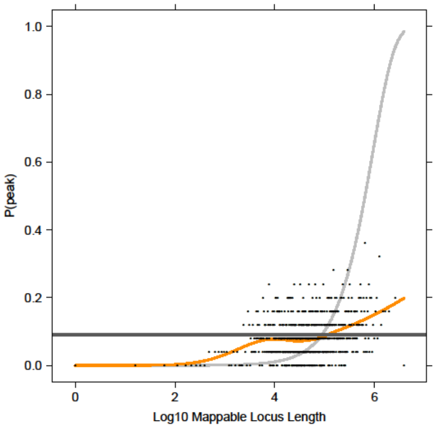
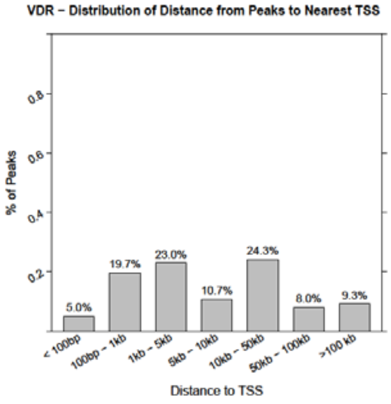
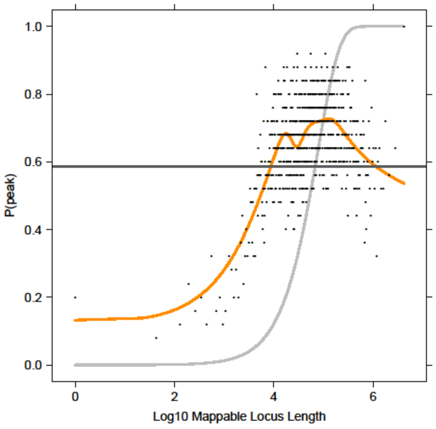
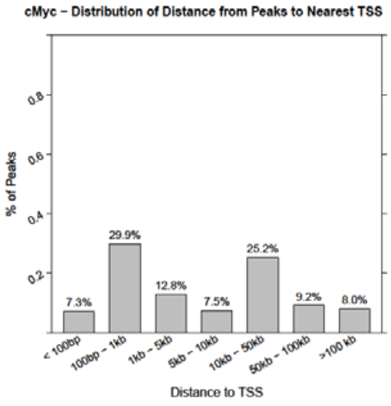


Figure 5.20. Comparison of the number of knots, choice of model (logistic regression, binomial smoothing spline), and $\log_{10}$ transformation of locus length on p-values for the top 15 GO terms for each dataset. We plot the $-\log_{10}$ p-values at knot values of 2, 5, 10, and 25 for the top 15 GO terms from each peak dataset, fit using the smoothing spline. ChIP-Enrich uses the model with 10 knots (green diamond.) All models are fit on $\log_{10}$ mappable locus length, with the exception of one model with 10 knots (red triangle.) Also plotted are three logistic regression models, which include progressively higher order $\log_{10}$ mappable locus length terms.

TF	Spline Fit	Distribution of Distance from Peak to Nearest TSS
GR 15,838 6,015 (31.7%)		
TCF4 10,413 5,782 (30.5%)		
Beta Catenin 10,324 5,445 (28.7%)		

TF	Spline Fit	Distribution of Distance from Peak to Nearest TSS
E2A 18,579 8,125 (42.8%)		
VDR 2,341 1,584 (8.4%)		
cMyc 30,902 12,535 (66.1%)		

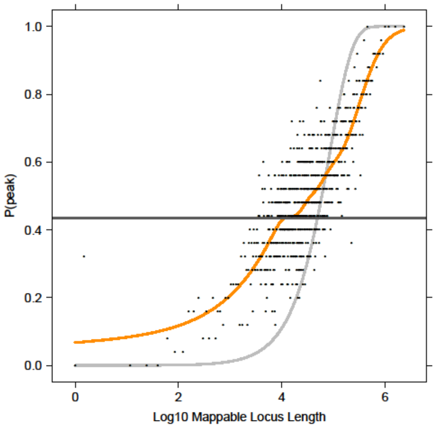
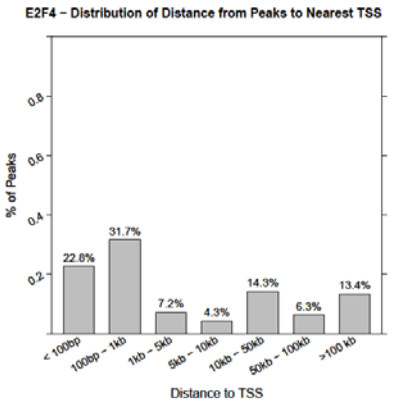
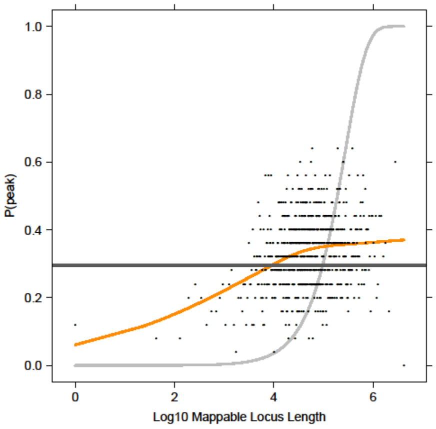
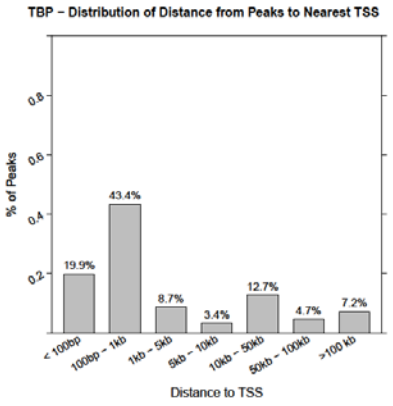
TF	Spline Fit	Distribution of Distance from Peak to Nearest TSS
E2F4 16,246 9,435 (49.7%)	 	
TBP 8,672 6,311 (33.3%)	 	

Figure 5.21. All 8 experimental ChIP-seq peak datasets. Each dataset is listed along with its spline fit plot and a histogram of the distance from each peak to the nearest TSS. The first column lists the transcription factor, the total number of peaks in the dataset, the number of genes to which a peak was assigned, the percentage of all genes assigned a peak, and the publication which introduced the dataset, in order. The second column shows the spline fit plot for each dataset. The third column shows the distribution of distance from each peak to the nearest TSS. The E2F4 dataset was introduced in (Lee et al. 2011), the GR dataset in (Reddy et al. 2009), the VDR dataset in (Heikkinen et al. 2011), and the remaining datasets in (Mokry et al. 2012).

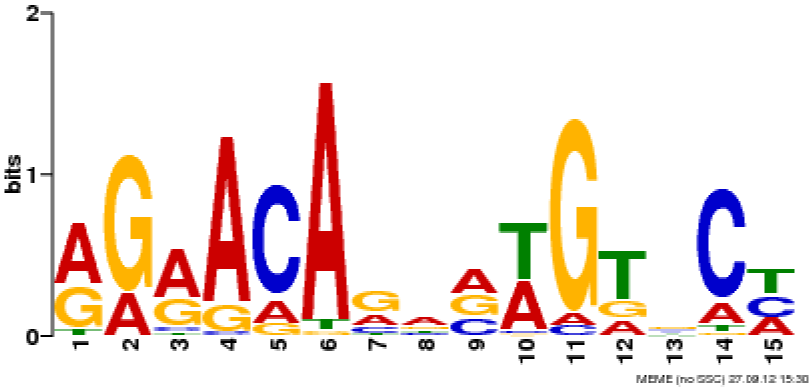


Figure 5.22. Most prevalent motif identified by MEME using the 460 peaks from the GR dataset assigned to genes in angiogenesis. The consensus sequence identified is 5'-AGAACAnnnTGTnCT-3', which best matches the known GRE motif consensus sequence.

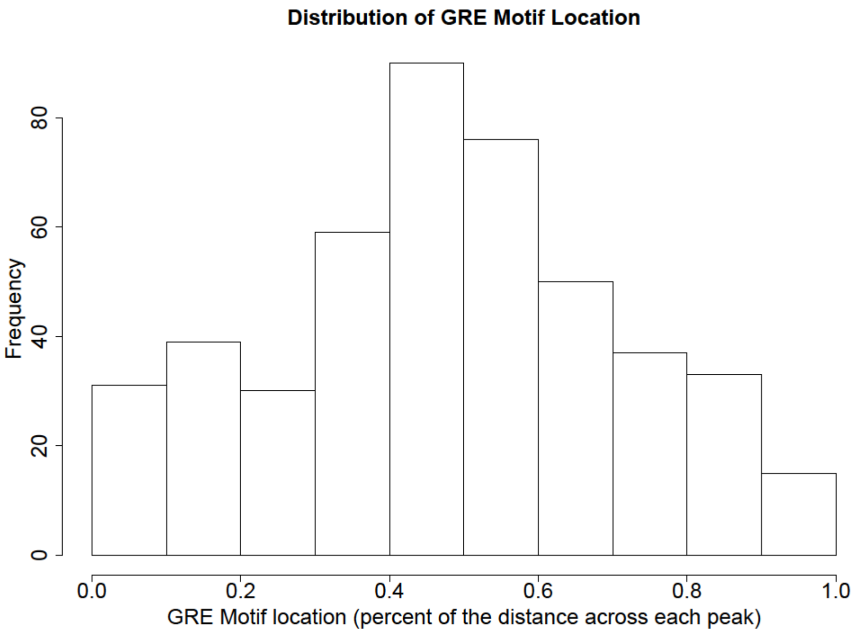


Figure 5.23. MEME analysis shows GRE elements tend to occur near the middle of GR peaks assigned to angiogenesis genes. The histogram represents the distribution of the GRE motif position within each peak, where the position is defined as the fraction of the total peak length at which the motif begins.

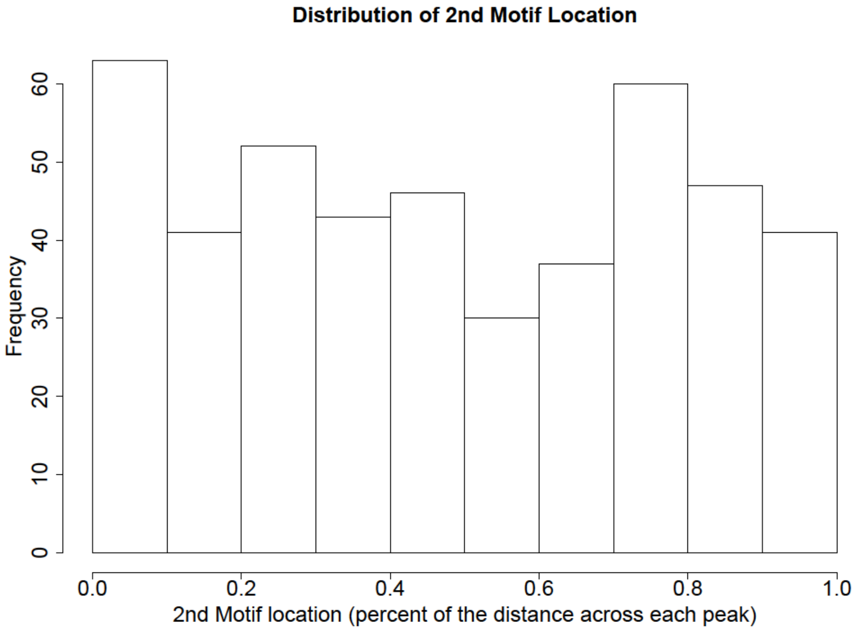
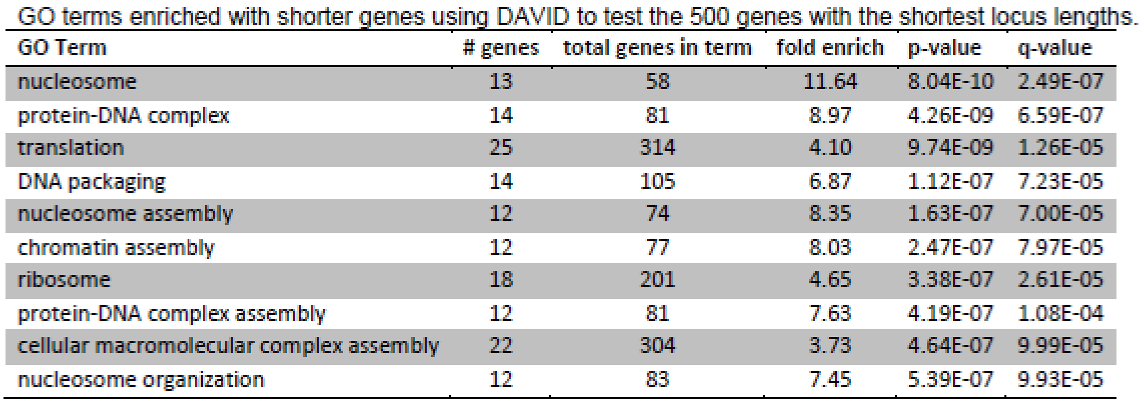


Figure 5.24. The 2nd (after GRE) identified motif from GR peaks assigned to angiogenesis is distributed randomly across the peaks. The histogram represents the distribution of the position of the 2nd motif within each peak, where the position is defined as the fraction of the total peak length at which the motif begins.

Publication	Protein	Total number of peaks called	Peak Caller	Sequencer	Aligner	Read length (bp)
Mokry et al. 2012	TCF4	10,413	CisGenome	SOLID	SHRiMP	36
Mokry et al. 2012	TBP	8,672	CisGenome	SOLID	SHRiMP	50
Mokry et al. 2012	Beta Catenin	10,324	CisGenome	SOLID	Maq	50
Mokry et al. 2012	E2A	18,579	CisGenome	SOLID	Maq	50
Mokry et al. 2012	cMyc	30,902	CisGenome	SOLID	Maq	50
Reddy et al. 2009	GR	15,838	MACS	Illumina GA1	ELAND	25
Heikkinen et al. 2011	VDR	2,341	MACS	Illumina GAII	Bowtie	36
Lee et al. 2011	E2F4	16,246	Parzen Window Algorithm Shivaswamy et al. 2008	Illumina GA1	ELAND	23-32

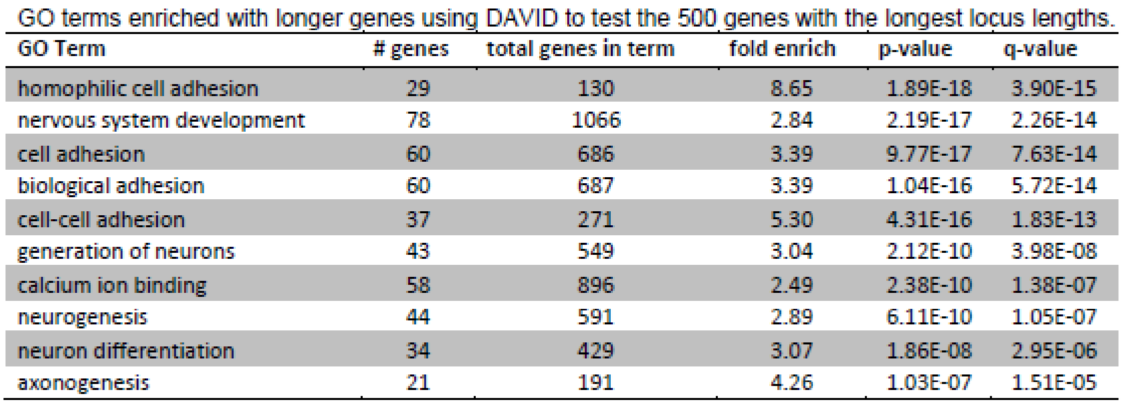
Table 5.4. Experimental ChIP-seq peak datasets selected for analysis. Each row corresponds to a separate ChIP-seq experiment as originally performed by the authors of the publication listed in the first column. Peaks were called by the authors and made publicly available for download.



GO terms enriched with shorter genes using DAVID to test the 500 genes with the shortest locus lengths.

GO Term	# genes	total genes in term	fold enrich	p-value	q-value
nucleosome	13	58	11.64	8.04E-10	2.49E-07
protein-DNA complex	14	81	8.97	4.26E-09	6.59E-07
translation	25	314	4.10	9.74E-09	1.26E-05
DNA packaging	14	105	6.87	1.12E-07	7.23E-05
nucleosome assembly	12	74	8.35	1.63E-07	7.00E-05
chromatin assembly	12	77	8.03	2.47E-07	7.97E-05
ribosome	18	201	4.65	3.38E-07	2.61E-05
protein-DNA complex assembly	12	81	7.63	4.19E-07	1.08E-04
cellular macromolecular complex assembly	22	304	3.73	4.64E-07	9.99E-05
nucleosome organization	12	83	7.45	5.39E-07	9.93E-05

Table 5.5. GO terms enriched with shorter genes using DAVID. The shortest 500 genes by locus length are tested with DAVID for enrichment in GO categories.



GO terms enriched with longer genes using DAVID to test the 500 genes with the longest locus lengths.

GO Term	# genes	total genes in term	fold enrich	p-value	q-value
homophilic cell adhesion	29	130	8.65	1.89E-18	3.90E-15
nervous system development	78	1066	2.84	2.19E-17	2.26E-14
cell adhesion	60	686	3.39	9.77E-17	7.63E-14
biological adhesion	60	687	3.39	1.04E-16	5.72E-14
cell-cell adhesion	37	271	5.30	4.31E-16	1.83E-13
generation of neurons	43	549	3.04	2.12E-10	3.98E-08
calcium ion binding	58	896	2.49	2.38E-10	1.38E-07
neurogenesis	44	591	2.89	6.11E-10	1.05E-07
neuron differentiation	34	429	3.07	1.86E-08	2.95E-06
axonogenesis	21	191	4.26	1.03E-07	1.51E-05

Table 5.6. GO terms enriched with longer genes using DAVID. The longest 500 genes by locus length are tested with DAVID for enrichment in GO categories.

GO ID	GO Category	Median Length of Genes	Number of Genes
GO:0002690	positive regulation of leukocyte chemotaxis	65665	30
GO:0008610	lipid biosynthetic process	67280	414
GO:0010463	mesenchymal cell proliferation	263619	32
GO:0000786	nucleosome	16390	61
GO:0005623	cell	66970	14283

Table 5.7. Gene Ontology terms selected for the simulation of type 1 error. Terms are shown graphically in Figure 5.11.

		Fisher's Test		Binomial Test		
Dataset	Gene Ontology Branch	Odds Ratio	P-value	P_{GO}	$P_{non\text{-}GO}$	P-value
GR	GO-BP	1.22	8.93E-09	0.76	0.72	2.84E-36
	GO-MF	1.24	3.36E-09	0.79	0.74	2.78E-58
	GO-CC	1.13	0.0021	0.83	0.80	2.87E-21
TCF4	GO-BP	1.34	1.76E-16	0.77	0.72	2.55E-35
	GO-MF	1.42	1.46E-20	0.80	0.74	8.43E-57
	GO-CC	1.36	2.10E-13	0.85	0.80	1.46E-36
Beta Catenin	GO-BP	1.32	1.35E-14	0.77	0.72	3.72E-28
	GO-MF	1.40	9.24E-19	0.80	0.74	2.87E-42
	GO-CC	1.33	8.32E-12	0.85	0.80	1.18E-36
E2A	GO-BP	1.28	2.81E-14	0.76	0.72	5.92E-44
	GO-MF	1.34	4.89E-18	0.79	0.74	4.83E-66
	GO-CC	1.36	6.52E-16	0.85	0.80	3.78E-65
VDR	GO-BP	1.38	1.45E-07	0.78	0.72	4.47E-12
	GO-MF	1.52	1.02E-10	0.82	0.74	7.17E-20
	GO-CC	1.35	2.82E-05	0.85	0.80	7.58E-09
CMYC	GO-BP	1.65	1.68E-50	0.77	0.72	1.12E-102
	GO-MF	1.79	4.40E-63	0.80	0.74	1.08E-161
	GO-CC	1.62	6.07E-37	0.85	0.80	9.12E-102
E2F4	GO-BP	1.48	3.99E-34	0.77	0.72	2.29E-57
	GO-MF	1.50	3.37E-33	0.80	0.74	7.20E-65
	GO-CC	1.41	1.05E-20	0.84	0.80	2.28E-43
TBP	GO-BP	1.55	3.06E-35	0.77	0.72	3.64E-24
	GO-MF	1.60	1.19E-37	0.79	0.74	2.34E-25
	GO-CC	1.47	3.82E-21	0.84	0.80	2.96E-16

Table 5.8. Enrichment of peaks within genes annotated in GO. For each dataset and GO branch, we used Fisher’s exact test and a binomial test to look for enrichment of peaks. We list the odds ratios and two-sided p-values for Fisher’s exact test. For the binomial test, we list the proportion of all peaks in the dataset within genes in the GO branch (P_{GO}), the proportion of all peaks occurring in the non-GO annotated genome ($P_{non\text{-}GO}$), and the one-sided binomial test p-value.

Peak Midpoint	Distance from Peak to TSS	Peak Assigned to Gene ID	Gene Symbol	Position of Motif in Peak	MEME Motif P-value
chr12:56241893	5158	4327	MMP19	238	1.11E-08
chr6:12370319	79791	1906	EDN1	386	2.13E-08
chr18:7955360	388047	5797	PTPRM	200	7.24E-08
chr2:20683352	36518	388	RHOB	200	1.03E-07
chr15:39720102	153177	7057	THBS1	144	1.03E-07
chr17:15866805	18575	136	ADORA2B	192	1.46E-07
chr15:90355621	2451	290	ANPEP	276	2.08E-07
chr5:153918051	60227	9421	HAND1	212	2.08E-07
chr12:92424587	115086	694	BTG1	256	2.08E-07
chr19:8439370	10360	51129	ANGPTL4	369	2.81E-07
chr4:177553014	160881	7424	VEGFC	123	2.81E-07
chr14:54593886	170332	652	BMP4	103	2.81E-07
chr8:27198448	15368	2185	PTK2B	328	3.77E-07
chr5:170866525	19859	8817	FGF18	177	3.77E-07
chr3:30671443	23450	7048	TGFBR2	182	3.77E-07
chr6:91238761	58003	6885	MAP3K7	7	3.77E-07
chr18:7414674	152639	5797	PTPRM	188	3.77E-07
chr3:30304510	343483	7048	TGFBR2	184	3.77E-07
chr9:33163157	4199	2683	B4GALT1	539	5.07E-07
chr18:7585790	18477	5797	PTPRM	117	5.07E-07
chr19:8433846	4836	51129	ANGPTL4	75	6.58E-07
chr9:139519756	37622	51162	EGFL7	46	8.42E-07
chr18:19859975	110560	2627	GATA6	119	8.42E-07
chr1:59454597	204812	3725	JUN	146	8.42E-07
chr18:19794875	45460	2627	GATA6	82	1.08E-06
chr17:37260244	47658	57125	PLXDC1	193	1.08E-06
chr12:92464269	75404	694	BTG1	648	1.08E-06
chr18:19847789	98374	2627	GATA6	62	1.08E-06
chr20:10975670	320976	182	JAG1	158	1.08E-06
chr18:7921116	353803	5797	PTPRM	97	1.08E-06
chr13:74002510	369369	688	KLF5	217	1.08E-06
chr6:138194636	6056	7128	TNFAIP3	215	1.35E-06
chr8:108626929	116675	284	ANGPT1	109	1.35E-06
chr13:73763675	130534	688	KLF5	213	1.35E-06
chr1:59395664	145879	3725	JUN	456	1.35E-06
chr20:10879354	224660	182	JAG1	156	1.35E-06
chr18:7871075	303762	5797	PTPRM	331	1.35E-06
chr17:64256584	31028	350	APOH	166	1.69E-06
chr17:15781667	66563	136	ADORA2B	131	1.69E-06
chr14:54273966	147304	652	BMP4	279	1.69E-06
chr17:64226839	1283	350	APOH	143	2.09E-06

chr2:96788196	6308	151	ADRA2B	390	2.09E-06
chr2:96764792	17096	151	ADRA2B	129	2.09E-06
chr20:872141	24819	51378	ANGPT4	179	2.09E-06
chr7:22713106	53659	3569	IL6	98	2.09E-06
chr3:78715827	352782	6091	ROBO1	214	2.09E-06
chr20:11177107	522413	182	JAG1	132	2.09E-06
chr15:89459311	2648	4240	MFGE8	79	2.56E-06
chr1:95018039	10626	2152	F3	237	2.56E-06
chr15:60678552	11633	302	ANXA2	172	2.56E-06
chr6:32212294	20450	4855	NOTCH4	170	2.56E-06
chr2:216325593	24802	2335	FN1	301	2.56E-06
chr8:55295732	74762	64321	SOX17	199	2.56E-06
chr3:30730775	82782	7048	TGFBR2	475	2.56E-06
chr18:7877805	310492	5797	PTPRM	200	2.56E-06
chr18:7941732	374419	5797	PTPRM	215	2.56E-06
chr5:76353615	27406	55109	AGGF1	179	3.12E-06
chr9:38040345	28865	6461	SHB	420	3.12E-06
chr1:110835273	46671	64783	RBM15	275	3.12E-06
chr6:132348915	76397	1490	CTGF	181	3.12E-06
chr18:19855589	106174	2627	GATA6	111	3.12E-06
chr9:33157034	10322	2683	B4GALT1	48	3.79E-06
chr14:75442400	19933	5228	PGF	140	3.79E-06
chr2:20725227	78393	388	RHOB	349	3.79E-06
chr2:20739618	92784	388	RHOB	273	3.79E-06
chr18:19624722	124693	2627	GATA6	218	3.79E-06
chr13:73823071	189930	688	KLF5	217	3.79E-06
chr7:100761707	8671	5054	SERPINE1	275	4.55E-06
chr3:30638915	9078	7048	TGFBR2	106	4.55E-06
chr17:1655957	9301	5176	SERPINF1	213	4.55E-06
chr2:216282359	18432	2335	FN1	360	4.55E-06
chr17:64225519	37	350	APOH	87	5.43E-06
chr1:27868925	52247	10163	WASF2	236	5.43E-06
chr8:27256345	73265	2185	PTK2B	287	5.43E-06
chr12:92637977	98304	694	BTG1	207	5.43E-06
chr4:86850602	823	83478	ARHGAP24	137	6.47E-06
chr15:60667018	23167	302	ANXA2	257	6.47E-06
chr18:57183770	180874	147372	CCBE1	171	6.47E-06
chr2:224459150	8067	7857	SCG2	149	7.67E-06
chr6:43781108	43163	7422	VEGFA	65	7.67E-06
chr17:59588981	55175	9496	TBX4	102	7.67E-06
chr6:68943559	402072	577	BAI3	75	7.67E-06
chr3:129331709	6127	23129	PLXND1	398	9.02E-06
chr18:7576805	9492	5797	PTPRM	169	9.02E-06
chr7:100755644	14734	5054	SERPINE1	351	9.02E-06
chr19:43002045	30616	634	CEACAM1	256	9.02E-06

chr10:94491276	41596	3087	HHEX	192	9.02E-06
chr21:36184017	76970	861	RUNX1	261	9.02E-06
chr15:33142813	132609	26585	GREM1	197	9.02E-06
chr1:59398732	148947	3725	JUN	488	9.02E-06
chr12:48213916	153	51564	HDAC7	1440	1.06E-05
chr6:12324605	34077	1906	EDN1	234	1.06E-05
chr4:56050664	58902	3791	KDR	114	1.06E-05
chr2:224353736	113481	7857	SCG2	95	1.06E-05
chr18:19600751	148664	2627	GATA6	145	1.06E-05
chr4:75410066	179207	2069	EREG	144	1.06E-05
chr19:6720814	152	718	C3	341	1.24E-05
chr1:37945567	5449	80149	ZC3H12A	747	1.24E-05
chr5:41194094	19573	729	C6	151	1.24E-05
chr19:42968211	21075	284340	CXCL17	232	1.24E-05
chr6:12314524	23996	1906	EDN1	720	1.24E-05
chr1:27855255	38577	10163	WASF2	329	1.24E-05
chr1:172715585	87401	356	FASLG	133	1.24E-05
chr3:30443639	204354	7048	TGFBR2	143	1.24E-05
chr13:74025639	392498	688	KLF5	161	1.24E-05
chr8:27172984	3986	2185	PTK2B	266	1.44E-05
chr20:30183872	9213	3397	ID1	671	1.44E-05
chr17:46626828	24982	3213	HOXB3	151	1.44E-05
chr15:60662859	27326	302	ANXA2	460	1.44E-05
chr1:59290886	41101	3725	JUN	92	1.44E-05
chr1:186547613	101946	5743	PTGS2	58	1.44E-05
chr4:75410899	180040	2069	EREG	79	1.44E-05
chr1:218794548	275873	7042	TGFB2	126	1.44E-05
chr10:34038130	414297	8829	NRP1	35	1.44E-05
chr12:54815103	2053	3678	ITGA5	395	1.68E-05
chr2:158723995	7628	90	ACVR1	119	1.68E-05
chr17:37261130	46772	57125	PLXDC1	79	1.68E-05
chr10:33505710	118123	8829	NRP1	135	1.68E-05
chr1:218819879	301204	7042	TGFB2	152	1.68E-05
chr7:100769965	413	5054	SERPINE1	84	1.94E-05
chr12:96590735	2529	2004	ELK3	191	1.94E-05
chr2:228045124	15844	1285	COL4A3	186	1.94E-05
chr1:59282097	32312	3725	JUN	240	1.94E-05
chr2:20692065	45231	388	RHOB	554	1.94E-05
chr3:30482535	165458	7048	TGFBR2	161	1.94E-05
chr2:129342314	266143	9394	HS6ST1	6	1.94E-05
chr8:108199658	310596	284	ANGPT1	164	1.94E-05
chr8:98670565	14159	92140	MTDH	139	2.24E-05
chr14:75406977	15490	5228	PGF	98	2.24E-05
chr4:123764360	16498	2247	FGF2	165	2.24E-05
chr1:22239980	23770	3339	HSPG2	155	2.24E-05

chr6:12348968	58440	1906	EDN1	182	2.24E-05
chr5:153918675	60851	9421	HAND1	118	2.24E-05
chr10:33483242	140591	8829	NRP1	318	2.24E-05
chr3:41013132	227809	1499	CTNNB1	656	2.24E-05
chr18:7876588	309275	5797	PTPRM	209	2.24E-05
chr20:11210622	555928	182	JAG1	450	2.24E-05
chr20:60930977	11391	3911	LAMA5	527	2.59E-05
chr6:43709688	28257	7422	VEGFA	175	2.59E-05
chr1:27848954	32276	10163	WASF2	120	2.59E-05
chr4:74570479	35743	3576	IL8	257	2.59E-05
chr1:172795577	167393	356	FASLG	196	2.59E-05
chr13:110772934	186562	1282	COL4A1	301	2.59E-05
chr3:78818200	250409	6091	ROBO1	176	2.59E-05
chr3:30384017	263976	7048	TGFBR2	40	2.59E-05
chr15:60685411	4774	302	ANXA2	36	2.97E-05
chr14:75430342	7875	5228	PGF	120	2.97E-05
chr1:95018997	11584	2152	F3	214	2.97E-05
chr4:177699883	14012	7424	VEGFC	222	2.97E-05
chr20:60925557	16811	3911	LAMA5	857	2.97E-05
chr1:86073053	26610	3491	CYR61	334	2.97E-05
chr2:20674986	28152	388	RHOB	178	2.97E-05
chr11:111999669	35171	3606	IL18	145	2.97E-05
chr1:86095002	48559	3491	CYR61	615	2.97E-05
chr12:1671842	54379	81029	WNT5B	170	2.97E-05
chr8:27271616	88536	2185	PTK2B	145	2.97E-05
chr18:7754948	187635	5797	PTPRM	325	2.97E-05
chr18:7797705	230392	5797	PTPRM	207	2.97E-05
chr18:7946872	379559	5797	PTPRM	18	2.97E-05
chr20:60944826	2458	3911	LAMA5	2184	3.41E-05
chr20:30181464	11621	3397	ID1	797	3.41E-05
chr21:36209035	51952	861	RUNX1	260	3.41E-05
chr14:62105882	56236	3091	HIF1A	126	3.41E-05
chr21:36338006	77019	861	RUNX1	49	3.41E-05
chr5:153972859	115035	9421	HAND1	187	3.41E-05
chr15:33132948	122744	26585	GREM1	254	3.41E-05
chr6:138056122	132458	7128	TNFAIP3	116	3.41E-05
chr10:89870305	247111	5728	PTEN	120	3.41E-05
chr2:236742752	333900	2637	GBX2	91	3.41E-05
chr20:39772144	5984	5335	PLCG1	210	3.90E-05
chr6:12298077	7549	1906	EDN1	200	3.90E-05
chr17:15867260	19030	136	ADORA2B	327	3.90E-05
chr2:129136780	60609	9394	HS6ST1	216	3.90E-05
chr8:55308363	62131	64321	SOX17	147	3.90E-05
chr18:7914772	347459	5797	PTPRM	159	3.90E-05
chr19:8420690	8320	51129	ANGPTL4	476	4.45E-05

chr8:108521339	11085	284	ANGPT1	146	4.45E-05
chr8:108537896	27642	284	ANGPT1	142	4.45E-05
chr22:30603141	39655	3976	LIF	268	4.45E-05
chr12:92492953	46720	694	BTG1	48	4.45E-05
chr2:46458015	66525	2034	EPAS1	266	4.45E-05
chr5:153963550	105726	9421	HAND1	383	4.45E-05
chr6:132384575	112057	1490	CTGF	203	4.45E-05
chr15:39724405	148874	7057	THBS1	34	4.45E-05
chr2:216587203	286412	2335	FN1	259	4.45E-05
chr19:8427011	1999	51129	ANGPTL4	241	5.07E-05
chr17:32579670	2625	6347	CCL2	359	5.07E-05
chr20:30429894	3526	2307	FOXS1	129	5.07E-05
chr1:155103904	3556	1942	EFNA1	191	5.07E-05
chr13:73628969	4172	688	KLF5	197	5.07E-05
chr17:45343703	12496	3690	ITGB3	142	5.07E-05
chr9:139416078	24160	4851	NOTCH1	301	5.07E-05
chr9:139414431	25807	4851	NOTCH1	484	5.07E-05
chr20:10686828	32134	182	JAG1	94	5.07E-05
chr6:138132719	55861	7128	TNFAIP3	145	5.07E-05
chr2:46460183	64357	2034	EPAS1	114	5.07E-05
chr13:73707859	74718	688	KLF5	94	5.07E-05
chr3:40981030	259911	1499	CTNNB1	320	5.07E-05
chr6:68891064	454567	577	BAI3	182	5.07E-05
chr20:30196350	3265	3397	ID1	192	5.75E-05
chr17:1669368	4110	5176	SERPINF1	115	5.75E-05
chr19:47818776	5673	728	C5AR1	188	5.75E-05
chr7:155653833	48866	6469	SHH	629	5.75E-05
chr2:228132745	103465	1285	COL4A3	254	5.75E-05
chr18:7756002	188689	5797	PTPRM	309	5.75E-05
chr3:79061124	7485	6091	ROBO1	116	6.52E-05
chr2:20638446	8388	388	RHOB	125	6.52E-05
chrX:100120168	9166	27035	NOX1	63	6.52E-05
chr1:22254198	9552	3339	HSPG2	20	6.52E-05
chr14:62109181	52937	3091	HIF1A	159	6.52E-05
chr6:132341552	69034	1490	CTGF	262	6.52E-05
chr9:27183921	74775	7010	TEK	45	6.52E-05
chr20:10942648	287954	182	JAG1	176	6.52E-05
chr3:30327822	320171	7048	TGFBR2	59	6.52E-05
chr19:6723404	2742	718	C3	58	7.38E-05
chr6:12296591	6063	1906	EDN1	209	7.38E-05
chr22:37484404	15289	164656	TMPRSS6	204	7.38E-05
chr22:30608608	34188	3976	LIF	226	7.38E-05
chr22:30601642	41154	3976	LIF	222	7.38E-05
chr1:22221283	42467	3339	HSPG2	138	7.38E-05
chr2:216363323	62532	2335	FN1	129	7.38E-05

chr1:59388131	138346	3725	JUN	161	7.38E-05
chr6:132428195	155677	1490	CTGF	113	7.38E-05
chr2:216483914	183123	2335	FN1	137	7.38E-05
chr13:73899733	266592	688	KLF5	109	7.38E-05
chr8:108208310	301944	284	ANGPT1	367	7.38E-05
chr1:218829932	311257	7042	TGFB2	590	7.38E-05
chr4:110842633	8594	1950	EGF	78	8.33E-05
chr1:37950470	10352	80149	ZC3H12A	157	8.33E-05
chr1:37927409	12709	80149	ZC3H12A	356	8.33E-05
chr22:35760470	16589	3162	HMOX1	32	8.33E-05
chr2:46542618	18078	2034	EPAS1	210	8.33E-05
chr15:60672045	18140	302	ANXA2	551	8.33E-05
chr19:42993225	39436	634	CEACAM1	22	8.33E-05
chr2:20696246	49412	388	RHOB	130	8.33E-05
chr15:60632711	57474	302	ANXA2	260	8.33E-05
chr1:218670114	151439	7042	TGFB2	63	8.33E-05
chr15:90358398	326	290	ANPEP	479	9.39E-05
chr1:22257721	6029	3339	HSPG2	536	9.39E-05
chr10:94457563	7883	3087	HHEX	89	9.39E-05
chr12:92457604	82069	694	BTG1	283	9.39E-05
chr2:224564419	97202	7857	SCG2	107	9.39E-05
chr18:7761570	194257	5797	PTPRM	307	9.39E-05
chr13:73833335	200194	688	KLF5	319	9.39E-05
chr2:46265541	258999	2034	EPAS1	147	9.39E-05
chr6:68942908	402723	577	BAI3	68	9.39E-05
chr11:20390141	4455	10553	HTATIP2	86	0.000106
chr22:35768101	8958	3162	HMOX1	28	0.000106
chr15:75985967	19222	1464	CSPG4	140	0.000106
chr6:43770479	32534	7422	VEGFA	140	0.000106
chr22:30606527	36269	3976	LIF	209	0.000106
chr9:101752048	45911	1306	COL15A1	205	0.000106
chr1:27867021	50343	10163	WASF2	386	0.000106
chr14:103655399	62736	7127	TNFAIP2	454	0.000106
chr10:33546966	76867	8829	NRP1	188	0.000106
chr15:33117136	106932	26585	GREM1	277	0.000106
chr12:1609591	116630	81029	WNT5B	141	0.000106
chr4:26147117	174214	3516	RBPJ	65	0.000106
chr18:19571262	178153	2627	GATA6	119	0.000106
chr8:108163843	346411	284	ANGPT1	306	0.000106
chr18:7921465	354152	5797	PTPRM	46	0.000106
chr18:7923290	355977	5797	PTPRM	166	0.000106

Table 5.9. Peaks assigned to genes in the angiogenesis GO term with a significant match to the GRE motif. Each peak is listed as the midpoint of the peak in chromosome:position format. The peak is assigned to the gene with the closest TSS. MEME gives the starting position of the highest ranked motif (GRE) within each peak, and a p-value for how well the peak sequence matches the motif sequence. Peaks are sorted according to the p-value for motif/sequence matching. Only those peaks with MEME p-value < 0.05 / 460 (Bonferroni adjustment) are presented.

Gene	GRE Motif Site	MEME P-value	Distance to TSS (bp)
Vegf-c	AGGACAAAATGTGCT	2.81E-07	31,087
Vegf-a	GGGACAGTGTGTACA	7.67E-06	28,257
Vegf-a	AGAACACTATGGACA	2.59E-05	32,534
Vegf-c	AGGACAGAAAGAAAT	2.97E-05	43,163
Vegf-a	GAGGCAGCATGTTCC	0.000106	160,881
Vegf-c	AGAGCAGACACTCCC	0.000118	157,073
Vegf-a	GGGCCAGACAGAACA	0.000497	14,012
Vegf-c	GGGCCAGCCACTGCA	0.00173	91,581

Table 5**.10. GRE motifs identified by MEME inside peaks assigned to VEGF-A and VEGF-C, known regulators of angiogenesis activity.** Five of the eight peaks closely matched the GRE consensus sequence (5'-AGAACAnnnTGTTCT-3') with a Bonferroni-adjusted p-value < 0.05 (rows highlighted in red.)

Chapter 6 Conclusion

Over the last few years, there has been an incredible expansion in the ability to interrogate the human genome. Genome-wide association studies have provided a comprehensive view of common SNPs and their association to hundreds of diseases and traits (Hindorff et al. 2009). With the expansion of efficient whole genome sequencing, there is now an even greater ability to assay both common and rare forms of variation, and in addition, the ability to investigate other interesting questions in genomics, such as the identification of genome-wide protein binding sites with ChIP-seq studies. These new technologies brought with them a number of interesting challenges.

Through GWAS, we and others have discovered that many associated variants are located outside the coding region of genes (Hindorff et al. 2009), making interpretation difficult as to how they might function in a genomic context to increase susceptibility to disease, or influence a trait. Researching the genes within the regions near associated variants is an essential step in not only identifying those genes that may be functionally related to the disease/trait, but also in furthering our understanding of the potential mechanisms through which the genes may operate. Given the large number of associated variants known, and the even greater number of genes within their genomic regions, it is a challenging task to systematically evaluate and research each of them.

Another challenge in GWAS is visualizing associated loci from GWAS and meta-analysis. An effective visualization combines many features of the human genome, including linkage disequilibrium, recombination rates, and functional annotation, with GWAS or meta-analysis results into a format that can be quickly inspected for patterns or insights. Through LocusZoom we make data exploration available to a wide audience and promote creation of plots that are consistent across studies and easily interpretable to researchers. These visualizations also allow for the detection of interesting patterns from multiple data sources. As an example, an associated variant could be in LD with other variants that have strong functional annotations and/or are located within multiple candidate genes. We can present this visually, reducing the need for a researcher to write queries or determine thresholds for LD. Actionable insights can also be gained from these visualizations. For example, often there are only a handful of genes are located within the recombination hotspots containing an associated variant, and an effective targeted sequencing strategy might be to focus attention on following up these genes first in follow-up efforts. Another scenario would be in identifying gaps of low SNP coverage near an association signal, potentially signaling the presence of a structural variant in the region. Many other patterns and insights likely exist in GWAS data, and future efforts at creating visualizations should prove to be interesting in ongoing studies.

In addition to these challenges, individual GWAS often are underpowered to detect common associated SNPs with low effect sizes. Combining GWAS through meta-analysis to increase power is an effective strategy to identify such SNPs (Zeggini et al. 2008; Teslovich et al. 2010; Voight et al. 2010). Another approach to identifying additional associated SNPs is to increase the number of SNPs genotyped, either at the discovery stage, or the replication stage. With previous genotyping platforms however, it was not possible to genotype many SNPs in replication samples, limiting identification of genome-wide significant SNPs to at most 10-30 in total (De la Vega et al. 2005; Gabriel et al. 2009). Development of specialized genotyping arrays, such as the Metabochip and ImmunoChip (Voight et al. 2012), provided the ability to follow-up thousands of SNPs from previous meta-analysis studies.

In chapter 2, I introduced software aimed at the challenge in GWAS of researching and understanding SNPs and their surrounding genomic regions in the context of a disease or trait. Snipper provides researchers the ability to quickly and exhaustively examine the genes at hundreds of associated loci. Since Snipper was designed to pull data directly from public databases, users will always receive up-to-date information for each gene. I provide a simple search method to collect the sizeable amount of information on each gene into a listing of where each search term matches, which allows a user to pinpoint the most interesting details regarding genes near their loci. I provide both a command-line and graphical interface to Snipper, which allows a broader range of researchers to use the program. Snipper is open source and freely available, and could easily be adapted to pull data from additional sources, such as tracks from the UCSC genome browser, Ensembl, tissue expression databases, or ENCODE (Kent et al. 2002; Fujita et al. 2011; Dunham et al. 2012; Flicek et al. 2012). Future work could also include developing a web interface to Snipper, where users could queue their lists of SNPs, and have their results emailed to them when the job completes.

In chapter 3, I introduced LocusZoom, a software tool for visualization of genome-wide association scan results. LocusZoom collapses numerous features of the genome into a single figure that can aid in the interpretation of associated regions. These features include the association signal (p-values), linkage disequilibrium and recombination rates, the position of SNPs and their location relative to genes, and functional annotation. By providing both a web based interface, and command-line standalone software, users at all levels of expertise are able to use LocusZoom, as well as extend the software for their own usage. Researchers have since adapted LocusZoom (and its related databases) for their own purposes, for example: with organisms other than human such as mouse (personal communication, Dan Gatti, 2011), to plot associated variants from the GWAS catalog (http://www.genome.gov/gwastudies/) (personal communication, Adam Locke, 2012), and to calculate linkage disequilibrium from populations beyond those provided in 1000 Genomes and HapMap (personal communication, Serena Sanna, 2010). On the web portal for LocusZoom we hope to continue adding additional meta-analyses and GWAS for new traits and diseases as they become available. This provides the unique opportunity for researchers to investigate their SNPs in other studies that are related to their trait for possible pleiotropy. For future work, I am currently modifying the software to plot fine-mapping scan results, as this has become a popular analysis since the Metabochip and other specialized chips were developed. I also intend to continue improving on our latest feature to plot associated SNPs from the GWAS catalog and from WikiGWA, a database of SNP-trait associations.

In chapter 4, I presented my work as part of the MAGIC consortium, where I meta-analyzed both previous GWAS and studies genotyped with the Metabochip to identify 4 new loci associated with 2-hr glucose. In addition, our group efforts identified 20 new loci associated with fasting glucose, and 17 with fasting insulin. We have raised the total number of loci associated with glycemic traits to 53, which now account for 4.8%, 1.2%, and 1.7% of the variance in FG, FI, and 2hGlu, respectively. Given that these three traits, and the many genes and pathways behind them, are important in overall glucose homeostasis, we expect that some of the loci associated with them are also linked to T2D. We observed that 33 loci (out of 53) from the three traits were also associated with increased T2D risk ($q < 0.05$). It is possible that loci overlapping both glycemic traits and T2D risk represent the best targets for further functional follow-up in the aim of improving our understanding of T2D etiology.

For 2-hr glucose we identified four new loci in addition to confirming the five previously known using a total sample size of 42,854 individuals. Three of the new association signals are near known loci for FG, HDL-cholesterol, and T2D, suggesting a potentially pleiotropic effect for some of these loci with other metabolic traits. From the total of 9 SNPs associated with 2-hr glucose in our analyses, 8 were also associated with T2D. For three of these SNPs, the 2-hr glucose raising allele was associated with a decreased risk of T2D, and this association perhaps acts through a mechanism involving fasting glucose, as the 2-hr glucose raising allele for each showed an association with lower fasting glucose levels. Gene based analyses also identified an additional 3 regions having an association with 2-hr glucose, suggesting that perhaps with greater sample sizes, SNPs in these regions may reach genome-wide significance in single variant analyses.

Further study of the loci from each of these traits is required to understand the mechanism through which they operate. Of particular interest to the study of T2D are the loci associated with both a glycemic trait and T2D. Fine-mapping studies currently being performed by MAGIC and DIAGRAM may help to narrow down the variants within each associated region, possibly providing more plausible functional variants for follow-up. Our gene-based analyses showed that there likely also exist additional associations to be confirmed, although greater sample sizes will be required. This will become possible as more studies are genotyped. We can also expect that genotyping samples of additional ancestries will yield new associations not detectable in European samples (Pulit et al. 2010).

Genome sequencing for detecting rare genetic variants is a natural next step to genome-wide association studies. Sequencing has also proven useful in many aspects of genome biology, such as DNA methylation (MeDIP-seq), RNA expression (RNA-seq), open chromatin (DNAse-seq), and many others. In the final chapter of my dissertation, I focused on ChIP-seq studies, which ascertain the binding locations of a protein across the genome. While the biological functions of the protein in these studies are often not known, it is possible to infer them from the functions of the genes near regions where the protein binds. This inference can be performed by testing for an enrichment of ChIP-seq peaks within sets of biologically related genes. In performing this testing, it is necessary to consider the potentially confounding effect of gene length, intergenic distance and other factors such as sequence mappability. These factors can increase the likelihood of observing a peak within long or highly mappable genes, which can lead to false detection of long or highly mappable sets of genes as enriched.

ChIP-Enrich is a gene set enrichment test for ChIP-seq data that can account for the potentially confounding effect of gene locus length and mappability. I showed that the assumptions made by Fisher's exact test (each gene has the same probability of observing ≥1 peak) and the binomial based test (probability of observing ≥1 peak is proportional to gene locus length) lead both tests to be highly anti-conservative. ChIP-Enrich uses an empirical approach to model the relationship of the probability of observing ≥1 peak in a locus to the locus length (and mappability, and its type 1 error rate is close to the expected rate over a range of simulations and permutations of experimental data. I provided supporting evidence from the published literature that angiogenesis, a function previously un-annotated to glucocorticoid receptor (GR), represents a true enrichment result for GR. I identified the glucocorticoid response element (GRE) (a known motif for GR) *de novo* using MEME from those peaks assigned to angiogenesis genes, and also found that the large majority of angiogenesis genes had a peak containing the GRE. These two facts support the conclusion that GR regulates genes involved in angiogenesis. Since GR peaks occurred distally to angiogenesis genes, this finding would not have been identified by only using peaks proximal to the transcription start site (TSS). I also demonstrated that limiting the analysis to peaks near the TSS in the GR dataset would result in missing many other biologically relevant GO terms.

ChIP-Enrich represents an important step forward for gene set enrichment testing with ChIP-seq data. Still, there remain many interesting avenues for future research. ChIP-Enrich considers loci with one or more peaks equally in the analysis. It remains an open question as to whether considering peak count per locus could yield different biological insights. ChIP-Enrich could potentially address this question by changing the outcome variable from the presence or absence of ≥1 peak, to the presence or absence of ≥ 2 peaks, or to any chosen number of peaks. An interesting analysis would be to consider which gene sets show greater enrichment under different thresholds for the count of peaks, as this could hint at potentially novel biology (e.g. the peak binds singly to certain sets of genes, but multiply to others.) Using a Poisson regression model, as opposed to a logistic model, could offer an alternative approach to modeling counts of peaks. Another avenue of research is to consider the construction of more accurate gene locus definitions that better capture the true regulatory domain of the gene. Given the wide array of data currently being generated by ENCODE (Dunham et al. 2012), we expect an improved picture of gene regulatory domains to emerge, and this could improve our ability to detect true gene set enrichment and make accurate predictions about the function of DNA binding proteins. Finally, ChIP-Enrich is applicable to many similar types of data, such as MeDIP-seq, and applying the method to these datasets could result in interesting biological insights that have not yet been identified.

Bibliography

Altshuler DL Durbin RM Abecasis GR Bentley DR Chakravarti A Clark AG Collins FS De la Vega FM Donnelly P Egholm M et al. 2010. A map of human genome variation from population-scale sequencing. *Nature* **467**(7319): 1061-1073.

Anderson CA Boucher G Lees CW Franke A D'Amato M Taylor KD Lee JC Goyette P Imielinski M Latiano A et al. 2011. Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nature genetics* **43**(3): 246-252.

Anderson SL, Coli R, Daly IW, Kichula EA, Rork MJ, Volpi SA, Ekstein J, Rubin BY. 2001. Familial dysautonomia is caused by mutations of the IKAP gene. *American journal of human genetics* **68**(3): 753-758.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics* **25**(1): 25-29.

Aulchenko YS, Pullen J, Kloosterman WP, Yazdanpanah M, Hofman A, Vaessen N, Snijders PJ, Zubakov D, Mackay I, Olavesen M et al. 2007. LPIN2 is associated with type 2 diabetes, glucose metabolism, and body composition. *Diabetes* **56**(12): 3020-3026.

Ayalasomayajula SP, Ashton P, Kompella UB. 2009. Fluocinolone inhibits VEGF expression via glucocorticoid receptor in human retinal pigment epithelial (ARPE-19) cells and TNF-alpha-induced angiogenesis in chick chorioallantoic membrane (CAM). *Journal of ocular pharmacology and therapeutics : the official journal of the Association for Ocular Pharmacology and Therapeutics* **25**(2): 97-103.

Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. 2009. MEME SUITE: tools for motif discovery and searching. *Nucleic acids research* **37**(Web Server issue): W202-208.

Barish GD, Narkar VA, Evans RM. 2006. PPAR delta: a dagger in the heart of the metabolic syndrome. *The Journal of clinical investigation* **116**(3): 590-597.

Barker A, Sharp SJ, Timpson NJ, Bouatia-Naji N, Warrington NM, Kanoni S, Beilin LJ, Brage S, Deloukas P, Evans DM et al. 2011. Association of genetic Loci with glucose levels in childhood and adolescence: a meta-analysis of over 6,000 children. *Diabetes* **60**(6): 1805-1812.

Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K. 2007. High-resolution profiling of histone methylations in the human genome. *Cell* **129**(4): 823-837.

Begum F, Ghosh D, Tseng GC, Feingold E. 2012. Comprehensive literature review and statistical considerations for GWAS meta-analysis. *Nucleic acids research* **40**(9): 3777-3784.

Benjamini Y, Hochberg Y. 1995. CONTROLLING THE FALSE DISCOVERY RATE - A PRACTICAL AND POWERFUL APPROACH TO MULTIPLE TESTING. *J R Stat Soc Ser B-Methodol* **57**(1): 289-300.

Bjorkqvist M, Fex M, Renstrom E, Wierup N, Petersen A, Gil J, Bacos K, Popovic N, Li JY, Sundler F et al. 2005. The R6/2 transgenic mouse model of Huntington's disease develops diabetes due to deficient beta-cell mass and exocytosis. *Human molecular genetics* **14**(5): 565-574.

Bradley SV, Hyun TS, Oravecz-Wilson KI, Li L, Waldorff EI, Ermilov AN, Goldstein SA, Zhang CX, Drubin DG, Varela K et al. 2007. Degenerative phenotypes caused by the combined deficiency of murine HIP1 and HIP1r are rescued by human HIP1. *Human molecular genetics* **16**(11): 1279-1292.

Brandstatter A, Peterson CT, Irwin JA, Mpoke S, Koech DK, Parson W, Parsons TJ. 2004. Mitochondrial DNA control region sequences from Nairobi (Kenya): inferring phylogenetic parameters for the establishment of a forensic database. *International journal of legal medicine* **118**(5): 294-306.

Burdon KP, Macgregor S, Hewitt AW, Sharma S, Chidlow G, Mills RA, Danoy P, Casson R, Viswanathan AC, Liu JZ et al. 2011. Genome-wide association study identifies susceptibility loci for open angle glaucoma at TMCO1 and CDKN2B-AS1. *Nature genetics* **43**(6): 574-578.

Bush WS, Dudek SM, Ritchie MD. 2010. Visualizing SNP statistics in the context of linkage disequilibrium using LD-Plus. *Bioinformatics* **26**(4): 578-579.

Butler M, McKay RA, Popoff IJ, Gaarde WA, Witchell D, Murray SF, Dean NM, Bhanot S, Monia BP. 2002. Specific inhibition of PTEN expression reverses hyperglycemia in diabetic mice. *Diabetes* **51**(4): 1028-1034.

Cai D, Yuan M, Frantz DF, Melendez PA, Hansen L, Lee J, Shoelson SE. 2005. Local and systemic insulin resistance resulting from hepatic activation of IKK-beta and NF-kappaB. *Nature medicine* **11**(2): 183-190.

Chapman K, Ferreira T, Morris A, Asimit J, Zeggini E. 2011. Defining the power limits of genome-wide association scan meta-analyses. *Genetic epidemiology* **35**(8): 781-789.

Chen YH, Liu CK, Chang SC, Lin YJ, Tsai MF, Chen YT, Yao A. 2008. GenoWatch: a disease gene mining browser for association study. *Nucleic acids research* **36**(Web Server issue): W336-340.

Cheung MS, Down TA, Latorre I, Ahringer J. 2011. Systematic bias in high-throughput sequencing data and its correction by BEADS. *Nucleic acids research* **39**(15): e103.

Cho MH, Boutaoui N, Klanderman BJ, Sylvia JS, Ziniti JP, Hersh CP, DeMeo DL, Hunninghake GM, Litonjua AA, Sparrow D et al. 2010. Variants in FAM13A are associated with chronic obstructive pulmonary disease. *Nature genetics* **42**(3): 200-202.

Chun KH, Choi KD, Lee DH, Jung Y, Henry RR, Ciaraldi TP, Kim YB. 2011. In vivo activation of ROCK1 by insulin is impaired in skeletal muscle of humans with type 2 diabetes. *American journal of physiology Endocrinology and metabolism* **300**(3): E536-542.

Cirulli ET, Goldstein DB. 2010. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet* **11**(6): 415-425.

Cui B, Zhu X, Xu M, Guo T, Zhu D, Chen G, Li X, Xu L, Bi Y, Chen Y et al. 2011. A genome-wide association study confirms previously reported loci for type 2 diabetes in Han Chinese. *PloS one* **6**(7): e22353.

Curtis RE, Yin J, Kinnaird P, Xing EP. 2012. Finding genome-transcriptome-phenome association with structured association mapping and visualization in genamap. *Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing*: 327-338.

Curtis RK, Oresic M, Vidal-Puig A. 2005. Pathways to the analysis of microarray data. *Trends in biotechnology* **23**(8): 429-435.

De la Vega FM, Lazaruk KD, Rhodes MD, Wenz MH. 2005. Assessment of two flexible and compatible SNP genotyping platforms: TaqMan SNP Genotyping Assays and the SNPlex Genotyping System. *Mutation research* **573**(1-2): 111-135.

Dennis G, Jr., Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA. 2003. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol* **4**(5): 3.

Dostert A, Heinzel T. 2004. Negative glucocorticoid receptor response elements and their role in glucocorticoid action. *Current pharmaceutical design* **10**(23): 2807-2816.

Draghici S, Khatri P, Tarca AL, Amin K, Done A, Voichita C, Georgescu C, Romero R. 2007. A systems biology approach for pathway level analysis. *Genome Res* **17**(10): 1537-1545.

Dunham I Kundaje A Aldred SF Collins PJ Davis CA Doyle F Epstein CB Frietze S Harrow J Kaul R et al. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**(7414): 57-74.

Dupuis J Langenberg C Prokopenko I Saxena R Soranzo N Jackson AU Wheeler E Glazer NL Bouatia-Naji N Gloyn AL et al. 2010. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nature genetics* **42**(2): 105-116.

Ehret GB Munroe PB Rice KM Bochud M Johnson AD Chasman DI Smith AV Tobin MD Verwoert GC Hwang SJ et al. 2011. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature* **478**(7367): 103-109.

Fajans SS, Bell GI, Polonsky KS. 2001. Molecular mechanisms and clinical pathophysiology of maturity-onset diabetes of the young. *The New England journal of medicine* **345**(13): 971-980.

Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S et al. 2012. Ensembl 2012. *Nucleic acids research* **40**(Database issue): D84-90.

Folkersen L, van't Hooft F, Chernogubova E, Agardh HE, Hansson GK, Hedin U, Liska J, Syvanen AC, Paulsson-Berne G, Franco-Cereceda A et al. 2010. Association of genetic risk variants with expression of proximal genes identifies novel susceptibility genes for cardiovascular disease. *Circulation Cardiovascular genetics* **3**(4): 365-373.

Franke A, McGovern DP, Barrett JC, Wang K, Radford-Smith GL, Ahmad T, Lees CW, Balschun T, Lee J, Roberts R et al. 2010. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nature genetics* **42**(12): 1118-1125.

Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM, Lindgren CM, Perry JR, Elliott KS, Lango H, Rayner NW et al. 2007. A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* **316**(5826): 889-894.

Freathy RM, Weedon MN, Melzer D, Shields B, Hitman GA, Walker M, McCarthy MI, Hattersley AT, Frayling TM. 2006. The functional "KL-VS" variant of KLOTHO is not associated with type 2 diabetes in 5028 UK Caucasians. *BMC medical genetics* **7**: 51.

Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D, Cline MS, Goldman M, Barber GP, Clawson H, Coelho A et al. 2011. The UCSC Genome Browser database: update 2011. *Nucleic acids research* **39**(Database issue): D876-882.

Furukawa N, Ongusaha P, Jahng WJ, Araki K, Choi CS, Kim HJ, Lee YH, Kaibuchi K, Kahn BB, Masuzaki H et al. 2005. Role of Rho-kinase in regulation of insulin action and glucose homeostasis. *Cell metabolism* **2**(2): 119-129.

Gabriel S, Ziaugra L, Tabbaa D. 2009. SNP genotyping using the Sequenom MassARRAY iPLEX platform. *Current protocols in human genetics / editorial board, Jonathan L Haines [et al]* **Chapter 2**: Unit 2 12.

Gamazon ER, Zhang W, Konkashbaev A, Duan S, Kistner EO, Nicolae DL, Dolan ME, Cox NJ. 2010. SCAN: SNP and copy number annotation. *Bioinformatics* **26**(2): 259-262.

Gaubatz S, Lindeman GJ, Ishida S, Jakoi L, Nevins JR, Livingston DM, Rempel RE. 2000. E2F4 and E2F5 play an essential role in pocket protein-mediated G1 control. *Molecular cell* **6**(3): 729-735.

Geley S, Fiegl M, Hartmann BL, Kofler R. 1996. Genes mediating glucocorticoid effects and mechanisms of their regulation. *Reviews of physiology, biochemistry and pharmacology* **128**: 1-97.

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J et al. 2004. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **5**(10): R80.

Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS. 2007. Quantifying similarity between motifs. *Genome Biol* **8**(2): R24.

Heid IM Jackson AU Randall JC Winkler TW Qi L Steinthorsdottir V Thorleifsson G Zillikens MC Speliotes EK Magi R et al. 2010. Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. *Nature genetics* **42**(11): 949-960.

Heikkinen S, Vaisanen S, Pehkonen P, Seuter S, Benes V, Carlberg C. 2011. Nuclear hormone 1alpha,25-dihydroxyvitamin D3 elicits a genome-wide shift in the locations of VDR chromatin occupancy. *Nucleic acids research* **39**(21): 9181-9193.

Higgins JP, Thompson SG. 2002. Quantifying heterogeneity in a meta-analysis. *Statistics in medicine* **21**(11): 1539-1558.

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America* **106**(23): 9362-9367.

Hiromura M, Choi CH, Sabourin NA, Jones H, Bachvarov D, Usheva A. 2003. YY1 is regulated by O-linked N-acetylglucosaminylation (O-glcNAcylation). *The Journal of biological chemistry* **278**(16): 14046-14052.

Hirosumi J, Tuncman G, Chang L, Gorgun CZ, Uysal KT, Maeda K, Karin M, Hotamisligil GS. 2002. A central role for JNK in obesity and insulin resistance. *Nature* **420**(6913): 333-336.

Hirschhorn JN, Daly MJ. 2005. Genome-wide association studies for common diseases and complex traits. *Nature reviews Genetics* **6**(2): 95-108.

Hollenberg AN, Susulic VS, Madura JP, Zhang B, Moller DE, Tontonoz P, Sarraf P, Spiegelman BM, Lowell BB. 1997. Functional antagonism between CCAAT/Enhancer binding protein-alpha and peroxisome proliferator-activated receptor-gamma on the leptin promoter. *The Journal of biological chemistry* **272**(8): 5283-5290.

Hurlbert MS, Zhou W, Wasmeier C, Kaddis FG, Hutton JC, Freed CR. 1999. Mice transgenic for an expanded CAG repeat in the Huntington's disease gene develop diabetes. *Diabetes* **48**(3): 649-651.

Ingelsson E, Langenberg C, Hivert MF, Prokopenko I, Lyssenko V, Dupuis J, Magi R, Sharp S, Jackson AU, Assimes TL et al. 2010. Detailed physiologic characterization reveals diverse mechanisms for novel genetic Loci regulating glucose and insulin metabolism in humans. *Diabetes* **59**(5): 1266-1275.

Irwin DM, Tan H. 2008. Molecular evolution of the vertebrate hexokinase gene family: Identification of a conserved fifth vertebrate hexokinase gene. *Comparative biochemistry and physiology Part D, Genomics & proteomics* **3**(1): 96-107.

Jayapandian M, Chapman A, Tarcea VG, Yu C, Elkiss A, Ianni A, Liu B, Nandi A, Santos C, Andrews P et al. 2007. Michigan Molecular Interactions (MiMI): putting the jigsaw puzzle together. *Nucleic acids research* **35**(Database issue): D566-571.

Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O'Donnell CJ, de Bakker PI. 2008. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics* **24**(24): 2938-2939.

Jonsson J, Carlsson L, Edlund T, Edlund H. 1994. Insulin-promoter-factor 1 is required for pancreas development in mice. *Nature* **371**(6498): 606-609.

Jorgenson E, Kvale M, Witte JS. 2009. VALID: visualization of association study results and linkage disequilibrium. *Genetic epidemiology* **33**(7): 599-603.

Kathiresan S, Willer CJ, Peloso GM, Demissie S, Musunuru K, Schadt EE, Kaplan L, Bennett D, Li Y, Tanaka T et al. 2009. Common variants at 30 loci contribute to polygenic dyslipidemia. *Nature genetics* **41**(1): 56-65.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. *Genome research* **12**(6): 996-1006.

Khatri P, Voichita C, Kattan K, Ansari N, Khatri A, Georgescu C, Tarca AL, Draghici S. 2007. Onto-Tools: new additions and improvements in 2006. *Nucleic acids research* **35**(Web Server issue): W206-211.

Kim JH, Karnovsky A, Mahavisno V, Weymouth T, Pande M, Dolinoy DC, Rozek LS, Sartor MA. 2012. LRpath analysis reveals common pathways dysregulated via DNA methylation across cancer types. *BMC genomics* **13**(1): 526.

Kim S, Sohn KA, Xing EP. 2009. A multivariate regression approach to association analysis of a quantitative trait network. *Bioinformatics* **25**(12): i204-212.

Kim S, Xing EP. 2009. Statistical estimation of correlated genome associations to a quantitative trait network. *PLoS genetics* **5**(8): e1000587.

Kuro-o M, Matsumura Y, Aizawa H, Kawaguchi H, Suga T, Utsugi T, Ohyama Y, Kurabayashi M, Kaname T, Kume E et al. 1997. Mutation of the mouse klotho gene leads to a syndrome resembling ageing. *Nature* **390**(6655): 45-51.

Lango Allen H Estrada K Lettre G Berndt SI Weedon MN Rivadeneira F Willer CJ Jackson AU Vedantam S Raychaudhuri S et al. 2010. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**(7317): 832-838.

Lee BK, Bhinge AA, Iyer VR. 2011. Wide-ranging functions of E2F4 in transcriptional activation and repression revealed by genome-wide analysis. *Nucleic acids research* **39**(9): 3558-3573.

Lee FS, Hagler J, Chen ZJ, Maniatis T. 1997. Activation of the IkappaB alpha kinase complex by MEKK1, a kinase of the JNK pathway. *Cell* **88**(2): 213-222.

Lehner B, Crombie C, Tischler J, Fortunato A, Fraser AG. 2006. Systematic mapping of genetic interactions in Caenorhabditis elegans identifies common modifiers of diverse signaling pathways. *Nature genetics* **38**(8): 896-903.

Leung KW, Pon YL, Wong RN, Wong AS. 2006. Ginsenoside-Rg1 induces vascular endothelial growth factor expression through the glucocorticoid receptor-related phosphatidylinositol 3-kinase/Akt and beta-catenin/T-cell factor-dependent pathway in human endothelial cells. *The Journal of biological chemistry* **281**(47): 36280-36288.

Liu JZ, McRae AF, Nyholt DR, Medland SE, Wray NR, Brown KM, Hayward NK, Montgomery GW, Visscher PM, Martin NG et al. 2010. A versatile gene-based test for genome-wide association studies. *American journal of human genetics* **87**(1): 139-145.

Logie JJ, Ali S, Marshall KM, Heck MM, Walker BR, Hadoke PW. 2010. Glucocorticoid-mediated inhibition of angiogenic changes in human endothelial cells is not caused by reductions in cell proliferation or migration. *PloS one* **5**(12): e14476.

Loos RJ Lindgren CM Li S Wheeler E Zhao JH Prokopenko I Inouye M Freathy RM Attwood AP Beckmann JS et al. 2008. Common variants near MC4R are associated with fat mass, weight and risk of obesity. *Nature genetics* **40**(6): 768-775.

Lyssenko V, Eliasson L, Kotova O, Pilgaard K, Wierup N, Salehi A, Wendt A, Jonsson A, De Marinis YZ, Berglund LM et al. 2011. Pleiotropic effects of GIP on islet function involve osteopontin. *Diabetes* **60**(9): 2424-2433.

Magi R, Morris AP. 2010. GWAMA: software for genome-wide association meta-analysis. *BMC bioinformatics* **11**: 288.

Maglott D, Ostell J, Pruitt KD, Tatusova T. 2011. Entrez Gene: gene-centered information at NCBI. *Nucleic acids research* **39**(Database issue): D52-57.

Manning AK Hivert MF Scott RA Grimsby JL Bouatia-Naji N Chen H Rybin D Liu CT Bielak LF Prokopenko I et al. 2012. A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nature genetics* **44**(6): 659-669.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A et al. 2009. Finding the missing heritability of complex diseases. *Nature* **461**(7265): 747-753.

McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN. 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature reviews Genetics* **9**(5): 356-369.

McGovern DP, Gardet A, Torkvist L, Goyette P, Essers J, Taylor KD, Neale BM, Ong RT, Lagace C, Li C et al. 2010. Genome-wide association identifies multiple ulcerative colitis susceptibility loci. *Nature genetics* **42**(4): 332-337.

McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, Bejerano G. 2010. GREAT improves functional interpretation of cis-regulatory regions. *Nature biotechnology* **28**(5): 495-501.

Meyer CF, Wang X, Chang C, Templeton D, Tan TH. 1996. Interaction between c-Rel and the mitogen-activated protein kinase kinase kinase 1 signaling cascade in mediating kappaB enhancer activation. *The Journal of biological chemistry* **271**(15): 8971-8976.

Mokry M, Hatzis P, Schuijers J, Lansu N, Ruzius FP, Clevers H, Cuppen E. 2012. Integrated genome-wide analysis of transcription factor occupancy, RNA polymerase II binding and steady-state RNA levels identify differentially regulated functional gene classes. *Nucleic acids research* **40**(1): 148-158.

Morris AP Voight BF Teslovich TM Ferreira T Segre AV Steinthorsdottir V Strawbridge RJ Khan H Grallert H Mahajan A et al. 2012. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature genetics*.

Mullin BH, Wilson SG, Islam FM, Calautti M, Dick IM, Devine A, Prince RL. 2005. Klotho gene polymorphisms are associated with osteocalcin levels but not bone density of aged postmenopausal women. *Calcified tissue international* **77**(3): 145-151.

Nakamura Y, Kaneto H, Miyatsuka T, Matsuoka TA, Matsuhisa M, Node K, Hori M, Yamasaki Y. 2006. Marked increase of insulin gene transcription by suppression of the Rho/Rho-kinase pathway. *Biochemical and biophysical research communications* **350**(1): 68-73.

Nicolino M, Claiborn KC, Senee V, Boland A, Stoffers DA, Julier C. 2010. A novel hypomorphic PDX1 mutation responsible for permanent neonatal diabetes with subclinical exocrine deficiency. *Diabetes* **59**(3): 733-740.

Nomiyama T, Perez-Tilve D, Ogawa D, Gizard F, Zhao Y, Heywood EB, Jones KL, Kawamori R, Cassis LA, Tschop MH et al. 2007. Osteopontin mediates obesity-induced adipose tissue macrophage infiltration and insulin resistance in mice. *The Journal of clinical investigation* **117**(10): 2877-2888.

Oguro R, Kamide K, Kokubo Y, Shimaoka I, Congrains A, Horio T, Hanada H, Ohishi M, Katsuya T, Okamura T et al. 2010. Association of carotid atherosclerosis with genetic polymorphisms of the klotho gene in patients with hypertension. *Geriatrics & gerontology international* **10**(4): 311-318.

Ohnishi M, Kato S, Akiyoshi J, Atfi A, Razzaque MS. 2011. Dietary and genetic evidence for enhancing glucose metabolism and reducing obesity by inhibiting klotho functions. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology* **25**(6): 2031-2039.

Olofsson LE, Orho-Melander M, William-Olsson L, Sjoholm K, Sjostrom L, Groop L, Carlsson B, Carlsson LM, Olsson B. 2008. CCAAT/enhancer binding protein alpha (C/EBPalpha) in adipose tissue regulates genes in lipid and glucose metabolism and a genetic variation in C/EBPalpha is associated with serum levels of triglycerides. *The Journal of clinical endocrinology and metabolism* **93**(12): 4880-4886.

Ovcharenko I, Loots GG, Nobrega MA, Hardison RC, Miller W, Stubbs L. 2005. Evolution and functional classification of vertebrate gene deserts. *Genome research* **15**(1): 137-145.

Park JH, Wacholder S, Gail MH, Peters U, Jacobs KB, Chanock SJ, Chatterjee N. 2010. Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nature genetics* **42**(7): 570-575.

Park PJ. 2009. ChIP-seq: advantages and challenges of a maturing technology. *Nature reviews Genetics* **10**(10): 669-680.

Paroni G, Seripa D, Panza F, Addante F, Copetti M, D'Onofrio G, Pellegrini F, Fontana L, Pilotto A. 2012. Klotho locus, metabolic traits, and serum hemoglobin in hospitalized older patients: a genetic association analysis. *Age (Dordr)* **34**(4): 949-968.

Pietilainen KH, Naukkarinen J, Rissanen A, Saharinen J, Ellonen P, Keranen H, Suomalainen A, Gotz A, Suortti T, Yki-Jarvinen H et al. 2008. Global transcript profiles of fat in monozygotic twins discordant for BMI: pathways behind acquired obesity. *PLoS medicine* **5**(3): e51.

Plesca D, Crosby ME, Gupta D, Almasan A. 2007. E2F4 function in G2: maintaining G2-arrest to prevent mitotic entry with damaged DNA. *Cell Cycle* **6**(10): 1147-1152.

Prokopenko I Langenberg C Florez JC Saxena R Soranzo N Thorleifsson G Loos RJ Manning AK Jackson AU Aulchenko Y et al. 2009. Variants in MTNR1B influence fasting glucose levels. *Nature genetics* **41**(1): 77-81.

Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, Boehnke M, Abecasis GR, Willer CJ. 2010. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* **26**(18): 2336-2337.

Pulit SL, Voight BF, de Bakker PI. 2010. Multiethnic genetic association studies improve power for locus discovery. *PloS one* **5**(9): e12600.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics* **81**(3): 559-575.

Raychaudhuri S, Plenge RM, Rossin EJ, Ng AC, Purcell SM, Sklar P, Scolnick EM, Xavier RJ, Altshuler D, Daly MJ. 2009a. Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. *PLoS genetics* **5**(6): e1000534.

Raychaudhuri S, Thomson BP, Remmers EF, Eyre S, Hinks A, Guiducci C, Catanese JJ, Xie G, Stahl EA, Chen R et al. 2009b. Genetic variants at CD28, PRDM1 and CD2/CD58 are associated with rheumatoid arthritis risk. *Nature genetics* **41**(12): 1313-1318.

Reddy TE, Pauli F, Sprouse RO, Neff NF, Newberry KM, Garabedian MJ, Myers RM. 2009. Genomic determination of the glucocorticoid response reveals unexpected mechanisms of gene regulation. *Genome research* **19**(12): 2163-2171.

Rhee EJ, Oh KW, Yun EJ, Jung CH, Lee WY, Kim SW, Baek KH, Kang MI, Park SW. 2006. Relationship between polymorphisms G395A in promoter and C1818T in exon 4 of the KLOTHO gene with glucose metabolism and cardiovascular risk factors in Korean women. *Journal of endocrinological investigation* **29**(7): 613-618.

Richards JB, Waterworth D, O'Rahilly S, Hivert MF, Loos RJ, Perry JR, Tanaka T, Timpson NJ, Semple RK, Soranzo N et al. 2009. A genome-wide association study reveals variants in ARL15 that influence adiponectin levels. *PLoS genetics* **5**(12): e1000768.

Rivals I, Personnaz L, Taing L, Potier MC. 2007. Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics* **23**(4): 401-407.

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

Rye M, Saetrom P, Handstad T, Drablos F. 2011. Clustered ChIP-Seq-defined transcription factor binding sites and histone modifications map distinct classes of regulatory elements. *BMC biology* **9**: 80.

Sartor MA, Mahavisno V, Keshamouni VG, Cavalcoli J, Wright Z, Karnovsky A, Kuick R, Jagadish HV, Mirel B, Weymouth T et al. 2010. ConceptGen: a gene set enrichment and gene set relation mapping tool. *Bioinformatics* **26**(4): 456-463.

Saxena R Hivert MF Langenberg C Tanaka T Pankow JS Vollenweider P Lyssenko V Bouatia-Naji N Dupuis J Jackson AU et al. 2010. Genetic variation in GIPR influences the glucose and insulin responses to an oral glucose challenge. *Nature genetics* **42**(2): 142-148.

Schmid J, Ludwig B, Schally AV, Steffen A, Ziegler CG, Block NL, Koutmani Y, Brendel MD, Karalis KP, Simeonovic CJ et al. 2011. Modulation of pancreatic islets-stress axis by hypothalamic releasing hormones and 11beta-hydroxysteroid dehydrogenase. *Proceedings of the National Academy of Sciences of the United States of America* **108**(33): 13722-13727.

Schmitt AO, Assmus J, Bortfeldt RH, Brockmann GA. 2010. CandiSNPer: a web tool for the identification of candidate SNPs for causal variants. *Bioinformatics* **26**(7): 969-970.

Scott RA Lagou V Welch RP Wheeler E Montasser ME Luan J Magi R Strawbridge RJ Rehnberg E Gustafsson S et al. 2012. Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. *Nature genetics*.

Shendure J, Ji HL. 2008. Next-generation DNA sequencing. *Nature biotechnology* **26**(10): 1135-1145.

Shimoyama Y, Nishio K, Hamajima N, Niwa T. 2009. KLOTHO gene polymorphisms G-395A and C1818T are associated with lipid and glucose metabolism, bone mineral density and systolic blood pressure in Japanese healthy subjects. *Clinica chimica acta; international journal of clinical chemistry* **406**(1-2): 134-138.

Skol AD, Scott LJ, Abecasis GR, Boehnke M. 2006. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nature genetics* **38**(2): 209-213.

Small GR, Hadoke PW, Sharif I, Dover AR, Armour D, Kenyon CJ, Gray GA, Walker BR. 2005. Preventing local regeneration of glucocorticoids by 11beta-hydroxysteroid dehydrogenase type 1 enhances angiogenesis. *Proceedings of the National Academy of Sciences of the United States of America* **102**(34): 12165-12170.

Soranzo N Sanna S Wheeler E Gieger C Radke D Dupuis J Bouatia-Naji N Langenberg C Prokopenko I Stolerman E et al. 2010. Common variants at 10 genomic loci influence hemoglobin A(1)(C) levels via glycemic and nonglycemic pathways. *Diabetes* **59**(12): 3229-3239.

Speliotes EK Willer CJ Berndt SI Monda KL Thorleifsson G Jackson AU Allen HL Lindgren CM Luan J Magi R et al. 2010. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature genetics* **42**(11): 937-948.

Spiegelman BM. 1998. PPAR-gamma: adipogenic regulator and thiazolidinedione receptor. *Diabetes* **47**(4): 507-514.

Spyrou C, Stark R, Lynch AG, Tavare S. 2009. BayesPeak: Bayesian analysis of ChIP-seq data. *BMC bioinformatics* **10**: 299.

Stoffers DA, Ferrer J, Clarke WL, Habener JF. 1997a. Early-onset type-II diabetes mellitus (MODY4) linked to IPF1. *Nature genetics* **17**(2): 138-139.

Stoffers DA, Zinkin NT, Stanojevic V, Clarke WL, Habener JF. 1997b. Pancreatic agenesis attributable to a single nucleotide deletion in the human IPF1 gene coding sequence. *Nature genetics* **15**(1): 106-110.

Strawbridge RJ Dupuis J Prokopenko I Barker A Ahlqvist E Rybin D Petrie JR Travers ME Bouatia-Naji N Dimas AS et al. 2011. Genome-wide association identifies nine common variants associated with fasting proinsulin levels and provides new insights into the pathophysiology of type 2 diabetes. *Diabetes* **60**(10): 2624-2634.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES et al. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **102**(43): 15545-15550.

Taher L, Ovcharenko I. 2009. Variable locus length in the human genome leads to ascertainment bias in functional inference for non-coding elements. *Bioinformatics* **25**(5): 578-584.

Tanaka T, Yamamoto J, Iwasaki S, Asaba H, Hamura H, Ikeda Y, Watanabe M, Magoori K, Ioka RX, Tachibana K et al. 2003. Activation of peroxisome proliferator-activated receptor delta induces fatty acid beta-oxidation in skeletal muscle and attenuates metabolic syndrome. *Proceedings of the National Academy of Sciences of the United States of America* **100**(26): 15924-15929.

Tang AT, Campbell WB, Nithipatikom K. 2012. ROCK1 feedback regulation of the upstream small GTPase RhoA. *Cellular signalling* **24**(7): 1375-1380.

Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. 1999. Systematic determination of genetic network architecture. *Nature genetics* **22**(3): 281-285.

Teslovich TM Musunuru K Smith AV Edmondson AC Stylianou IM Koseki M Pirruccello JP Ripatti S Chasman DI Willer CJ et al. 2010. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**(7307): 707-713.

Thornblad TA, Elliott KS, Jowett J, Visscher PM. 2007. Prioritization of positional candidate genes using multiple web-based software tools. *Twin research and human genetics : the official journal of the International Society for Twin Studies* **10**(6): 861-870.

Upton GJG. 1992. Fisher's Exact Test. *Journal of the Royal Statistical Society Series A (Statistics in Society)* **155**(3): 395-402.

Utsugi T, Ohno T, Ohyama Y, Uchiyama T, Saito Y, Matsumura Y, Aizawa H, Itoh H, Kurabayashi M, Kawazu S et al. 2000. Decreased insulin production and increased insulin sensitivity in the klotho mutant mouse, a novel animal model for human aging. *Metabolism: clinical and experimental* **49**(9): 1118-1123.

Voight BF, Kang HM, Ding J, Palmer CD, Sidore C, Chines PS, Burtt NP, Fuchsberger C, Li Y, Erdmann J et al. 2012. The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS genetics* **8**(8): e1002793.

Voight BF Scott LJ Steinthorsdottir V Morris AP Dina C Welch RP Zeggini E Huth C Aulchenko YS Thorleifsson G et al. 2010. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nature genetics* **42**(7): 579-589.

Wang J, Zhuang J, Iyer S, Lin X, Whitfield TW, Greven MC, Pierce BG, Dong X, Kundaje A, Cheng Y et al. 2012. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome research* **22**(9): 1798-1812.

Wang K, Dickson SP, Stolle CA, Krantz ID, Goldstein DB, Hakonarson H. 2010. Interpretation of association signals and identification of causal variants from genome-wide association studies. *American journal of human genetics* **86**(5): 730-742.

White MF. 1998. The IRS-signalling system: a network of docking proteins that mediate insulin action. *Molecular and cellular biochemistry* **182**(1-2): 3-11.

Wijesekara N, Konrad D, Eweida M, Jefferies C, Liadis N, Giacca A, Crackower M, Suzuki A, Mak TW, Kahn CR et al. 2005. Muscle-specific Pten deletion protects against insulin resistance and diabetes. *Molecular and cellular biology* **25**(3): 1135-1145.

Willer CJ, Li Y, Abecasis GR. 2010. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**(17): 2190-2191.

Willer CJ Speliotes EK Loos RJ Li S Lindgren CM Heid IM Berndt SI Elliott AL Jackson AU Lamina C et al. 2009. Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nature genetics* **41**(1): 25-34.

Withers DJ, Burks DJ, Towery HH, Altamuro SL, Flint CL, White MF. 1999. Irs-2 coordinates Igf-1 receptor-mediated beta-cell development and peripheral insulin signalling. *Nature genetics* **23**(1): 32-40.

Withers DJ, Gutierrez JS, Towery H, Burks DJ, Ren JM, Previs S, Zhang Y, Bernal D, Pons S, Shulman GI et al. 1998. Disruption of IRS-2 causes type 2 diabetes in mice. *Nature* **391**(6670): 900-904.

Wood SN. 2006. *Generalized additive models : an introduction with R*. Chapman & Hall/CRC.

Wood SN. 2010. mgcv: GAMs with GCV/AIC/REML smoothness estimation and GAMMs by PQL. *R package version*: 1.6-2.

Wood SN. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73**(1): 3-36.

Wu Z, Rosen ED, Brun R, Hauser S, Adelmant G, Troy AE, McKeon C, Darlington GJ, Spiegelman BM. 1999. Cross-regulation of C/EBP alpha and PPAR gamma controls the transcriptional pathway of adipogenesis and insulin sensitivity. *Molecular cell* **3**(2): 151-158.

Xu C, He J, Jiang H, Zu L, Zhai W, Pu S, Xu G. 2009. Direct effect of glucocorticoids on lipolysis in adipocytes. *Mol Endocrinol* **23**(8): 1161-1170.

Yamagata K, Oda N, Kaisaki PJ, Menzel S, Furuta H, Vaxillaire M, Southam L, Cox RD, Lathrop GM, Boriraj VV et al. 1996. Mutations in the hepatocyte nuclear factor-1alpha gene in maturity-onset diabetes of the young (MODY3). *Nature* **384**(6608): 455-458.

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW et al. 2010a. Common SNPs explain a large proportion of the heritability for human height. *Nature genetics* **42**(7): 565-569.

Yang Q, Kottgen A, Dehghan A, Smith AV, Glazer NL, Chen MH, Chasman DI, Aspelund T, Eiriksdottir G, Harris TB et al. 2010b. Multiple genetic loci influence serum urate levels and their relationship with gout and cardiovascular disease risk factors. *Circulation Cardiovascular genetics* **3**(6): 523-530.

Yano A, Fujii Y, Iwai A, Kageyama Y, Kihara K. 2006. Glucocorticoids suppress tumor angiogenesis and in vivo growth of prostate cancer cells. *Clinical cancer research : an official journal of the American Association for Cancer Research* **12**(10): 3003-3009.

Young MD, Wakefield MJ, Smyth GK, Oshlack A. 2010. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol* **11**(2): R14.

Yu CY, Mayba O, Lee JV, Tran J, Harris C, Speed TP, Wang JC. 2010. Genome-wide analysis of glucocorticoid receptor binding regions in adipocytes reveal gene network involved in triglyceride homeostasis. *PloS one* **5**(12): e15188.

Yujiri T, Sather S, Fanger GR, Johnson GL. 1998. Role of MEKK1 in cell survival and activation of JNK and ERK pathways defined by targeted gene disruption. *Science* **282**(5395): 1911-1914.

Yuneva MO, Fan TW, Allen TD, Higashi RM, Ferraris DV, Tsukamoto T, Mates JM, Alonso FJ, Wang C, Seo Y et al. 2012. The metabolic profile of tumors depends on both the responsible genetic lesion and tissue type. *Cell metabolism* **15**(2): 157-170.

Zarrabeitia MT, Hernandez JL, Valero C, Zarrabeitia AL, Ortiz F, Gonzalez-Macias J, Riancho JA. 2007. Klotho gene polymorphism and male bone mass. *Calcified tissue international* **80**(1): 10-14.

Zeggini E, Scott LJ, Saxena R, Voight BF, Marchini JL, Hu T, de Bakker PI, Abecasis GR, Almgren P, Andersen G et al. 2008. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nature genetics* **40**(5): 638-645.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W et al. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**(9): R137.