Candidate-Based Approaches to Identify Genetic Variation Influencing Type 2 Diabetes and Quantitative Traits

Kyle Jeffrie Gaulton

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Curriculum of Genetics and Molecular Biology

Chapel Hill

2010

Approved By:

Ethan Lange, Ph.D.

Jason Lieb, Ph.D.

Karen Mohlke, Ph.D.

Todd Vision, Ph.D.

Kirk Wilhelmsen, M.D., Ph.D.

ABSTRACT

Candidate-based approaches to identify genetic variation influencing type 2 diabetes and quantitative traits

(Under the direction of Karen Mohlke)

Type 2 diabetes (T2D) is a metabolic disorder characterized by insulin resistance and impaired insulin secretion that affects more than 20 million Americans, although the genetic component of the disorder is largely unknown. Individual genetic susceptibility to type 2 diabetes and other complex traits is the result of variation that is both common in human populations and rare, *de novo* and inherited mutations. We adopted a diverse set of genetics, genomics and informatics approaches to prioritize candidate genomic regions and variants and perform in-depth, targeted analysis of their contributions to type 2 diabetes susceptibility and related trait variability. Our initial efforts focused on the selection of candidate genes relevant to a complex trait by developing a metric to weight the relevance of functional gene annotations to the known biology of a trait. We used this method to select candidate genes for type 2 diabetes and performed a T2D case-control and quantitative trait association study in 2,335 Finnish individuals from the FUSION study. After follow-up in additional samples, we identified several variants that might contribute to T2D susceptibility. Genomic regions associated with plasma levels of HDL cholesterol and triglycerides were re-sequenced in individuals with

trait-extreme values. Our analysis revealed a denser set of common and rare functional target variants including several non-synonymous, 3' UTR, and non-coding SNPs and indels. Finally, we utilized two approaches to identify candidate functional non-coding variants that may directly contribute to trait susceptibility. First, we used Formaldehyde-assisted isolation of regulatory elements (FAIRE) coupled with high-throughput sequencing to identify nucleosome-depleted regions in pancreatic islets. We used islet FAIRE-seq data to identify SNPs associated with T2D that potentially alter islet transcriptional regulation. A SNP in *TCF7L2,* rs7903146, was located in a FAIRE-seq site and demonstrated allelic differences in islet chromatin openness and enhancer activity, suggesting that it may contribute functionally to T2D susceptibility. Second, we used transcription factor binding site motifs to computationally predict variants that have allelic differences in regulatory activity. Taken together, these results suggest that identifying candidate genomic regions can successfully enrich for variation important for type 2 diabetes and other complex traits.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

293T	human embryonic kidney cell line
3C	chromatin conformation capture
5C	chromatin conformation capture carbon copy
832/13	rat insulinoma-derived cell line
ANOVA	analysis of variance
ARMD	age-related macular degeneration
ASP	affected sib-pairs
AUC	area under curve
BIND	biomolecular interaction database
BMI	body-mass index
BP	base pair
CAD	coronary artery disease
CAESAR	Candidate search and rank
CEU	Utah residents from Western and Northern European ancestry

ChIP	chromatin immunoprecipitation
ChIPOTle	Chomatin immunoprecipitation On Tiled arrays
CIDR	Center for Inherited Disease Research
cM	centimorgan
CHR	chromosome
CNV	copy number variant
CORE	cluster of open regulatory elements
DB	database
DGI	diabetes genetics initiative
DIAGEN	Diabetes genetics study
DIAGRAM	Diabetes genetics replication and meta-analysis consortium
DNA	Deoxyribonucleic acid
DHS	DNase I hypersensitivity
DSP	digital signal processing
HOMA-B	homeostasis model adjustment, beta-cell function
HOMA-IR	homeostasis model adjustment, insulin resistance

ENCODE	ENCyclopedia Of DNA Elements
EST	expressed sequence tag
eVOC	expression vocabulary ontology
FAIRE	Formaldehyde-assisted isolation of regulatory elements
FG	fasting glycemia
FUSION	Finland-United States Investigation of NIDDM genetics study
GAD	genetic association database
gDNA	genomic deoxyribonucleic acid
GM12878	Lymphoblastoid cell line
GO bp	gene ontology biological process
GO mf	gene ontology molecular function
GOA	gene ontology annotation
GWA	genome-wide association study
HapMap	International Haplotype Map project
HDL-C	high-density lipoprotein cholesterol
HeLa-S3	Human epithelial carcinoma cell line

HepG2	Hepatocellular carcinoma cell line
hME	homogeneous MassEXTEND assay (Sequenom)
HPRD	human protein reference database
HUNT	Nord-Trondelag Health Study
HUVEC	Human Umbilical Vein Endothelial Cells
HWE	Hardy-Weinberg equilibrium
IBD	identity by descent
IBS	Identity by state
Indel	insertion / deletion
IPRO	InterPro
K562	Human myelogenous leukaemia cell line
KB	kilobase
KEGG	Kyoto encyclopedia of genes and genomes
LD	linkage disequilibrium
LDL-C	low-density lipoprotein cholesterol
MACH	Markov chain haplotyping

MAF	minor allele frequency
MAQ	mapping and assembly with qualities
MB	megabase
METSIM	Metabolic syndrome in man study
MGD	mouse genome database
MIN6	mouse insulinoma cell line
MODY	Mature Onset Diabetes of the Young
MP	mammalian phenotype ontology
NGT	normal glucose tolerant
NIDDM	non-insulin dependant diabetes mellitus
NMI	Non-mendelian inheritance
OMIM	online mendelian inheritance in man
PCR	polymerase chain reaction
PPI	protein-protein interaction
Q-Q	Quantile-quantile
QTL	Quantitative trait locus

RMA	robust multichip average
RNA	Ribonucleic acid
ROC	Receiver-operating characteristic
SD	standard deviation
SEM	standard error of the mean
SNP	single nucleotide polymorphism
SNR	signal to noise ratio
T1D	type 1 diabetes
T2D	type 2 diabetes
TFBS	transcription factor binding site
TG	triglyceride
TSS	transcription start site
UTR	untranslated region
WC	waist circumference
WHR	waist-to-hip ratio
WTCCC	Wellcome Trust case control consortium

Chapter I

Introduction

The genetics of complex traits

The majority of phenotypic variability between humans is the result of both genetic and environmental factors that are neither sufficient nor necessary determinants of the phenotype[1]. This complex pattern of inheritance is found in both traits that are common to all individuals (quantitative), such as anthropometric traits and plasma levels of biomarkers, as well as those that are only found in a subset of the population (qualitative), including the majority of cardiovascular [2], metabolic [3], autoimmune [4] and psychiatric disorders [5-7]. As the both impact on human health and the genetic heritability for these disorders is substantial, there is considerable interest in understanding the underlying genetic factors. However, these factors are just beginning to be identified.

Genetic diversity in humans includes factors of both vastly different allele frequencies and genomic sizes [8]. On a macro scale are entire chromosomal deletions or duplications, chromosomal rearrangements and copy number variable (CNV) regions that can be several megabases in size. CNVs can also be a few base pairs, such as microsatellites, as can insertions / deletions (indels). At the smallest scale are single base pair indels or changes (SNPs). SNPs account for the majority of human genetic variation by frequency, as there are roughly 10 million that are common (minor allele frequency > .01) in human

populations as well as millions more rare SNPs [9]. Despite changing only a single base, SNPs likely contribute a great amount to phenotypic differences between individuals. SNPs can functionally cause differences in protein products (nonsynonymous substitutions), transcript stability, transcript expression level, transcript splicing, transcription factor binding, chromatin accessibility, RNA folding and secondary structure, among many other mechanisms [10].

One traditional mechanism to identify factors influencing genetic disorders is family-based linkage analysis using polymorphic markers such as microsatellites to identify stretches of the genome that individuals with a disease share through common ancestry (identity by descent - IBD) and that might contain variants that contribute to disease susceptibility [11]. However, for complex traits, the effect size of the contributing variants is often small, and thus difficult to identify through IBD approaches [11]. In addition, the relatively low resolution of linkage peaks coupled with the polygenic inheritance of complex traits means that a large percentage of the genome is often covered by these analyses.

Alternatives are approaches that use identity by state (IBS), direct tests of a genetic marker for trait association. Technological limitations previously restricted the number of variants that could be interrogated for disease association, requiring the selection of a small number of known variants for study[11]. Candidate gene approaches have typically been used to identify variants of interest, which rely on prior biological knowledge of genes to determine a potential relationship to disease [12]. Informatics-assisted methods can help consolidate biological data into gene annotation to assist candidate selection [13-

[18]. Many are designed to assist positional cloning of a linkage peak, although for complex traits few convincing linkage peaks have been identified [19]. Further, predictive quality of these methods is restricted to available biological annotation data and the known etiology of a disorder, and for many complex traits the underlying biological mechanisms are poorly understood [20].

Several advances have made large-scale, more comprehensive IBS studies possible. First, the International Haplotype Map (HapMap) project catalogued common SNPs in four human populations[21]. This work helped consolidate knowledge of the existence and allele frequencies of many SNPs based on previous human genome sequencing studies[22], as well as patterns of linkage disequilibrium (LD) between them [21]. Second, advances in technology made it possible to genotype thousands to hundreds of thousands of markers at once [23-26]. Third, computational methods to impute untyped markers allowed genotype inference for many of the remaining HapMap markers not directly present on a genotyping chip [27]. The resulting candidate gene and genome-wide association (GWA) studies of common SNPs in population-based cohorts have identified numerous susceptibility loci for a wide variety of complex traits and disorders [28].

The next generation of complex trait genetics involves both population-based and phenotype guided sequencing studies to identify a broader spectrum of allele frequencies and types of variation influencing trait susceptibility [28]. The primary aim of the former is to identify a denser set of linkage disequilibrium patterns that can be used to fill in association studies of an incomplete marker set. Sequencing studies aimed at uncovering LD patterns have traditionally targeted specific genomic regions of interest, based on

candidate gene selection or the results of previous genetic association. Advances in sequencing technology, however, have made large scale, even full-genome, sequencing of many samples a technical possibility. One of the goals of the 1000 Genomes Project (www.1000genomes.org) is to perform low-pass full-genome sequencing of an ethnically diverse set of individuals to catalog SNP, indel and CNV variation down to population allele frequencies as low as .005, and may obviate the need for *de novo* sequencing for LD discovery completely.

Medical sequencing projects enrich for rare alleles that are not present or present in low frequency in the general population yet may contribute to individual phenotypes. Many previous candidate gene studies focused on coding regions, where variants affecting protein products are easier to predict [29-33]. As statistical power to detect association with rare, often private, variants is low, these studies often group variants either by disease status or trait value. Recently, sequencing of the entire coding sequence (exome) of individuals has been used to find mutations causing a Mendelian disorder [34]. Full genome sequencing of phenotyped samples will eventually allow interrogation of the entire profile of variation in individuals with complex disorders.

The genetics of type 2 diabetes

Diabetes mellitus is a heterogeneous collection of disorders characterized by high levels of blood glucose, which can cause heart disease, stroke, neuropathy, blindness and reduced life expectancy[35]. The most prevalent form of diabetes is type 2 (T2D), previously known as non-insulin dependant diabetes mellitus (NIDDM), which accounts for 90-95% of cases and affects more than 20 million people in the United States [9]. High

blood glucose levels in type 2 diabetes are sustained by peripheral tissue resistance to insulin, which normally triggers glucose uptake, and reduced insulin secretion by the beta cells of the pancreatic islets[36].

Type 2 diabetes has a complex pattern of inheritance, influenced by numerous factors, such as diet, weight, amount of physical activity, smoking, sleep patterns, as well as those that are heritable [3,35]. Evidence supportive of the heritability of T2D includes increased concordance in monozygotic compared to dizygotic twins, monogenic forms of the disorder, and common risk factors [37,38].

Genome-wide scans of family cohorts had previously identified numerous regions of the genome linked to type 2 diabetes[39-44], although many of these regions did not consistently replicate across populations or identify convincing association signals upon positional cloning, and a meta-analysis across many T2D linkage studies resulted in only modest signals [45,46]. However, one follow-up study of a linkage peak identified on chromosome 10q in Icelandic individuals successfully localized to a variant in intron 3 of the transcription factor *TCF7L2* with a large relative risk[47]. Candidate gene association studies also similarly often identified only modest associations with SNP variants, and amino acid changing substitutions in *PPARG*[48] and *KCNJ11*[49] were until recently among the only variants with convincing replication across populations and large sample sizes. Variants near *WFS1* [50] and *TCF2* [51,52] have subsequently been confirmed through candidate gene studies in large samples.

Genome-wide studies have implicated a much larger set of genes in T2D susceptibility, including many that were previously unreported through linkage or candidate gene

studies. The first reported T2D genome-wide association (GWA) scan implicated variants at five susceptibility loci that include *TCF7L2,* and novel loci near the genes *SLC30A8, IDE-KIF11-HHEX, LOC387761,* and *EXT-ALX4* [53]. Three companion GWA studies replicated evidence for *PPARG, KCNJ11, TCF7L2, SLC30A8, IDE-KIF11-HHEX,* and provided new evidence for *CDKAL1, CDKN2A/B, IGF2BP2,* and *FTO* [54-56]. Additional first-generation GWA studies provided additional evidence for *TCF7L2*, *CDKAL1* and *SLC30A8* [57-62]. Subsequent studies have implicated variants near *MTNR1B* [63,64], *IRS1* [65], a region on chromosome 11q [66], and *KCNQ1* [67,68]. More recently, meta-analyses of the results of several genome-wide association studies following by replication in more than 50,000 samples identified six additional loci, including variants near *JAZF1*, *CDC123/CAMKD1*, *TSPAN8/LGR5*, *THADA, ADAMTS9,* and *NOTCH2* [69].

In total, more than 20 independent genetic loci are known to harbor common risk alleles for type 2 diabetes, although the additive effects of these variants explain only 5-10% of trait heritability[70,71]. Many of these risk variants appear to influence aspects of pancreatic beta cell function leading to impaired insulin secretion [70,72]. In addition, numerous loci have been identified that influence quantitative traits related to T2D pathogenesis including body-mass index (BMI), plasma lipid level (HDL-C, LDL-C, TG), and plasma fasting glucose and insulin levels. Association studies of T2D-related quantitative traits are likely to help identify T2D susceptibility loci. For example, of 18 loci recently identified to influence fasting glucose or insulin level, variants at five also demonstrated novel association with type 2 diabetes (*ADCY5, PROX1, GCK, GCKR,* and *DGKB/TMEM195)* [73], and several BMI and lipid level quantitative trait loci overlap known T2D susceptibility loci [74]. For the vast majority of loci, however, the direct

mechanisms of how they contribute to T2D susceptibility or quantitative trait variability is unknown.

Finland-United States Investigation of NIDDM (FUSION) genetics study

The Finland-United States Investigation of NIDDM genetics (FUSION) study aims to map and identify variants influencing susceptibility to T2D and related quantitative traits in the Finnish population [75]. There are several advantages to using Finns as a model population for genetic study. The majority of Finland belongs to one linguistic and ethnic group formed by a small founder population that was linguistically, culturally and geographically isolated. Clinical aspects of T2D in Finland are similar to those in other European populations, suggesting that susceptibility loci identified in Finnish samples may be more generally applicable.

Initially, FUSION ascertained Finnish affected sib-pair (ASP) families to perform a genome-wide linkage scan for T2D susceptibility loci. The sample collection criteria included selecting probands of 35-60 years of age with no family history of type 1 diabetes, where at least one sibling was also affected by T2D and one parent was normal glucose tolerant (NGT). In addition, a set of extended families was ascertained with non-diabetic spouses, siblings and offspring, as well as a set of elderly controls. Blood samples and clinical measurements were collected from all study participants in 21 cities distributed throughout Finland.

Linkage and association analysis for type 2 diabetes were performed on 580 FUSION ASP families and controls genotyped on roughly 400 microsatellite markers spaced

across the genome [76]. Regions on chromosomes 20, 11 and 6q showed the strongest evidence of linkage, and a region on chromosome 22 had the most significant association; however, none of the results reached accepted thresholds for genome-wide significance. The T2D and NGT control individuals were also separately analyzed for QTL linkage using a set of T2D-related quantitative traits [77]. Among numerous QTLs were several that overlapped those identified in the T2D analysis on multiple chromosomes. An additional linkage analysis of 275 ASP Finnish families (FUSION 2) was performed and the results were combined with the initial linkage analysis. Regions on chromosome 6, 11, 14 and X were most interesting across both studies, and were fine-mapped using a denser set of microsatellite markers [41].

In several candidate gene association studies, FUSION identified modest T2D association in Finns with variants in *HNF4A* [78], four genes known to cause MODY [79] [78] [80], *PPARG, KCNJ11*, *ENPP1, SLC2A2, PCK1, TNF, IL6* [79], and *TCF7L2* [81].

A two-stage genome-wide association study of type 2 diabetes with 2,335 Stage 1 and 2,473 Stage 2 samples from the FUSION and Finrisk 2002 studies confirmed associations with variants in *PPARG, KCNJ11, TCF7L2, HHEX, SLC30A8, FTO*, contributed to novel associations with variants near *CDKAL1*, *CDKN2A/B, IGF2BP2*, and identified an association with variants in an intergenic region of chromosome 11q12 [55]. FUSION has since contributed to a meta-analysis of T2D GWA results with the DGI and WTCCC consortiums to help identify susceptibility variants near *JAZF1*, *CDC123/CAMKD1*, *TSPAN8/LGR5*, *THADA, ADAMTS9,* and *NOTCH2* [69].

FUSION has also contributed to the identification of loci influencing quantitative trait variability including BMI [82], blood pressure [83], waist circumference (WC) and waist-to-hip ratio (WHR) [84], plasma lipid level (HDL-C, LDL-C, TG) [85,86], fasting glucose and insulin level, measures of beta-cell function (homeostasis model adjustment (HOMA)-B) and insulin resistance (HOMA-IR) [73], and height [87].

Currently, FUSION has 2,335 samples with GWA data and 4,937 samples for replication studies. In addition, approximately 14,000 samples from the HUNT, METSIM and DIAGEN studies are also used for replication of quantitative trait results.

Determining the functional basis for complex trait association

Variants associated with a complex trait do not necessarily contribute functionally to differences in trait susceptibility and can be in linkage disequilibrium with true functional variant(s) [11]. While this is advantageous for genetic studies that interrogate an incomplete set of markers, the ensuing process of sorting functional variants from those merely inherited on the same haplotypes can be non-trivial. Due to limited recombination events in the human genome, associated variants can often be in linkage disequilibrium with many additional variants. For example, at one locus on chromosome 12 associated with high-density lipoprotein cholesterol (HDL-C) level, there are over 50 HapMap SNPs in high LD (r^2>.8) spanning a region of approximately 200 kb [86]. In addition, the completion of the 1000 Genomes Project will uncover on average about three times more variants in LD than in HapMap [21]. Many loci do not contain an obvious functional trait-associated variant, such as a frameshift early in a protein, splice site, or non-synonymous substitution predicted to be deleterious. Given that it is not efficient to

test all variants for differential activity in the laboratory, it is often necessary to prioritize between variants to guide functional follow-up studies.

Variants influencing transcriptional regulation are likely to contribute to complex trait variability [88] at sites driving gene expression proximal to transcription start sites (TSS) (promoters), distal sites increasing or decreasing the amount of expression (enhancers / silencers), or sites blocking the activity of distal regulatory elements (insulators) [89]. The accessibility of regulatory DNA sequence to DNA-binding proteins that control transcriptional regulation at these sites is largely controlled by chromatin packaging into nucleosomes, which consist of DNA wrapped around a core of histone proteins. Histones contain N-terminal tails that can be post-translationally modified to change the dynamics of how histones interact with DNA to help modulate regulatory activity. Specific histone modifications have been shown to demarcate active and repressed regions of transcriptional regulation [90]. For example, lysine residue 4 of histone H3 is often methylated either to mark enhancer (mono-methylation) or promoter (tri-methylation) regions [90]. Another hallmark of transcriptional regulation is that nucleosomes are evicted from active regulatory sites, making the underlying DNA more accessible to regulatory proteins [91].

Identification of regulatory elements has benefited from this improved knowledge of how the human genome encodes and organizes regulatory information and the recent development of high-throughput experimental techniques to exploit this knowledge [89]. Chromatin immunoprecipitation (ChIP) can identify genomic locations where specific proteins are bound to DNA, and when coupled with microarray hybridization (ChIP-chip)

or high-throughput sequencing (ChIP-seq), can be used to find these DNA-protein interactions on a genome-wide scale [92 93]. ChIP can also be performed to identify epigenetic information such as specifically modified histone residues [90 93]. Additional techniques such as DNase I hypersensitivity (DHS) [94] and Formaldehyde-assisted isolation of regulatory elements (FAIRE) [95] identify regions of the genome not bound to nucleosomes, and can both identify sites of transcriptional regulation on a genome-wide scale (DNase-seq [96], FAIRE-seq). The employment of these techniques across a series of tissue types, genotypes, and environmental conditions will uncover a denser set of regulatory elements to facilitate study of how transcriptional regulation is affected by variants influencing complex traits.

Computational tools have also been developed to identify functional non-coding elements, many by predicting the genomic locations of where transcriptional regulatory factors bind (TFBS). Transcription factors often bind sequences in degenerate patterns, making *in silico* binding site prediction difficult [97]. Information in a set of related sequences, for example known binding sites for a transcription factor, can be consolidated into frequency matrices of each base at each binding site position, and these binding site motifs can then be used to find similarly matching sequences that might represent novel binding sites [97]. Databases catalog motifs for a large number of transcription factors derived from literature of known binding sites (JASPAR [98] and TRANSFAC [99]) or studies that profile binding to oligonucleotide microarrays (UniPROBE [100]). Prediction of individual TFBS using motifs alone, however, has low sensitivity and specificity [101]. To increase predictive quality, methods have been developed that exploit genomic features of *in vivo* TFBS. First, the sequence surrounding

functional elements is often conserved between both between closely and distantly related species [102-106]. Second, TFBS are found in clusters, especially those for factors that are expressed in the same tissues [107-111]. Computational predictions of functional elements can be considered complimentary to functional genomics approaches, as the latter are restricted to tested experimental conditions and the technological limitations of the assay.

Chapter II

Computational selection of biologically-relevant candidate genes for complex traits

Abstract

Motivation: Identification of the genetic variation underlying complex traits is challenging. The wealth of information publicly available about the biology of complex traits and the function of individual genes permits the development of informatics-assisted methods for the selection of candidate genes for these traits.

Results: We have developed a computational system named CAESAR that ranks all annotated human genes as candidates for a complex trait by using ontologies to semantically map natural language descriptions of the trait with a variety of gene-centric information sources. In a test of its effectiveness, CAESAR successfully selected 7 out of 18 (39%) complex human trait susceptibility genes within the top 2% of ranked candidates genome-wide, a subset that represents roughly 1% of genes in the human genome and provides sufficient enrichment for an association study of several hundred human genes. This approach can be applied to any well-documented mono- or multi-factorial trait in any organism for which an annotated gene set exists.

Availability: CAESAR scripts and test data can be downloaded from http://visionlab.bio.unc.edu/caesar/

Introduction

Unlike Mendelian traits, in which a mutation in one gene is causative, or oligogenic traits, where several genes are sufficient but not necessary, complex traits are caused by variation in multiple genetic and environmental factors, none of which are sufficient to cause the trait [20]. The contribution of any given gene to a complex trait is usually modest. In addition, complex traits often encompass a variety of phenotypes and biological mechanisms, making it difficult to determine which genes to study [112].

As a result, traditional methods of genetic discovery, such as linkage analysis and positional cloning, while widely successful in identifying the genes for Mendelian traits, have had more limited success in identifying genes for complex traits. Candidate gene studies have had encouraging success, yet this approach requires an effective method for deciding *a priori* which genes have the greatest chance of influencing susceptibility to the trait [113]. Recent advances in genotyping technology have provided researchers with the ability to test association in hundreds of genes relatively quickly, and even the entire genome through a genome-wide association study. Genome-wide association studies are promising, yet not always economically feasible or statistically desirable [114]. Therefore, one of the greatest challenges in disease association study design remains the intelligent selection of candidate genes.

To this end, we have developed a computational methodology, named CAESAR (CAndidatE Search And Rank), that uses text and data mining to rank genes according to potential involvement in a complex trait. CAESAR exploits the knowledge of complex traits in literature by using ontologies to semantically map the trait information to gene and protein-centric information from several different public data sources, including tissue-specific gene expression, conserved protein domains, protein–protein interactions, metabolic pathways and the mutant phenotypes of homologous genes. CAESAR uses four possible methods of integration to combine the results of data searches into a prioritized candidate gene list. In effect, CAESAR mimics the steps a researcher would undertake in selecting candidate genes, albeit faster, potentially more thoroughly, and in a more quantitative manner.

CAESAR represents a novel selection strategy in that it combines text and data mining to associate genetic information with extracted trait knowledge in order to prioritize candidate genes. In contrast to a number of existing approaches [15-17] gene selection is not limited to one or more genomic regions, as all genes annotated in one of our databases are potential candidates. CAESAR is ultimately designed for traits in which the relevant biological processes may not be well understood and potentially hundreds of reasonable candidate genes exist.

The potential benefits to a researcher in adopting a computational approach to gene selection such as CAESAR include the ability to quickly and systematically process several hundred thousand biological annotations, many of which require highly specialized domain expertise to interpret. This benefit will continue to grow in importance

as the volume and technical detail of annotation data increases. Relevant gene annotations can easily escape human consideration due to biases that investigators bring to the task of prioritization and that are difficult to overcome even by conscious effort. This is particularly valuable for complex traits, which may be affected by a wider array of biological processes, some of which may not have been directly implicated by previous studies. CAESAR also reports the evidence supporting the prioritization rank of each gene, allowing an investigator to trace the line of reasoning and to exercise his or her own judgment as to its validity. Thus, it can be seen as a very sophisticated aid to manual prioritization.

Though designed to help with the design of an association study involving a few hundred genes, CAESAR can also be used to prioritize a smaller number of candidates within a region of linkage, or to prioritize among polymorphisms annotated with ranked genes that show significant association in a genome-wide study.

We have tested CAESAR on 18 susceptibility genes for 11 common complex traits in humans including type 1 and type 2 diabetes mellitus, schizophrenia, Parkinson's disease, cardiovascular disease, age-related macular degeneration, rheumatoid arthritis and celiac disease. Test genes were ranked higher than 95.7% of all ranked genes on average, and higher than 99.7% in the best case.

Methods

CAESAR is comprised of three main steps. First, previously implicated genes mentioned in the input text are identified and ontology terms are ranked based on their similarity to

an input text. Second, genes are ranked for each data source independently based on the relevance of the ontology terms with which they are annotated. Third, the individual gene lists are integrated to provide a single ranked list of candidate genes that combines evidence from all data sources. We refer to these three steps as text mining, data mining and data integration, respectively. The approach of CAESAR is presented as a schematic diagram in **Figure 2.1a**.

CAESAR requires a user-defined body of text (referred to as a corpus) as input. This text is ideally an authoritative and comprehensive source of biological knowledge about the trait of interest. If an online Mendelian inheritance in man (OMIM)[115] identifier is supplied, CAESAR will use the OMIM record as input. Alternately, the user can provide any other body of text, for instance one or more review articles.

Since the corpus is written in natural language, the information must be converted to machine-readable form. This is done in two ways. First, human gene symbols are identified within the corpus. If an OMIM record is used as input, gene identifiers can be extracted directly from the OMIM database. Otherwise, gene symbols are extracted by matching to a reference list. Genes are weighted based on frequency of occurrence in the corpus, f_g, where the weight c_g of extracted gene g is calculated as f_g divided by the sum of all f_g across n total extracted genes. The reference list of standard names, symbols, database identifiers and corresponding mouse homologs for each gene is compiled from Entrez Gene [116] and Ensembl [117]. The extracted genes are assumed to be relevant to the biology of the trait, but do not necessarily contribute to the genetic variation of the trait.

Second, the corpus is used to quantify the relevance of terms within several different biomedical ontologies. Four ontologies are used as part of CAESAR, the gene ontology biological process (GO bp) and molecular function (GO mf) [118], the mammalian phenotype ontology (MP) [119] and the eVOC anatomical ontology [120] (**Table 2.1**). Relevance is quantified using a similarity search under a vector-space model [121], as follows (**Figure 2.2**). For each ontology, the individual terms are split into separate documents containing the term name and term description if available. These documents together comprise a document database, or search space, against which the corpus is queried (**Figure 2.2a**). The corpus and each document are converted to vectors $v_i = < w_{i1}, w_{i2}, \ldots , w_{in} >$ with dimensionality equal to the size of the word space n, which is the total number of unique words in the document database. Commonly used stop words such as 'and' and 'the' are removed from the word space. Each element of the vector for document i is calculated as $w_{ij} = e_{ij}$, where e_{ij} is the number of occurrences of word j in the document.

The similarity of the corpus to each document is calculated as the cosine of the angle between the vectors, which is equal to the dot product of the vectors divided by the product of the magnitudes of the vectors. A larger cosine indicates vectors with greater similarity. Using this measure, ontology terms are weighted based on their similarity to the corpus (**Figure 2.2c**), where the weight c_t of term t is directly equal to the cosine.

Eight sources of gene-centric information are used to map ranked ontology terms to the genes annotated with them (**Figure 2.1b**). The resulting output is eight lists of gene scores, one for each functional category.

Mammalian phenotype ontology terms are used to query the mouse genome database (MGD) [122] for genes producing a given phenotype when mutated and to query the genetic association database (GAD) [123] for genes showing positive evidence of association with a phenotype in a human population. The eVOC anatomical ontology terms are used to query the UniProt database [124] for genes expressed in a given tissue. Gene ontology terms are used to query the gene ontology annotation database (GOA)[125] for genes annotated with a given gene ontology biological process or molecular function term. Finally, the extracted genes are used to query the biomolecular interaction network database (BIND) [126] and the human protein reference database (HPRD) [127] for genes encoding proteins that interact with the protein products of the extracted genes, query the Kyoto encyclopedia of genes and genomes (KEGG) pathway database [128] for other genes involved in the same human cellular pathways and query the InterPro protein domain database (IPro) [129] for genes sharing conserved protein domains with the extracted genes.

The user may also optionally input one or several genomic sequence regions to include genes in chromosomal regions implicated through genetic linkage as an additional list of genes (**Figure 2.1b**).

The score r_{ij} of gene i for source j is then calculated as either the maximum, sum or mean of the weights of the k matching ontology terms or extracted genes $c_1 \ldots c_k$. The three alternatives weigh the combined evidence for relevance in different ways, as described below for data integration from multiple sources.

The gene scores from the eight sources are integrated to produce one combined score for each gene. Integration is accomplished using one of four methods. Each method

represents a different approach that an investigator might choose when manually prioritizing candidate genes on the basis of evidence from several data sources.

The first three methods involve taking the maximum, sum or mean of the *z*-transformed r_{ij} scores for each gene. The maximum favors genes with strong evidence from one data source, the sum favors genes with evidence in many data sources and the mean favors genes with strong evidence only, penalizing genes with any weak evidence. The maximum, mean and sum are referred to as int1, int2 and int3, respectively. Transformed scores are calculated as $z_{ij} = (r_{ij} - x_j)/s_j$, where x_j is the mean and s_j the SD of the scores from source *j*. The combined score $\Phi_{.,i}$ is then obtained by calculating the maximum

$$\phi_{\text{int1},\,i} = \max z_{ij}$$

average

$$\phi_{\text{int2},\,i} = \sum_{j=0}^{n} z_{ij}/n$$

or sum

$$\phi_{\text{int3},\,i} = \sum_{j=0}^{n} z_{ij}$$

of the transformed scores for gene *i*.

The fourth method, referred to as int4, differs from the other three by considering both the score of a gene within a data source as well as the number of genes returned for that data source. First, a transformed score s_{ij} is obtained.

$$s_{ij} = \frac{r_{ij}}{\sum_{i=0}^{n} r_{ij}}$$

The transformed gene scores are then summed together to provide a final score for each gene.

$$\phi_{\text{int4},\,i} = \sum_{j=1}^{J} s_{ij} \frac{g_j}{G}$$

where g_j is the number of genes returned for source j and

$$G = \sum_{j=1}^{J} g_j$$

The CAESAR algorithms were written using Perl version 5.8.1 and Java version 1.4.2. The vector space similarity searches were performed using a modified version of the Perl module Search::VectorSpace by Maciej Ceglowski (http://www.perl.com/pub/a/2003/02/19/engine.html). Databases and ontology schemas were downloaded and parsed into XML under a custom XML schema. Intermediate text and data-mining results were also stored as XML under the same schema.

To assess the ability of CAESAR to choose valid candidates, 18 test genes were selected from recently published reports providing strong evidence of statistical association with

known complex human disorders. The test genes included *CTLA4* [130], *PTPN22* [131], *PTPN22* [132], *SUMO4* [133], *FCRL3* [134], *ENTH* [135], *EN2* [136], *TCF7L2* [47], *CFH* [137], *LOC387715* [138], *LTA4H* [139], *C2* [140], *CFB* [140], *NPSR1* [141], *MYO9B* [142], *IL2RA* [143], *SEMA5A* [144] and *LOC439999* [145].

Each disorder required a custom corpus, either an OMIM record or one or more review articles describing the biology of the disorder (**Table 2.2).** Review articles were selected by searching PubMed [146] for articles published before the year of discovery of each gene association. Where multiple suitable review articles were available, the texts were concatenated to produce the corpus. We removed any direct reference to the testing gene in the input text. In addition, entries in the GAD containing the test genes were removed. Thus, the input data closely mimicked the state of knowledge prior to the discovery of the positive association between the disease and the test gene.

In the case of age-related macular degeneration (ARMD), positive associations for the two test genes, *CFB* and *C*2, were reported after the discovery of *CFH* as a susceptibility gene for the disease. Due to the absence of a suitable review article incorporating the discovery of *CFH*, results for these two test genes employ the ARMD OMIM corpus only.

A common way of summarizing the performance of previous candidate gene selection algorithms is to calculate 'fold enrichment', which is the total number of ranked genes divided by the rank of the test gene. Fold enrichment must be interpreted with caution, because it is not calculated relative to random expectation. Nonetheless, we report this statistic in order to facilitate comparison with other methods.

Results

We tested the performance of the algorithm on a set of test genes previously reported to be associated with 11 complex human diseases (**Table 2.2**). For each disease, we selected one or more genes for which recent population genetic studies have reported a significant association with the disease phenotype. Nearly 15,000 genes had sufficient information from one or more data sources to be ranked. **Table 2.2** summarizes results of the 18 test genes by separately considering tests using review articles and OMIM records as input, although not all genes were tested using both input types. In order to report the success of CAESAR using all 18 genes, we combined review article tests for 16 genes with OMIM record tests for 2 genes, *CFB* and *C2*, which were not tested using review articles (see Methods section). The following results using all 18 test genes are thus not summarized in **Table 2.2.**

First, we evaluated the choice of data-mining method for determining the score r_{ij} of each gene i for each data source j (see Methods section). The distributions of the ranks are shown in **Figure 2.3a**. Each data-mining method used the int4 integration method (data for other integration methods not shown). The maximum method had a smaller median rank (549.5) than both the sum (1353) and mean (1020) methods.

Second, we evaluated the four different methods for the integration of data from different sources (**Figure 2.3b**). Int4 yielded the smallest median rank (549.5) compared to the results for int1 (max), int2 (mean) and int3 (sum), which were 1488, 2594 and 1201, respectively. Furthermore, int4 had smaller upper and lower quartile ranks than

int1, int2 and int3. We thus report the results for the maximum data-mining and int4 integration method in what follows.

Overall, 16 of 18 test genes were ranked with a median rank of 549.5 and 67-fold average enrichment. Seven of the 18 test genes (39%) were ranked higher than 98% of all ranked genes for the trait in question, while five (28%) ranked in the 99th percentile. The highest rank seen in our tests was 44 for *CFB*, a susceptibility gene for age-related macular degeneration, which corresponds to a 293-fold enrichment. Two of the genes, *LOC387715* and *LOC439999*, were unranked due to a lack of information on these genes in any of the data sources.

We compared the observed distribution of the ranks for the 18 test genes to that expected by chance, which is a minimal test for the effectiveness of the method. The expected mean percentile for a random gene would be 50. The observed mean percentile is 80.5 and, under a binomial expectation, the 95% confidence interval is 66–95. Thus, the observed distribution of ranks for the test genes is significantly displaced relative to random expectation.

We next examined the effect of the choice of corpus on the ranks for the test genes. Using review article corpus tests only, 14 of 16 test genes were ranked, with a median rank of 725 and 54-fold average enrichment. Six of the 16 test genes (37.5%) ranked in the 98th percentile, while four (25%) ranked in the 99th percentile (**Table 2.2**).

For comparison, we selected for each disease the relevant records from the OMIM database. For all tests the int4 method was used (**Table 2.1**). The test for candidate genes

of myocardial infarction was omitted because the OMIM record for this disease is only 100 words in length, which would be insufficient for reliably scoring a large number of ontology terms. Of the remaining 17 genes tested, 15 had sufficient information to be ranked. The median rank was 879 with an average 43-fold enrichment. The best performance was observed for *CFB*, with 293-fold enrichment. Three of the 17 test genes (17.6%) ranked in the 98th percentile of all ranked genes, while 2 of 17 (11.8%) ranked in the 99th percentile. Only one gene, *SEMA5A*, had a dramatically improved rank relative to that obtained using a corpus of published review articles. Thus, the ranks for the test genes using OMIM records, while still clearly an improvement over random expectation, are in most cases inferior to those obtained using review articles.

We examined whether the length of the input text could help explain the difference in performance between the two types of input text. The length of each corpus was measured as the number of words excluding stop words and non-word characters. There was no significant correlation between the length of the corpus and the rank obtained for each test gene (Spearman's rho = –0.21 , $P = 0.27$).

CAESAR is dependent on available annotations to rank genes. Therefore, the preferential ranking of well-annotated genes is a potential source of bias in the results. We addressed this issue in two ways, by measuring the effect of both breadth and depth of annotation on gene rank. We first measured the correlation between gene rank and the breadth of annotation, or the number of sources for which a gene is annotated, across each integration method. Using the default methods (max and int4), there is a strong correlation ($\rho = –0.75$), as shown in **Figure 2.4**. By comparison, again using the max

method, int2 ($\rho = -0.15$) and int3 ($\rho = -0.06$) showed little correlation, while int1 showed modest correlation ($\rho = -0.47$).

We next addressed the correlation between gene rank and annotation depth by considering the number of GO annotations (biological process + molecular function) per gene. For each data-mining method, and using int4 for data integration, we calculated the mean number of GO terms for genes ranked within the top 98th percentile (max: 7.2 ± 4.1; avg: 6.2 ± 3.7; sum: 9.8 ± 5.3) and found this to be significantly higher than the mean number of GO terms across all ranked genes (4.6 ± 2.9) for all three data methods (two-tailed, unpaired *t*-tests, *P*-values $<2 \times 10^{-16}$).

Data sources used by CAESAR include diverse available sources of gene-centric information; however, non-independence among data sources could also potentially bias the results. To address this issue, we measured the average correlation between the ranked gene lists for each tested trait using the review article corpus (**Table 2.3**). The majority of the sources show a mild, yet significant, correlation. No two data sources show a correlation greater than $\rho = 0.43$. Several pairs of sources show very weak negative correlations.

Discussion

The extraordinary amount of biological information available in the published literature and in publicly available databases about complex human diseases, on the one hand, and genes and their protein products, on the other, is well suited to the *in silico* identification of candidate genes for disease. The approach is enabled by ontologies that provide a

semantic mapping between the natural language description of diseases and traits, and the functional annotation of genes and their products. It is further enabled by the availability of well-curated pathway and protein-interaction datasets, and a wide variety of functional information about not only the genes themselves, but also their homologs in model organisms. The approach implemented in CAESAR can, in principle, be applied to any complex trait in any organism for which similar information resources exist.

CAESAR relies on human expert knowledge in order to function effectively, but it does not require that the user actually possess all of this knowledge. At a minimum, the user needs to select a relevant corpus, but much more user intervention is possible. The user may manually modify the scores from the text-mining step and/or introduce genes in addition to those that were extracted from the corpus. The final rankings may be modified based on user perceptions of the importance of particular data sources. The user may also restrict the algorithm to consider only certain genomic regions or particular sets of genes. While it is not advisable to eliminate human judgment and oversight of the candidate gene selection process, due to the volume and the complexity of the information involved, semi-automated methods such as CAESAR may well outperform an unaided expert. At the very least, CAESAR provides a quantitative starting point for which the assumptions are clear and the user's biases are minimized.

The success of CAESAR in any given instance is due both to factors that are, at least to some extent, under the user's control and those that are not. The user's choice of a corpus that accurately reflects the biology of the trait is clearly of critical importance. In our experiments, we found that review articles generally, though not always, yielded better

results than OMIM records. The explanation for this difference is not clear; it does not appear to be due to differences in corpus length.

Other factors under the user's control are algorithmic, e.g. how to calculate a score for a gene within a data source and to rank genes across multiple data sources. The variety of simple methods used here can, in some cases, lead to substantially different rankings. One example is *NPSR1*, which had ranks of 749 and 2751 using int1 and int2, respectively. Four different data sources (GO bp, GO mf, IPro and tissue) report information on *NPSR1*, and the scores vary from high to low. Int1, which calculates the maximum, favors genes with a high score in one data source regardless of the others, whereas the low scores are detrimental to the final rank using int2, which calculates the average. Each of the methods can be justified (see Method section), and it is not clear a priori which should be superior.

Overall, we found that the best results on the test set were obtained using a corpus of review articles, the maximum method for combining scores for a gene within a data source, and the int4 method for data integration across multiple sources. However, other combinations of parameters were superior for particular test genes. On the basis of our test results, we have selected the 'max' data-mining and 'int4' data-integration methods to be the default settings for CAESAR. The OMIM record, if available, is used as the input text by default, though our results suggest that one or more review articles should be used instead, or in addition, when possible.

A number of factors affecting CAESAR's success are outside of the user's control. One is the depth of biological knowledge about the complex trait under study and the extent to

which this knowledge has been recorded. Another is the extent to which ontologies can be used to mediate between trait-centric and gene-centric information sources. For example, anatomical ontologies are available for mammals, but not yet for all organisms. Even where an ontology exists, certain terms may not exist, have listed synonyms, or be sufficiently well defined.

The process of extracting gene names from unstructured text is also error-prone [147], especially when using older bodies of text containing outdated gene names and symbols. Gene extraction is complicated further by the fact that genes often share symbols with other genes and non-gene acronyms.

Perhaps most importantly, CAESAR depends on the availability of functional information. Approximately half of the unique entries in our reference set remained unranked for any trait due to lack of annotation, including two of the test genes, *LOC387715* and *LOC439999*. As the total number of ranked genes depends on the number of ontology terms that are mapped from the corpus, the success of CAESAR for a given trait depends on the information content of the corpus. One tested trait, myocardial infarction, did not have a sufficiently informative OMIM record. Therefore, CAESAR is limited to genes and traits for which there is sufficient information in the form of annotations and text descriptions, respectively. To the extent that this reflects incomplete knowledge of genes and traits, it is a limitation shared by all candidate gene approaches. The lack of gene-centric information, at least, can be partially overcome by including additional data sources from map-based studies, systematic functional genomic screens and other model systems in which homologs may have been characterized.

Given the importance of including a wide variety of functional information, CAESAR could be enhanced by the inclusion of additional data sources. A particularly valuable source would be data from transcription profiling experiments, which would provide information on a large proportion of genes that are lacking information from other sources. Inclusion of this data will be challenging, however, as the datasets available are diverse and heterogeneous, and it is not clear how best to score the relevance of a particular expression pattern to a trait.

Inclusion of additional data sources could potentially raise the issue of non-independence among them. Although no two data sources used in this study are highly correlated, most of them have a significant weak correlation. CAESAR does not currently correct for non-independence during the data-integration step.

A variety of *in silico* methods for candidate gene selection have previously been reported, though most have been designed and tested to prioritize positional candidates. Gene-Seeker [17] selected candidates in a given genomic region through web-based data mining of expression and phenotype databases. This approach enriched for disease genes in 10 monogenic disorders, providing at best 25- and 7-fold enrichment on average. POCUS [15] exploited functional similarities between genes at two or more loci to predict candidates, requiring no user input beyond the genomic regions of interest. It provided 12-, 29- and 42-fold enrichment on average for three test loci of increasing size and at best provided 81-fold enrichment. Perez-Iratxeta *et al.* (2002) used literature mining to associate pathology with GO terms and then used these terms to rank candidate genes. The authors created artificial loci containing an average of 300 genes for testing and

found 10-fold enrichment on average and, at best, 38-fold enrichment. The correct disease gene was present in their enriched set for 50% of the loci. Freudenberg and Propping (2002) computed similarity-based clusters of known disease genes based on phenotypic sharing between diseases. Their method selected the correct disease gene in roughly two-thirds of the cases, on average resulting in 10-fold enrichment, and in the top one-third of the cases resulting in 33-fold enrichment. Franke *et al*. (2006) developed a functional network of human genes to select candidate genes found in pathways with known disease genes. They constructed artificial loci that contained on average 100 genes, and found 20- and 10-fold enrichment on average in 27 and 34% of tested genes, respectively.

More recently, SUSPECTS [16] and ENDEAVOUR [14] have been developed for application to more complex traits. Both of these systems prioritized genes using a combination of annotation and sequence features based on similarity to a training set. SUSPECTS was able to identify a test gene in artificial loci on average within the top 13% of candidates, a 7-fold enrichment. In half the cases, the test gene was in the top 5% of candidates, a 20-fold enrichment. ENDEAVOUR tested both monogenic and polygenic (complex) disorders using a test set of 200 genes. Over all tested disorders, ENDEAVOUR provided 9-fold enrichment on average and 200-fold enrichment at best. Considering polygenic disorders only, ENDEAVOUR provided 5-fold enrichment on average and 18-fold enrichment at best.

The measure of success for an approach such as CAESAR ultimately depends on the specific application. Our goal has been the enrichment of candidates within the top 2% of ranked genes, which represents roughly the top 1% of genes in the human genome. Given

the number of functionally annotated human genes, this corresponds to 250–300 genes, which is a reasonable number to include in a high-resolution SNP association study for a complex disease in human populations. Our results suggest that approximately one-third to one-half of the genes previously associated with complex human disease would be included in this enriched candidate set. With a complex trait, for which the true effectors are only partially known, it is difficult to quantify the number of true and false positives. Nonetheless, assuming all genes outside of our test set are negatives, we can calculate sensitivity as TP/(TP+FN) and specificity as TN/(TN+FP), where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives and FN is the number of false negatives. Considering positives to be the top 2% of ranked genes, we obtained an overall sensitivity of 39% and specificity of 98% for our test set. Other measures of success may be relevant for different applications, such as prioritizing SNPs for follow-up work from a genome-wide association study. By standard measures, CAESAR compares favorably with other methods, even though we use a test set of genes associated with complex rather than monogenic or oligogenic diseases. The highest (293) and average (67) fold enrichment obtained with CAESAR are greater than those reported for other systems.

CAESAR makes use of a relatively small trait-specific corpus, comprised of one to several review articles, and a large body of gene-centric information. A similar approach could be useful for other applications involving semantic mediation between larger corpora or sets of corpora.

In conclusion, CAESAR can successfully mine large amounts of biological information to guide the selection of candidate genes for complex diseases in humans. Applications include selection of candidate genes for association or re-sequencing studies, prioritization of candidates for functional genomics experiments, or evaluation of results from linkage and genome-wide association studies. The approach may be extended to select candidates for complex traits in other organisms for which similar informatics resources are available. No computational system can select candidate genes with certainty; however, when used as a guide, CAESAR is a useful tool for candidate gene prioritization.

Contributions

A version of this work has been published previously as: Kyle J Gaulton, Karen L Mohlke, Todd J Vision. A computational system to select candidate genes for complex human traits. *Bioinformatics.* 2007 May 1;23(9):1132-40.

KJG, KLM, TJV designed the study and wrote the manuscript. KJG wrote scripts and performed data analysis.

Chapter III

Comprehensive association study of type 2 diabetes and related quantitative traits with common variation in 222 candidate genes

Abstract

OBJECTIVE: Type 2 diabetes is a common complex disorder with environmental and genetic components. We used a candidate gene-based approach to identify single nucleotide polymorphism (SNP) variants in 222 candidate genes that influence susceptibility to type 2 diabetes**. RESEARCH DESIGN AND METHODS**: In a case-control study of 1,161 type 2 diabetic subjects and 1,174 control Finns who are normal glucose tolerant, we genotyped 3,531 tagSNPs and annotation-based SNPs and imputed an additional 7,498 SNPs, providing 99.9% coverage of common HapMap variants in the 222 candidate genes. Selected SNPs were genotyped in an additional 1,211 type 2 diabetic case subjects and 1,259 control subjects who are normal glucose tolerant, also from Finland. **RESULTS**: Using SNP- and gene-based analysis methods, we replicated previously reported SNP-type 2 diabetes associations in PPARG, KCNJ11, and SLC2A2; identified significant SNPs in genes with previously reported associations (ENPP1 [rs2021966, P = 0.00026] and NRF1 [rs1882095, P = 0.00096]); and implicated novel genes, including RAPGEF1 (rs4740283, P = 0.00013) and TP53 (rs1042522, Arg72Pro, P = 0.00086), in type 2 diabetes susceptibility. **CONCLUSIONS**: Our study provides an effective gene-based approach to association study design and analysis. One or more of the newly implicated genes may contribute to type 2 diabetes pathogenesis. Analysis of

additional samples will be necessary to determine their effect on susceptibility.

Introduction

Type 2 Diabetes (T2D) is a metabolic disorder characterized by insulin resistance and pancreatic β-cell dysfunction, and is a leading cause of morbidity and mortality in the USA and worldwide. The incidence of T2D is rapidly increasing with 1.5 million new cases documented in the United States in 2005 alone, and the number of affected individuals worldwide is expected to double in the next 50 years [148]. While environmental factors play a major role in predisposition to T2D, substantial evidence supports the influence of genetic factors on disease susceptibility. For example, the twin concordance rate is an estimated 34% for monozygotic twins and 16% for dizygotic twins [38]. However, the underlying genetic variants are just beginning to be identified [149] .

Numerous published reports have identified association between T2D and common genetic variants in human populations [79,150,151]; however, until very recently, variants in only a few genes have been consistently replicated across populations and with large sample sizes. Among these are the Pro12Ala (rs1801282) variant in peroxisome proliferator-activated receptor gamma (*PPARG*) [152], the Glu23Lys (rs5210) variant in the potassium channel gene *KCNJ11* [49], and several variants in the Wnt-receptor signaling pathway member *TCF7L2* [153].

Recent genome-wide studies have implicated many previously unreported genes in T2D susceptibility. The first reported genome-wide association (GWA) scan implicated

variants at five susceptibility loci that include *TCF7L2,* and novel loci near the genes *SLC30A8, IDE-KIF11-HHEX, LOC387761,* and *EXT-ALX4* [53]. Three companion GWA studies, including one by our group, replicated evidence for *PPARG, KCNJ11, TCF7L2, SLC30A8, IDE-KIF11-HHEX,* and provided new evidence for *CDKAL1, CDKN2A-CDKN2B, IGF2BP2, FTO*, and a region of chromosome 11 with no annotated genes [54-56]. Additional GWA studies [57-62] provided additional evidence for *TCF7L2*, *CDKAL1* and *SLC30A8*. The candidate genes *WFS1* [50] and *TCF2* [51,52] have also been confirmed in large samples, bringing the current list of T2D susceptibility loci to at least 10.

The recent discovery of these loci still explains only a small fraction (estimated 2.3%) of the overall risk of T2D [54]. Therefore, novel susceptibility genes remain to be identified through increasingly comprehensive analyses of both individual genes and the entire genome.

The Finland-United States Investigation of NIDDM genetics (FUSION) study aims to identify variants influencing susceptibility to T2D and related quantitative traits in the Finnish population [75]. FUSION has previously identified modest T2D association in Finns with variants in *HNF4A* [78], four genes known to cause MODY [79 78 80], *PPARG, KCNJ11*, *ENPP1, SLC2A2, PCK1, TNF, IL6* [79], and *TCF7L2* [81], in addition to the loci identified in the GWA studies.

As a complementary approach to GWA studies, which are conducted without *a priori* biological hypotheses, we sought to perform an in depth analysis of >200 genes likely to

influence susceptibility to T2D and quantitative trait variation that we selected by applying CAESAR (CandidAtE Search And Rank), a text and data-mining algorithm [154]. We aimed to analyze the full spectrum of HapMap-based common variation in each of these candidate genes. The combination of high-throughput genotyping, linkage disequilibrium (LD) information from HapMap [21], the ability to impute un-genotyped variants [27], and the improved functional annotation of the genome make possible in-depth candidate gene based association analysis.

Methods

Sample selection

The Stage 1 sample set consisted of 2,335 Finnish individuals from the FUSION [75] and Finrisk 2002 [155 156] studies (**Table 3.1, Table 3.6**). The sample included 1,161 individuals with T2D and 1,174 normal glucose tolerant (NGT) controls. Diabetes was defined according to 1999 World Health Organization criteria of fasting plasma glucose concentration ≥ 7.0 mmol/L or 2-h plasma glucose concentration ≥ 11.1 mmol/L, by report of diabetes medication use, or based on medical record review. Normal glucose tolerance was defined as having fasting glucose < 6.1 mmol/L and 2-h glucose < 7.8 mmol/L. 120 FUSION offspring with genotyped parents were included for quantitative trait analysis; all offspring were NGT except one T2D individual who was included in the case sample.

Stage 2 consisted of 2,473 Finnish individuals (**Table 3.1, Table 3.7**), and included 1,215 individuals with T2D and 1,258 NGT controls [55]. 56 duplicate samples were used for

quality control.

The sample sets are identical to those used in the FUSION GWA study [55]. Study protocols were approved by local ethics committees and/or institutional review boards, and informed consent was obtained from all study participants.

Gene selection

A total of 222 candidate genes were selected for study using two strategies. 217 candidate genes were selected using CAESAR, an algorithm that prioritizes candidate genes for complex human traits [154]. CAESAR prioritizes candidate genes for complex human traits by semantically mapping trait relevant natural language descriptions to a variety of functional annotation sources. The trait relevant input text used here was four concatenated T2D review articles [157-160]. Given a trait relevant input text, CAESAR uses text- and data-mining to extract information from the input text in two ways.

First, terms in four biomedical ontologies (gene ontology biological process and gene ontology molecular function [161], eVOC anatomy [162], and mammalian phenotype ontology [163]) were ranked based on vector-space similarity to the input text [121]. For each ontology term, the term, including its description, and the input text were represented as two separate word vectors excluding common stop words, and the similarity of the word frequency in the two vectors was measured as the cosine of the angle between the vectors. A stronger similarity results in a higher rank. The ranked ontology terms were then used to query four functional annotation databases for genes annotated with the

ranked terms. We queried the gene ontology annotation (GOA) database [164], UniProt [124], the mouse genome database (MGD) [122], the genetic association database (GAD) [123] to create four lists of annotated genes.

Second, we independently compiled a list of genes of interest. Unlike the procedure described in CEASAR, in which the input text is mined for gene symbols, we empirically chose seven genes related to T2D. *PPARG, HNF4A,* and *KCNJ11* were selected based on prior evidence of T2D association in FUSION and other studies [78-80], and *PPARGC1A, PPARGC1B, ESRRA,* and *GABPA* were selected based on evidence of T2D-relevant transcriptional regulation [165-167]. These seven genes were then used to query functional databases for genes sharing annotations with the extracted genes. We used InterPro [129], the Kyoto encyclopedia of genes and genomes (KEGG) pathway database [128], and combined data from the interaction databases biomolecular interaction network database (BIND) [126] and the human protein reference database (HPRD) [127] to create three lists of annotated genes. Each extracted gene was given the same weight.

We added two gene lists to the seven described above: (a) genes located on chromosome 10 from ~120-130 cM, a region implicated by a meta-analysis of linkage genome scans [168] and (b) genes with evidence of *cis*-acting variation affecting gene expression [169] .

For each list, annotated genes are weighted based on calculating the maximum, sum, or average of the matching terms for that gene. For the MGD and *cis*-acting variant gene lists, the 'sum' of term scores was used to favor genes with evidence from several

ontology terms. For the remaining seven gene lists scores were generated by taking the 'maximum' term score for each gene to favor genes with strong evidence from one ontology term.

For the integration step, we weighted each of the nine resulting lists by the strength of their biological evidence; MGD, GO biological process, and UniProt data were considered to have the strongest biological evidence and were weighted highest. The weighted scores across all lists were then summed together for each gene to provide a final score for that gene. In total, 10,760 genes with annotation evidence were ranked and thus prioritized for relevance as T2D candidate genes. Genes were chosen from the top of the prioritized list, except 22 high-ranking genes were excluded based partially on strong negative evidence of association from previous reports, an absence of known SNPs in the gene, or numerous SNPs in the gene. The last choice was made before imputation techniques were available and exclusion of 11 genes with >20 SNPs allowed inclusion of additional lower-ranked genes that required less additional tagSNP genotyping. The genes excluded for this criterion were *CACNA1D*, *CACNA1E*, *CACNA1C*, *RFX3*, *NRG1*, *SHC3*, *RARB*, *PRKCE*, *PFKP*, *SNAP25*, and *RORA*.

Five genes were not ranked high enough to have been included using CAESAR. *ENPP1, HFE, WFS1*, and *ZNHIT3* were included because each had one or more SNPs associated with T2D ($p < .1$) in prior study of a subset of FUSION samples [79] (and unpublished data); in addition, *ENPP1* and *WFS1* had been previously studied as T2D candidate genes. *CAPN10* was included because it had been previously studied by FUSION [170] and

others [171 172].

SNP selection

We defined the 'transcribed region' of each of the 222 candidate genes as the sequence including the first exon of any transcribed isoform through the last exon of any transcribed isoform, and we aimed to capture variation up to 10 kb upstream and 5 kb downstream of the transcribed region (-10kb/+5kb). In this process we allowed SNPs to be located as far as 50 kb upstream and 50 kb downstream (-50kb/+50kb) of the transcribed region if they tagged a -10kb/+5kb SNP at r^2>.8.

We initially identified 2,312 SNPs from the Illumina Infinium™ II HumanHap300 BeadChip. A SNP was included if it was located within (a) the -10kb/+5kb region or (b) the 50 kb upstream and 50 kb downstream of the transcribed region of a candidate gene (-50kb/+50kb) and it tagged one or more HapMap SNPs within the -10kb/+5kb region at an r^2 threshold of .8 based on HapMap CEU genotypes. We previously demonstrated that the HapMap CEU data are a sufficient resource to select tagSNPs for the Finnish population [173].

To more comprehensively evaluate each gene, we selected additional SNPs for genotyping on an Illumina GoldenGate panel. 1,405 HapMap tagSNPs not present on the HumanHap300 BeadChip were selected to tag additional HapMap SNPs in each -10kb/+5kb region at an r^2 threshold of .8. We used a tiered selection process to select tagSNPs that were prioritized based on proximity to a candidate gene and functional

annotation. Annotation categories were non-synonymous variants, variants close to exon/intron boundaries, and variants in 8-species conserved regions (multiz8way) [174]. We selected 122 additional non-synonymous and potential splice-site SNPs with dbSNP minor allele frequency (MAF) >.05 that were not present in HapMap.

When HapMap release 21 became available, we re-selected tagSNPs from the HumanHap300 BeadChip by first identifying 10,762 common (MAF > .05) HapMap SNPs present within the -10kb/+5kb gene regions. For each SNP in the -10kb/+5kb gene region we included the best tag from the -50kb/+50kb gene region; the best tag for a genotyped SNP was itself. 3,428 SNPs were identified in this manner, including all successfully genotyped tagSNPs not on the HumanHap300 BeadChip. We also included eight SNPs that had been previously genotyped in candidate gene studies on a smaller subset of FUSION samples [79].

All reported SNP and gene positions are based on NCBI Build 35 (hg17).

Genotyping

317,503 SNPs were genotyped at the Center for Inherited Disease Research (CIDR) on the HumanHap300 BeadChip using the Illumina Infinium™ II assay protocol [55]. 99.7% (2,585 / 2,592) of the genotyped samples were successful at a call rate > 97.5%. Genotypes were obtained for 99.4% (315,635 / 317,503) of the SNPs and for the 317,503 SNPs there was a genotype consistency rate of 99.994% based on 24,990,942 duplicate genotype pairs (79 duplicate samples).

1,527 SNPs were genotyped in partnership with the Mammalian Genotyping Core at the University of North Carolina using the Illumina GoldenGate assay. 99.7% (2,586 / 2,592) of the genotyped samples were successful at a call rate > 97.5%, and genotypes were obtained for 97.2% (1,484 / 1,527) of the SNPs. GoldenGate SNPs were manually clustered, and in this process 40 of the 79 duplicate samples were used to assist and improve clustering; the remaining 39 duplicate samples were used to estimate the error rate. For the 1,527 SNPs there was a genotype consistency rate of 99.979% based on 61,905 duplicate genotype pairs.

Eight SNPs that were previously genotyped on a subset of our Stage 1 samples [79] using the Illumina GoldenGate assay were genotyped on the remaining samples using the Sequenom homogeneous MassEXTEND (hME) assay. Genotypes were obtained for all eight SNPs. Between the two platforms there was a genotype consistency rate of 99.889% based on 904 duplicate genotype pairs. Four SNPs were genotyped to validate imputed p-values; three were genotyped using Applied Biosystems TaqMan allelic discrimination assays and the fourth (rs2021966) using the hME assay. Genotypes were obtained for all four SNPs and there was a genotype consistency rate of 100% based on 295 duplicate genotype pairs.

31 SNPs were genotyped on Stage 2 samples using the hME assay. Genotypes were obtained for 29 of the SNPs and there was a genotype consistency rate of 100% based on 1,518 duplicate genotype pairs.

We applied quality control criteria to exclude and flag SNPs. SNPs were excluded from analysis if they (a) failed a test for Hardy-Weinberg equilibrium (HWE) at a p-value threshold of $1x10^{-6}$ using unrelated subjects, (b) had four or more duplicate errors or non-Mendelian inheritance (NMI) errors, or (c) had a sample success rate < 90%. Stage 2 data for two SNPs was included even though they had sample success rates < 90% (87.8% for rs858341 and 85.0% for rs4843165). Flagged SNPs were not removed from analysis, but were carefully examined if found interesting. A SNP was flagged if it (a) failed HWE at a threshold of $1x10^{-3}$, (b) had two or more duplicate or NMI errors, (c) had a sample success < 95%, or (d) had atypical clustering patterns. Twenty-two non-polymorphic SNPs were excluded from analysis. Two SNPs overlapped between GWA and UNC genotyping; thus, a total 3,531 genotyped SNPs were included in analysis.

Imputation

We used MACH, a computationally efficient hidden Markov model based algorithm (29, 38) to impute genotypes in FUSION samples for 7,498 common (MAF > .05) HapMap SNPs present in the target regions but not genotyped in our study. MACH combines our genotype data with phased chromosomes for the HapMap CEU samples and then infers the unknown FUSION genotypes probabilistically by searching for similar stretches of flanking haplotype in the HapMap CEU reference sample [27]. In this process, we used the GWA and Golden Gate genotype data from SNPs in the extended gene regions (-50 kb/+50 kb). For each individual at each imputed SNP, we calculated an expected allele count based on the average of allele counts for 90 iterations of the imputation algorithm. We assessed the quality of the results for each SNP by calculating the ratio of the

observed variance of scores across samples to the expected variance given the imputed allele frequency of the SNP (estimated r^2). SNPs with an estimated $r^2 \leq .3$ were excluded from further analysis. To improve the quality of imputation near the ends of the target regions, we used at least 1Mb of flanking genotype information to impute SNPs in target regions.

Coverage of HapMap SNPs

Coverage was calculated as the percentage of all common (MAF > .05) HapMap Release 21 CEU SNPs in the -10kb/+5kb gene regions that are tagged by a genotyped SNP at an r^2 threshold of at least .8.

T2D association analysis

Genotyped SNPs were tested for T2D association using logistic regression under additive (p_{add}), dominant, and recessive genetic models with adjustment for 5-year age category, sex, and birth province. Imputed SNPs were tested for T2D association using logistic regression under an additive model (p_{impute}) with the expected allele count in place of the allele count and adjusted for the same covariates. This approach takes into account the degree of uncertainty of genotype imputation in a computationally efficient manner by replacing allele counts (0, 1, 2) at the marker locus by predicted allele counts based on estimated probabilities of 0, 1, or 2 copies of a SNP allele [27].

We accounted for carrying out multiple correlated tests using the p-value Adjusted for Correlated Tests (p_{ACT}) method [176]. The p_{ACT} method was used to correct the minimum p-

value among (a) tests of three genetic models for a single SNP (p_{SNP}) and (b) multiple SNPs and models across a gene region (p_{gene}).

To evaluate each genotyped SNP, we adjusted the minimum p-value across the three tested models using p_{ACT} and termed the adjusted p-value p_{SNP}. We performed permutation testing to verify the p_{SNP} results. We empirically determined the experiment-wide significance of p_{SNP} by permuting case/control status 1,000 times. We calculated the power of our experiment to detect T2D association of the *TCF7L2* SNP rs7903146 in the Stage 1 or Stage 1+2 samples based on an experiment-wide p-value = 6.3×10^{-5}, disease prevalence =.1, risk allele frequency =.18, and OR=1.37, based on Table 1 of Scott et al. [55], using CaTS [177].

For each gene region we identified the minimum T2D association p-value for all models and SNPs and adjusted this p-value for the multiple tests using p_{ACT} and termed the adjusted p-value p_{gene}. Six pairs of the 222 candidate genes had adjacent or overlapping gene regions and were thus combined for the gene analysis: *C3-TRIP10*, *CHUK-PKD2L1, KCNJ11-ABCC8*, *LTA-TNF, NR1H3-SPI1,* and *NR5A1-NR6A1*. We estimated the study-wide significance of an observed number of significant SNPs by comparing to the appropriate binomial distribution and using a one-sided test of significance. To remove potential bias from the test for excess significance, two sets of genes were separately excluded: (a) seven genes with SNP(s) showing prior evidence of association in FUSION samples: *ENPP1, IL6, KCNJ11, PCK1, PPARG, SLC2A2, TNF*; and (b) five genes not selected by CAESAR: *CAPN10, ENPP1, HFE, WFS1, ZNHIT3*.

The genomic control λ values were 1.03 for genotyped SNPs (p_{add}) and 1.04 for imputed SNPs (p_{impute}) (**Figure 3.1**) [69].

We determined the independence of significant association signals in genes by including one SNP as a covariate in logistic regression and reassessing the evidence for association with the other SNPs.

Quantitative trait analysis

We tested all genotyped and imputed SNPs for association with 20 T2D-related quantitative traits including, in controls only: fasting insulin, fasting glucose, homeostasis model adjustment, and fasting free fatty acid; and in all samples: body-mass index, weight, waist circumference, hip circumference, waist to hip ratio, waist to height^2 ratio, total cholesterol, HDL cholesterol, LDL cholesterol, triglyceride level, cholesterol to HDL ratio, triglyceride to HDL ratio, diastolic blood pressure, systolic blood pressure, pulse, and pulse pressure.

For cases and controls separately, we regressed the quantitative trait variables on age, age^2, sex, birth province, and study indicator, and transformed the residuals of each quantitative trait to approximate normality using inverse normal scores, which involves ranking the residual values and then converting these to z-scores according to quantiles of the standard normal distribution. We then carried out association analysis on the residuals. To allow for relatedness, regression coefficients were estimated in the context

of a variance component model that also accounted for background polygenic effects (40). For genotyped SNPs we tested for association using the residuals under an additive model. For imputed SNPs we tested for association using the residuals and the expected allele count in place of the allele count under an additive model. For traits analyzed in both cases and controls, results were combined using meta-analysis in which we calculated a z-statistic summarizing the p-value, in its magnitude, and the direction of effect, in its sign. We then calculated an overall z-statistic as a weighted average of the T2D and NGT statistics and calculated the corresponding p-value. Weights were proportional to the square-root of the number of individuals in each sample and were selected such that the squared weights summed to 1.

Results

We studied 222 candidate genes for T2D association in our Stage 1 sample of 1,161 T2D cases and 1,174 NGT controls from the FUSION study (**Table 3.1**). Of 10,762 target HapMap SNPs (MAF > .05) in the –10 kb/+5 kb gene regions, the 3,531 genotyped SNPs cover 10,299 (95.7%) SNPs at an r^2 threshold of .8. This represents an improvement over the genome-wide HumanHap300 genotyped SNPs, which alone cover 79.0% of the target SNPs at $r^2 \geq .8$ (**Table 3.2**). 3,187 of the 3,531 genotyped SNPs are located in the -10kb/+5kb regions. Of the remaining 7,575 ungenotyped target SNPs, 7,498 were successfully imputed. Altogether, 99.9% of all target variation was genotyped, imputed, or tagged ($r^2 \geq .8$) by an analyzed SNP.

We evaluated the significance of genotyped SNPs in each gene region after correcting for multiple SNPs tested while accounting for the LD between SNPs, designated p_{gene} [176]. Given six pairs of adjacent genes (see Methods), we analyzed 216 distinct gene regions for T2D association (**Table 3.8**). SNPs in four gene regions were significantly associated with T2D at $p_{gene} < .005$: rs11183212 in *ARID2* ($p_{gene} = .0029$), rs2235718 in *FOXC1* ($p_{gene} = .0028$), rs8069976 in *SOCS3* ($p_{gene} = .0037$), and rs222852 in *SLC2A4* ($p_{gene} = .0024$), although no p_{gene} result reached study-wide significance of .00023, a threshold determined using a Bonferroni correction. SNPs in 19 genes were significant at $p_{gene} < .05$, including SNPs in three genes previously implicated in T2D susceptibility in FUSION (6) (**Table 3.3**). There was an excess of significant p_{gene} results at both thresholds: four at $p_{gene} < .005$ (p = .024); 19 at $p_{gene} < .05$ (p = .013). The excess of significant results at $p_{gene} < .005$ is maintained after excluding (a) seven genes showing prior evidence of association with any SNP in FUSION samples (p = .022) or (b) five genes not selected by CAESAR (p = .022), as no excluded genes were significant at that threshold (see Methods).

To evaluate all 3,531 genotyped SNPs (**Table 3.9**), we permuted the case/control status to estimate whether an excess of significant results was observed. 214 SNPs showed significant T2D association at a p_{SNP} threshold of .05, and of these, 26 were associated at a p_{SNP} threshold of .005 (**Table 3.4**); there was modest but not significant excess at both of these p_{SNP} thresholds (observed=214, expected=183.3, p = .09; observed=26, expected=18.9, p = .12, respectively). The most significant p_{SNP} value of 3.6×10^{-4} was observed for rs11183212, an intronic SNP in the *ARID2* gene, but when compared to an

empirical distribution of the most significant p-values this SNP does not reach a study-wide significance threshold of 6.3×10^{-5} based on 1,000 permutations. In the combined Stage 1 and Stage 2 sample, we have >99% power (80% in Stage 1 alone) to detect the most strongly associated previously observed T2D SNP, rs7903146 in *TCF7L2* [53-56], at a study-wide significance level, and substantially less power to detect T2D-associated SNPs with smaller effect sizes.

Nineteen of the 216 gene regions have at least one SNP significantly associated with T2D at $p_{SNP} < .005$; among these, Pro12Ala (rs1801282) in *PPARG* ($p_{SNP} = .0025$) was the only SNP that matched or was in high LD ($r^2 \geq .8$) with a previously reported variant, given the available HapMap LD information. Imputation identified 421 additional SNPs in 59 genes significantly associated with T2D ($p_{impute} < .05$, **Table 3.9**), including SNPs in 10 genes that did not contain a significant genotyped SNP ($p_{SNP} > .05$). We genotyped four of these initially imputed SNPs that were both significantly associated with T2D ($p_{impute} < .05$) and for which the imputation-based p-value was at least five times more significant than that for any nearby genotyped SNP; three of the four SNPs had highly concordant imputed and genotyped p-values (**Table 3.11**).

We selected for follow-up genotyping in Stage 2 samples 24 SNPs that were either significant at $p_{SNP} < .005$ or, if a non-synonymous variant, significant at $p_{SNP} < .01$ (**Table 3.1**). The most significant SNPs in the combined Stage 1 and Stage 2 samples were rs4740283 in *RAPGEF1 (*$p_{SNP} = .00013$*)*, rs2021966 in *ENPP1* ($p_{SNP} = .00026$*)*, Arg72Pro (rs1042522) in *TP53* ($p_{SNP} = .00086$), and rs1882095 in *NRF1* ($p_{SNP} = .00096$). In total,

16 SNPs were significant at $p_{SNP} < .05$ in the combined Stage 1 and Stage 2 samples (**Table 3.4**).

To evaluate the effect of body-mass index (BMI), we included BMI as an additional covariate in the analysis of the additive model for all genotyped and imputed SNPs. Of the 11 SNPs originally significant at $p_{add}<.001$, all were within a 10-fold difference after correction (**Table 3.12**). In addition, of the 16 SNPs significant at $p_{add}<.001$ after correction, three were more than 10-fold different than before correction, including several SNPs at the *TRIP10*/*C3* locus (**Table 3.13**).

Four genotyped and 30 imputed SNPs were strongly associated ($p < .0001$) with one or more of 20 quantitative traits after combining case and control subjects by meta-analysis (see Methods) (**Table 3.5, Table 3.14**). Variants in *APOE* and *PPARA* showed strong evidence of association with serum lipid levels, confirming previous reports [178,179]. Strong novel associations ($p<1x10^{-5}$) were observed for rs4912407 in *PRKAA2* with triglyceride level ($p = 3.68x10^{-6}$), rs10517844 in *CPE* with HDL level ($p = 2.07x10^{-5}$), and rs4689388 in *WFS1* with LDL level ($p = 5.30x10^{-5}$). We followed-up genotyped SNPs significantly associated ($p < .0001$) with one or more quantitative traits by genotyping the Stage 2 samples. No SNP showed study-wide significance in the combined Stage 1 and Stage 2 samples (**Table 3.5**).

Discussion

In this study we evaluated the evidence for T2D association for SNPs in 222 candidate

genes and provide a framework for thorough analysis of association of common variation to disease using gene-based functional annotation, HapMap LD information, and imputation of genotypes. This framework could be used in the context of a genome-wide association study or an independent investigation of candidate genes. We replicated previous T2D association with SNPs in *PPARG, KCNJ11,* and *SLC2A2;* identified significant SNPs in genes previously implicated in T2D risk, *NRF1* and *ENPP1*; and identified additional genes that may influence susceptibility to T2D and related quantitative traits including *RAPGEF1* and *TP53*. While some of the genes may be significant by chance, one or more may represent true susceptibility genes. We expect true susceptibility genes identified in our sample set will in many cases be shared in additional populations, as the FUSION GWA study identified many of the same risk alleles as other GWA studies on European populations [53-56,62].

To assess the role of the 222 genes in susceptibility to T2D, we attempted to assess complete coverage of common (MAF > .05) SNPs in the HapMap CEU database. The coverage of common HapMap CEU SNPs across all 222 candidate genes using genotyped SNPs was 95.7%, a 16.7% percent improvement over the coverage of 79.0% based on the Illumina HumanHap300 genome-wide panel (**Table 3.2**). HapMap provides excellent coverage of common variation in European samples; however, there are additional non-HapMap SNPs in these gene regions [21]. Of the 122 genotyped SNPs not in HapMap, 10 were not tagged at an r^2 threshold of .8 by a HapMap SNP, indicating that some of the non-HapMap variation is better covered in our study than the GWA panel.

Our most strongly T2D-associated SNP in the Stage 1 and Stage 2 samples was SNP rs4740283 (p_{SNP} = .00013), located 4 kb downstream of Rap guanine nucleotide exchange factor 1 (*RAPGEF1*). *RAPGEF1* is a ubiquitously expressed gene involved in insulin signaling [180] and Ras-mediated tumor suppression [181]. rs4740283 is in strong LD with SNPs in the coding region, and may affect either a regulatory element or protein function. Variation in this gene may contribute to susceptibility through reduced ability of peripheral tissues to absorb glucose in response to insulin.

The second strongest associated SNP in the Stage 1 and Stage 2 samples was Arg72Pro in *TP53* (rs1042522, p_{SNP} = .00086), originally identified by imputation, subsequently genotyped, and not well tagged by any originally genotyped SNP (maximum r^2 = .27 with rs2909430). *TP53* encodes the tumor-suppressor protein p53, and the Arg72Pro variant has a functional role in the efficiency of p53 in inducing apoptosis, possibly through reduced localization to the mitochondria [182]. The risk allele Arg72 has higher apoptotic potential consistent with a possible link between increased pancreatic beta-cell apoptosis, impaired insulin secretion and T2D.

We observed significant association with SNPs in two genes previously implicated in T2D susceptibility, nuclear respiratory factor 1 (*NRF1*) and the insulin-dependent facilitated glucose transporter *SLC2A2*. *NRF1* helps regulate mitochondrial transcription and oxidative phosphorylation [167], which has a known role in insulin resistance, and the associated *NRF1* variant, rs1882095, is located 1 kb downstream of the gene and not in modest LD (r^2>.6) with any HapMap SNP. In *SLC2A2* we found supporting evidence in

Stage 1 for the non-synonymous variant Thr110Ile (rs5400) (p_{SNP} = .0065), as well as a previously unreported variant, rs10513684 (p_{SNP} = .0046). The rs10513684 signal became slightly more significant after Stage 2 genotyping (p_{SNP} = .0023); however, the signal was attenuated (p =.18) after inclusion of Thr110Ile in the analysis.

Among the most significant T2D associated SNPs is rs2021966 in *ENPP1* (p_{SNP} = .00026). SNPs in high LD with rs2021966 are located in intron 1, in a region of strong multi-species conservation containing a pseudogene but no known transcripts. Previous studies of *ENPP1* have reported associations with rs1044498 and with a related three-SNP haplotype (rs1044498, rs1799774, rs7754561) and support a modest role in T2D susceptibility, possibly acting through obesity [183]. In our study, rs1044498 (p_{SNP} = .16) and rs7754859 (p_{SNP} = .18, r^2=1 with rs7754561) were not significantly associated with T2D (rs1799774 was not tested). The newly identified variants are in very low LD with rs1044498 (r^2 < .05).

Although we observed significant quantitative trait associations in previously implicated genes (*APOE* and *PPARA* with serum lipid levels), no quantitative trait associations became more significant after addition of Stage 2 samples (**Table 3.5**). This is likely due in part to the small number of SNPs selected for follow-up. Stage 2 genotyping of SNPs less significant in Stage 1 samples will be necessary to establish whether any novel SNPs contribute to quantitative trait variability.

In any gene-based study, the definition of gene boundaries is critical but by necessity somewhat arbitrary. We defined a gene region as 10 kb upstream of the first known exon

through 5 kb downstream of the last known exon in an attempt to capture the majority of nearby regulatory elements influencing a gene. Regulatory elements, however, can often be found up to several hundred kb away from a gene [184]. We evaluated whether a broader definition of a gene had a substantial effect on the p_{gene} results by testing extended gene regions, 50 kb upstream and 50 kb downstream of transcribed regions, by including HumanHap300 SNPs from these regions in our analysis. Using the extended gene boundaries, the insulin gene *INS* would be the most significant gene in our study (p_{gene} = .0019), driven by SNP rs10743152 (p_{SNP} = .00015) located 13 kb upstream of the first exon. Other genes that had significant SNPs (p_{gene} < .05) only in the extended gene region were *MAP2K1, CDK4,* and *IRF4*.

Even using the narrow gene boundaries, several SNPs in our study may influence expression or function of other nearby or even more distant genes. Recent genome-wide association studies have confirmed novel susceptibility variants downstream of *HHEX,* a gene selected for this study by CAESAR [53-56]; the reported SNPs are located outside of the narrow gene region (-10 kb/+5 kb) in a large LD block that includes *KIF11* and *IDE*, and we only detected nominal significance in the narrow *HHEX* region (p_{SNP} = .037 for rs12262390). For some genes, the extent of LD surrounding significant SNPs implicates flanking genes. For example in *ARID2*, rs35115 (p_{SNP} = .0067) is located in intron 7 but also tags the non-synonymous variant rs7315731 in *SFRS2IP* (r^2 = .93). These examples demonstrate that defining a gene boundary requires a balance between capturing all possible SNPs influencing the gene and introducing SNPs that may be more functionally relevant to other genes. A more sophisticated approach to establish gene boundaries that

defines each gene boundary separately by considering the genomic context around the gene may be helpful in future gene-based approaches.

Gene-based approaches to interpreting the results of candidate gene and even genome-wide association studies are important because most variation influencing susceptibility to T2D and other common complex traits is currently expected to be gene-centric, although the definition of a gene is constantly evolving. Detailed coverage of the common variation in these genes represents a critical requirement for an effective and thorough gene-based study. Here we have identified genes significantly associated with T2D and related quantitative traits that are attractive targets for future replication studies. Confirmation in a larger sample set and meta-analyses across studies will be important to help determine the role of these genes.

Contributions

A version of this work has been published previously as: Kyle J Gaulton, Cristen J Willer, Yun Li, Laura J Scott, Karen N Conneely, Anne U Jackson, William L Duren, Peter S Chines, Narisu Narisu, Lori L Bonnycastle, Jingchun Luo, Maurine Tong, Andrew G Sprau, Elizabeth W Pugh, Kimberly F Doheny, Timo T Valle, Goncalo R Abecasis, Jaakko Tuomilehto, Richard N Bergman, Francis S Collins, Michael Boehnke, Karen L Mohlke. Comprehensive association study of type 2 diabetes and related quantitative traits with 222 candidate genes. *Diabetes.* 2008 Nov;57(11):3136-44.

MB, FSC, RNB, JT designed the FUSION study. KJG, KLM designed the candidate gene study. KJG, KLM, CJW, LJS, FSC, MB wrote the manuscript. KJG, CJW, YL,

LJS, KNC, AUJ, PSC performed data analysis. KJG, JL, MT, NN, AGS, EWP, KFD performed genotyping.

Chapter IV

Targeted sequencing of the HDL cholesterol level associated loci *GALNT2* and *MVK/MMAB* and triglyceride level associated loci *MLXIPL, TRIB1* and *ANGPTL3* in Finnish individuals with trait values in the tails of the trait distribution

Abstract

Recent genome-wide association studies have identified many novel risk loci for plasma lipid levels. To identify additional genetic variants contributing to individual trait variability at these loci, we sequenced three novel loci associated with triglyceride level, *MLXIPL, TRIB1,* and *ANGPTL3* and two loci associated with HDL cholesterol level, *GALNT2* and *MVK/MMAB* in 188 Finnish individuals with trait values in the tails of the trait distribution. We identified common variants in linkage disequilibrium with known associated SNPs that represent additional functional targets, including several non-synonymous, 3' UTR, and non-coding variants in predicted hepatic regulatory regions. In addition, between 10% -15% of variants in linkage disequilibrium with what at a given threshold were indels. Among less common and rare variants, we identified several with trait association that may independently contribute to trait values, including a variant in intron 1 of *TRIB1* and rare variants unique to low HDL-C individuals in the *GALNT2* 3' UTR. This study represents the first sequencing effort of these lipid-level associated loci in phenotypically-selected samples.

Introduction

Plasma concentrations of lipids such as HDL cholesterol, LDL cholesterol and triglycerides are important risk factors for the development of coronary artery disease, the leading cause of morbidity and mortality in industrialized nations [185]. In the past several years, genome-wide association studies have identified many common variants contributing to individual variation in plasma lipid levels, many of which were previously unknown [85,86].

For triglyceride level, recently identified loci include variants near *MLXIPL, ANGPTL3* and *TRIB1* [86]. Both *MLXIPL* and *ANGPTL3* have links to triglyceride metabolism. *MLXIPL* encodes a transcription factor that binds carbohydrate response element motifs upstream of genes involved in triglyceride synthesis[186], and *ANGPTL3* encodes a protein that inhibits lipoprotein lipase (LPL) that in turn regulates triglyceride metabolism [187]. *TRIB1* regulates mitogen-activated protein kinases and helps control vascular smooth muscle proliferation [188], although how this function may contribute to triglyceride metabolism is unclear. As the associated variants in this region are 25 kb downstream of *TRIB1*, the contribution to triglyceride level variability may involve another gene or transcript altogether.

Novel susceptibility loci for HDL-C level include variants near *GALNT2* and *MMAB/MVK* [86]. The *MMAB/MVK* associated region spans greater than 350 kb and includes two bi-directionally organized genes, *MMAB* and *MVK,* that are both known to be regulated by SREBP2 [189]. The protein product of *MVK* catalyzes an early step in

cholesterol biosynthesis[189], and the product of *MMAB* is involved in cholesterol degradation[190]. In addition, HDL-C associated variants at this locus are correlated with levels of *MMAB* transcript in hepatocytes[191,192]. The association signal at *GALNT2* is localized to an approximately 14 kb region in intron 1 of the gene. *GALNT2* is a glycosyltransferase involved in O-linked oligosaccharide biosynthesis [193], and the mechanism by which this gene may influence HDL-C level is currently unknown.

Identification of susceptibility loci influencing lipid levels has subsequently created a new set of challenges, including identification of the variant(s) functionally responsible for the association signal and identification of other variants in these genomic regions influencing trait susceptibility. Targeted re-sequencing of associated regions and/or nearby genes in a small to moderate number of samples should identify the full suite of common variants potentially responsible for the association, and genotyping the more complete set of associated SNPs in a large sample could help localize the signal [194]. The 1,000 Genomes Project should facilitate susceptibility variant discovery by uncovering a more complete set of common SNPs, insertions-deletions (indels) and copy number variants (CNVs) across the genome [28]. Targeted functional study of candidate variants may successfully identify likely functional culprits, as has been accomplished previously for other GWA loci [195,196].

Deeper re-sequencing efforts are aimed at identifying rare variants not easily captured through linkage disequilibrium (LD)-based approaches and that may contribute independently to trait susceptibility. One widely used approach is to re-sequence

genomic regions in phenotypically-selected groups of individuals, such as those at phenotypic extremes of the trait distribution, and then identify highly penetrant variants preferentially found in one group[29,31,32,197,198]. These studies have thus far been primarily restricted to re-sequencing coding regions, which requires less sequencing and at which predictions of variants with a functional effect can focus on protein changes.

We sequenced genomic regions representing three loci associated with TG level and two loci associated with HDL-C level in normoglycemic individuals with high (>95th percentile) and low (<5th percentile) TG or HDL-C level. Common SNPs and indels were in LD with previously associated variants that represent additional functional targets at these loci. We also identified less common (MAF<.05) transcribed and non-coding variants that may contribute to trait variability.

Methods

Sample selection for sequencing

Samples were selected from 2,335 Finnish individuals part of the FUSION [199] and Finrisk 2002 [41,200] studies that had been previously genotyped on 315,000 SNPs using the Illumina HumanHap300 BeadChip [201]. Although FUSION is a study of type 2 diabetes, we chose to exclude individuals from sequencing affected with type 2 diabetes to remove the influence of diabetes status on lipid levels. We also excluded individuals whose trait values at the time of clinical exams could have been influenced by lipid-lowering medications. Study protocols were approved by local ethics committees and/or institutional review boards, and informed consent was obtained from all study

participants. Serum HDL cholesterol and triglycerides were quantified from samples as described previously [199]

The 188 samples in the HDL-C set consisted of 94 individuals that had the highest or lowest HDL-C levels. The low HDL-C set had a mean HDL-C level of 0.87±0.13 mmol/L, and the high HDL-C set had a mean HDL-C level of 2.30±0.25 mmol/L. The 188 samples in the TG set consisted of 94 individuals that had the highest or lowest TG levels. The low TG set had a mean TG level of 0.61±0.081 mmol/L, and the high TG set had a mean TG level of 2.87±0.72 mmol/L (**Table 4.1**).

A total of 322 Finnish individuals were selected for re-sequencing from the FUSION and Finrisk 2002 studies, as there was overlap in the set of samples selected for HDL and TG loci; 143 samples are unique to the HDL-C set, 134 are unique to the TG set, and 55 samples are included in both sets.

Follow-up samples (Stage 2) consisted of 18,860 individuals from the FUSION, DIAGEN, METSIM and HUNT studies.

Genomic region selection

We selected five novel loci associated with HDL-C and TG level in more than 20,000 individuals [86]; two independent novel loci associated with HDL-C, chromosome 12 including *MMAB* and *MVK* and chromosome 1 in intron 1 of *GALNT2*, and three

independent novel loci associated with TG, chromosome 7 near *MLXIPL*, chromosome 8 downstream of *TRIB1*, and chromosome 1 near *ANGPTL3*.

For each locus we selected the regions for re-sequencing based on three criteria. First, we included the region spanning the associated SNPs at each locus, empirically determined from GWA data. Second, we included the best candidate genes near each associated locus. For *MVK*, *MMAB, MLXIPL,* and *TRIB1* the full gene including all isoforms was selected; *GALNT2* was too large to re-sequence the entire gene so in this case only exons were selected. Third, we used published genome-wide data profiling histone modifications and variants (H3K4me1, H3K4me3, H3K9me1, H2A.z) [202], DNase hypersensitivity [96] and DNA binding proteins (CTCF, RNA Pol II) [202] to select regions likely to harbor regulatory elements.

In total, 175 kB at the HDL-C and TG associated loci was targeted for re-sequencing: 119 kB of gene regions, 33 kB of associated regions lying outside of the candidate gene boundaries, and 22 kB of regions containing functional regulatory data.

Sequencing and SNP detection

Sequencing was performed at the J. Craig Venter Institute as part of the National Heart Lung and Blood Institute Re-sequencing and Genotyping Service.

PCR reactions were performed with genomic DNA and products were analyzed by DNA sequencing. PCR reactions were categorized as "strict" if the amplicon Tm was less than

or equal to 82C, or "high_gc" if the amplicon Tm exceeded 82C and used different parameters. DNA sequencing was performed on ABI 3730xl DNA Analyzers. Initial base calling for sequence chromatograms was done using ABI kb V1.2. Sequence chromatograms were filtered using custom digital signal processing (DSP) software to attenuate "dye-blob", primer off-by-one, and PCR stutter artifacts. Filtered sequence chromatograms were base called and .poly files generated using a customized version of TraceTuner, calibrated for ABI 3730xl, POP7, and BDTv3.1.

Mixed bases were called, using custom software, by comparing areas, heights, and locations of minor peaks versus major peaks identified in the .poly file. The high signal-to-noise-ratio (SNR) clear range was identified for each chromatogram. Sequence chromatograms were analyzed for the presence of heterozygous indels. Sequence chromatograms containing heterozygous indels were computationally split into long haplotype and short haplotype versions. PCR primer sequences were screened from the chromatogram sequence using cross_match. Chromatogram sequences were assembled together with the reference amplicon sequence using phrap. Mixed bases (i.e., SWMKUY) were called when bi-directional reads agreed. Variations (substitution, insertion, or deletion) between the reference amplicon sequence and assembly consensus sequence were calculated and recorded.

Variants identified through computational methods were manually inspected if they were singletons, non-synonymous coding changes, heterozygous indels, or, if sequenced on multiple amplicons, had inconsistent calls between amplicons.

We observed a 21-bp deletion (rs6143660) as part of an independent sequencing project. We genotyped rs6143660 on 87 FUSION samples to determine its frequency and relation to HDL-C associated SNPs. Primers for amplification were selected using Primer3: Forward: 5'-CTCATCTTTGCACACGAAGG-3' Reverse: 5'-GAGACCCTGAGTGTGAGGCT-3'. The products for chromosomes with and without the deletion were 91 bp and 112 bp, respectively. The products were run on a 3% low melting point agarose gel and were scored by hand.

Stage 1 and Stage 2 genotyping of selected variants identified in sequencing was performed using the Sequenom homogeneous MassEXTEND (hME) assay.

Quality control

We excluded 67 SNPs identified in re-sequencing that failed Hardy-Weinberg equilibrium ($p < 1x10^{-3}$) and/or had low genotyping success (<75%). When comparing genotypes for 86 SNPs that were sequenced on more than one amplicon, SNPs were excluded that had more than 2 discrepancies.

For imputation, SNPs with MAF > .01 were excluded if they had an MACH imputation quality score (r^2 hat) < .3. SNPs with MAF < .01 were excluded if they had an r^2 hat < .5.

Variants in LD with trait associated SNPs

We identified published index SNPs for TG level associated loci: rs17145738 (*MLXIPL*), rs1748195 (*ANGPTL3*) and rs2954029 (*TRIB1*), and HDL-C level associated loci: rs2144300 (*GALNT2*) and rs2338104 (*MMAB/MVK*) [86]. We then identified SNPs in HapMap CEU release 22 in strong LD ($r^2 > .8$) with each index SNP. Using this set of SNPs (termed 'HapMap associated' SNPs), we then identified variants from sequence data and 1000 Genomes Project data in high ($r^2>.9$), moderate ($r^2>.5$) and low LD ($r^2>.2$) with these SNPs. 1000 Genomes Project LD files were created from the December 2009 pilot release of 60 CEU samples.

Selection of Stage 2 variants

We selected variants follow-up (Stage 2) genotyping using several criteria. All amino acid changing variants (non-synonymous, frameshift) not in HapMap were selected. A subset of variants with MAF < .05 were also selected that showed preliminary evidence of association in Stage 1 FUSION samples and also had HepG2 annotation. 6 of 12 variants selected for Stage 2 were successfully genotyped.

Quantitative trait association analysis

We tested all genotyped and imputed variants for Stage 1 association with HDL cholesterol and triglyceride level. We regressed the quantitative trait variables on age, age^2, sex, birth province, type 2 diabetes affection status, and study indicator, and transformed the residuals of each quantitative trait to approximate normality using inverse normal scores, which involves ranking all trait values and then converting these to z-scores according to quantiles of the standard normal distribution. We then carried out

association analysis on the residuals. To allow for relatedness, regression coefficients were estimated in the context of a variance component model that also accounted for background polygenic effects. We tested for association using the residuals and the expected allele count from imputation under an additive model.

For follow-up analysis, variants were tested for association separately for each Stage 2 study population (FUSION, METSIM, DIAGEN, HUNT). Results from Stage 1 and each Stage 2 study were then combined using meta-analysis.

Excess of group-unique variants

Starting with all 261 variants with one allele found only in individuals belonging to one trait group ('group-unique variants'), we looked for windows with significant deviations from the expected number of high and low-unique variants. We tested window sizes of 200 bp, 400 bp, 1000 bp and 2000 bp, where each new window started at the position of each group-unique variant. Thus, for each window size, the number of tested windows is equal to the number of group-unique variants. The significance of individual windows was determined by comparing the ratio of observed high or low-unique variants to the ratio obtained in each of 10,000 permutations of high/low status. *P*-values represent the number of permutations with a greater deviation in ratio of high-unique/low-unique variants in either direction. A total of 1,044 (261 variants x 4 window sizes) windows were tested, resulting in a *P* value of 4.8×10^{-5} for experiment-wide significance.

Results

Re-sequencing of loci

We selected three loci associated with triglyceride (TG) level, and two loci associated with high-density lipoprotein cholesterol (HDL-C) level for re-sequencing in 94 samples at each tail (5%/95%) of the distribution of TG or HDL-C trait values from the Finland-United States Investigation of NIDDM genetics (FUSION) study[86] (**Table 4.1**). Although FUSION is a study of type 2 diabetes, we chose to exclude individuals from sequencing affected with type 2 diabetes. Due to limited resources, regions selected for re-sequencing differed between the loci, always including at least exons of the closest candidate gene(s), and at some loci also including the region of association and regions predicted to contain regulatory elements based on data from CD4+ T cells (see Methods). In total, 130,466 (75%) of 174,595 targeted base pairs were successfully re-sequenced in essentially all 188 individuals (**Table 4.2**).

874 total variants (818 SNPs and 56 indels of 1 to 38 nucleotides) were identified (**Table 4.3**). 60% of these variants were novel, and 60% of novel variants had a minor allele frequency (MAF) less than 1%. After filtering variants for low genotyping success, HWE failure, and discrepancies between overlapping amplicons (see Methods), we further assessed genotype quality of the 807 remaining variants by measuring concordance with existing genotypes for sequenced samples [201]. Re-sequenced genotype concordance was 100% with 25 variants directly genotyped by Illumina Infinium, Illumina GoldenGate or Sequenom platforms, and 98.6% with 147 variants imputed from Illumina Infinium genotypes using HapMap reference haplotypes.

We then used the re-sequenced samples as reference haplotypes and imputed genotypes for all non-singleton, bi-allelic variants in 2,457 FUSION samples (Stage 1) [141], and tested these variants for HDL-C or TG association (see Methods). Follow-up genotyping (Stage 2) consisted of 18,860 samples from the FUSION, DIAGEN, METSIM and HUNT studies (See Methods).

Novel functional targets in LD with known trait-associated variants

Prior to sequencing, the set of common variants associated with HDL-C level or TG level at these loci primarily was restricted to variants present in HapMap. We identified HapMap SNPs in high LD ($r^2 > .8$) with the published index SNP for each locus, and considered them highly likely to be associated with the trait (see Methods; **Table 3.4**). Using this set of SNPs present in HapMap (termed 'HapMap associated' SNPs), we searched for additional (non-HapMap) variants identified by sequencing and determined to be in LD with a HapMap associated SNP. Across all five loci, 20 non-HapMap variants were in high LD ($r^2 > .9$) with a HapMap associated SNP (28 in $r^2 > .5$, 68 in $r^2 > .2$) (**Table 3.4**). 10-15% of variants identified at these r^2 thresholds were indels. We compared the patterns of LD identified in our sequencing data to those obtained from pilot data on 60 CEU samples from the 1000 Genomes Project (www.1000genomes.org; see Methods). Using 1000 Genomes Project data identified a larger set of non-HapMap variants in LD with a HapMap associated SNP at each locus, which is expected because resequencing was incomplete (**Table 3.4**). 87% of SNPs identified by resequencing to be in LD ($r^2 > .2$) with a HapMap associated variant were also identified at this threshold using 1000 Genomes Project data. When comparing LD patterns using the subset of

variants identified using both approaches, pairwise r^2 values exhibit strong correlation (Pearson's r = .92).

Several non-HapMap variants in moderate LD with what were located in transcribed regions (**Table 4.5**). One of these variants is a previously identified non-synonymous substitution in *MLXIPL* (Q241H, rs3812316) [203]. We found a novel non-synonymous substitution in exon 1 of *MMAB* (R18H, rs10774775, MAF = .31), in moderate LD with HapMap associated SNPs (r^2=.5 with rs7134594). This SNP is predicted by SIFT to be damaging, and it is in perfect LD with synonymous substitution R19 (rs10774774) at the adjacent base pair. In addition, we identified a common 9-bp deletion (rs34483103) in the 3' UTR of *ANGPTL3* (r^2 = 1 with HapMap associated variant rs11207997).

Differences in transcriptional regulatory activity may be the functional basis for association at susceptibility loci, especially in regions at which the entire association signal spans no known genes (near *TRIB1*) or is completely intronic (*GALNT2*). To identify non-coding, possible regulatory SNPs, we annotated variants with publicly available datasets predictive of transcriptional regulation in HepG2 cells [102,204-207] (**Table 4.6**). At the *GALNT2* locus, 8 HapMap SNPs are in high LD (r^2 > .8) with HDL-C associated variant rs2144300 [86], and an additional 9 in high LD were identified through sequencing (see Methods). Of these 17 variants, five (rs2144300, rs6143660, rs17315646, rs4846914, and rs10127775) are located in a ~1 kb genomic region that contains many HepG2 non-coding annotations, including open chromatin (FAIRE and DNase), transcription factor binding (cMYC, SREBP1A, and RNA Pol II) and histone

modification (H3K4me3) (**Figure 4.1**). These SNPs represent *a priori* the strongest candidates for regulatory variants at this locus.

Quantitative trait association

Of 116 variants with nominal evidence of association ($P < .05$) with either TG or HDL-C level, the majority represent variants in LD ($r^2 > .2$) with previously reported association signals (**Table 4.6**). Among variants in lower LD with a previously associated SNP, 21 were significant ($P < .05$) including 9 with MAF less than .01 (**Table 4.7**).

To follow up preliminary evidence of association and/or functional annotation, a subset of these 21 SNPs were genotyped in Stage 2 samples (see Methods). A common variant in *GALNT2* (rs56217501, MAF=.1) with modest Stage 1 HDL-C association ($P = .005$) was in low LD ($r^2<.2$) with HapMap associated SNPs at this locus (**Table 4.7**); however, after genotyping in Stage 2 samples evidence of HDL-C association was not significant ($P = .27$) (**Table 4.8**). Among less common variants, the most significant result after Stage 2 genotyping was rs72647336 in intron 1 of *TRIB1* ($P = 1x10^{-4}$ with TG level), located in a genomic region highly conserved between species (**Table 4.8**). The rs72647336 signal appears to be independent of the GWA-identified index SNP rs17321515 (Stage 1 $P = .017$; conditional on rs17321515, $P = .023$).

Rare alleles unique to one trait group in associated regions

We identified 261 variants for which the rare allele was present only in individuals in the high trait value group or low trait value group but not both. Among six coding variants

with MAF < .05 four were only found in one trait group: a frameshift mutation in *MMAB* (rs72650181), and non-synonymous mutations in *MLXIPL* (R841W; rs66489924, predicted to be damaging by SIFT [208]), *MVK* (V377I; rs28934897, reported in patients with Hyper-IgD syndrome [209,210]), and *ANGPTL3* (N151D; rs72649574) (**Table 4.9**). Among these variants, only rs28934897 and rs72650181 were successfully genotyped in Stage 1 and Stage 2 samples; both were no longer significantly associated (*P*>.4) with HDL-C (see Methods; **Table 4.8**).

Given large, contiguous re-sequenced regions for each locus, we used a sliding window approach to identify genomic regions with an excess of rare variants unique to one group (**Table 4.10**; see Methods). The greatest excess was in a 2 kb window spanning the 3' UTR of *GALNT2* (11 low HDL-C individuals with a variant, 1 high HDL-C individual with a variant), although this excess was not significant after correction for multiple windows tested.

Discussion

Our sequencing at five GWA loci identified additional common variant(s) expected to show evidence of association with HDL or TG based on LD. The number of new common variants we identified by sequencing varied greatly between loci. For example, at an r^2 threshold of .9, the number of likely HDL-C associated variants at *GALNT2* doubled, while at *TRIB1* only three additional variants were added to 17 TG-associated HapMap variants. One of the major goals of the 1000 Genomes Project is to catalog all common variation in European populations, which should render re-sequencing studies to

determine LD patterns unnecessary in these populations [28]. Our data suggest that using 1000 Genomes Project data, even at a pilot stage of 60 CEU samples, is sufficient to capture much of the same information about LD patterns as we observed with 188 samples, and even more given that targeted sequencing coverage was lower.

With a set of common variants significantly associated with a trait comes the challenge of identifying which variants are immediate candidates for functional study. Frameshift variants early in a coding region are obvious functional targets, as are some but not all non-synonymous substitutions and 3' UTR variants. However, some transcribed variants have no effect on a transcript or protein, and other variants on the same haplotypes may also affect function. For example, while we identified a non-synonymous substitution in *MMAB* (R18H; rs10774775) that is predicted damaging by SIFT [208] in moderate LD with HDL-C associated SNPs, previous studies have shown that the associated SNPs are correlated with hepatocyte transcript levels of *MMAB*. Therefore, it may be a combination of biological effects influencing HDL-C levels at this locus.

Non-coding annotations may usefully predict likely regulatory variants, especially at loci where the associated variants do not include any coding variants [211]. In our study, open chromatin, ChIP, and histone modification data generated in HepG2 cells, which may be relevant to cholesterol and triglyceride biology, identified a 1 kb portion of *GALNT2* intron 1 that contains 5 likely HDL-C-associated variants which warrant further study. However, for other loci currently available annotation data might not be as useful for prioritization. For example, the region roughly 20 kb downstream of *TRIB1* associated

with TG level had no striking annotations to suggest that any of the associated variants are strong *a priori* regulatory candidates. It may be that the contribution of this locus to TG level involves regulatory elements either not identified in HepG2 cells or with these specific techniques, or another biological mechanism altogether.

Sequencing samples selected from the trait extremes should enrich for the discovery of rare alleles with an effect on trait value that might be missed or is not present in databases or other sequenced individuals. However, to detect significant associations with rare variants either effects or sample sizes need to be substantial [28]. One approach to increase sample size is to impute the new rare genotypes in samples that have an existing genotype scaffold; this approach is limited by reduced accuracy of imputation at low allele frequencies.

Another approach is to directly genotype lower frequency variants in larger sample sets, although large numbers of variants with preliminary evidence of association coupled with low power to detect real associations requires genotyping many low frequency variants that may be false-positives. In our study we selected variants to follow-up based on coding and non-coding annotation, an approach that is limited by the quality and depth of available annotation. One annotated variant we selected for follow-up, rs72647336 in intron 1 of *TRIB1*, was significantly associated with TG level at $P = 1x10^{-4}$ in 16,000 samples. As this variant did not reach genome-wide significance ($P < 5x10^{-8}$), it is unclear whether it or a variant in strong LD functionally contributes to trait variability. The need for large sample sizes for rare variants is also complicated by variants that may

be population-specific and thus not present in replication samples from different populations. For example, a frameshift mutation in *MMAB* (rs72650181) with an allele frequency of .0002 in FUSION samples was monomorphic in all other studies.

A third option to detect significant, rare variant associations is to group rare variants together, an approach that has been successful in re-sequencing studies of coding regions [197]. Identifying meaningful groups of non-coding variants is a more complex problem, confounded by the under-annotation of the non-coding genome and the difficulty in determining *a priori* the directional effect of a variant. For example, we identified a modest excess of low HDL-C individuals with variants in the 3' UTR of *GALNT2*. 3' UTR variants may influence transcript stability through differences in miRNA binding, and we searched for predicted miRNA sites at these variants. Several variants unique to low HDL-C individuals overlapped a predicted miRNA site; however, predictions of how these variants might influence binding of the miRNA, if at all, are needed to help determine whether they are altering transcript stability and thus phenotypic output in the same direction. Finally, given the likely smaller effect sizes of many non-coding variants, much larger sample sets than the 188 sequenced in this project would need to be sequenced to tease out subtle effects through grouping analysis.

Given the sequencing capabilities afforded by next-generation technologies, it is now feasible to consider the availability of whole-genome re-sequencing data on a large number of samples selected based on phenotype. While the amount of data generated by

these efforts will greatly eclipse the scope of this project, fundamental questions remain about how to optimally identify and follow-up variants influencing trait variability.

Contributions

Kyle J Gaulton, Tanya Teslovich, Cristen J Willer, Lori Bonnycastle, Anne U Jackson, Laura J Scott, Peter S Chines, Narisu Narisu, Amy Swift, Mario Morken, Timothy Stockwell, Dana Busam, Samuel Levy, Francis Collins, Mike Boehnke, Karen Mohlke

KJG and KLM designed the study and wrote the manuscript. KLM, MB, FSC designed the FUSION study. KJG, TT, CJW, AUJ, LJS performed data analysis. TS, DB and SL performed Sanger sequencing. LB, PSC, NN, AS and MM performed replication genotyping

Chapter V

Mapping regions of open chromatin using FAIRE

Abstract

Tissue-specific transcriptional regulation is central to human disease. To identify regulatory DNA active in human pancreatic islets, we profiled chromatin by formaldehyde-assisted isolation of regulatory elements coupled with high-throughput sequencing (FAIRE-seq). We identified ~80,000 open chromatin sites. Comparison of FAIRE-seq data from islets to that from five non-islet cell lines revealed ~3,300 physically linked clusters of islet-selective open chromatin sites, which typically encompassed single genes that have islet-specific expression. We mapped sequence variants to open chromatin sites and found that rs7903146, a *TCF7L2* intronic variant strongly associated with type 2 diabetes, is located in islet-selective open chromatin. We found that human islet samples heterozygous for rs7903146 showed allelic imbalance in islet FAIRE signals and that the variant alters enhancer activity, indicating that genetic variation at this locus acts in *cis* with local chromatin and regulatory changes. These findings illuminate the tissue-specific organization of *cis*-regulatory elements and show that FAIRE-seq can guide the identification of regulatory variants underlying disease susceptibility.

Introduction

Pancreatic islets are groups of endocrine cells that secrete insulin, glucagon, and other polypeptide hormones. Beta cells, the predominant islet cell type, secrete insulin in response to glucose. Insulin, in turn, promotes cellular glucose uptake [212]. Defects in beta cell mass or function consequently result in diabetes, a leading cause of blindness, kidney failure, heart disease, and premature death[213].

The transcriptional regulation of genes that function in islet cells has major implications for human diabetes. In Type 1 diabetes, which results from the autoimmune destruction of beta-cells, one of the major research goals is to generate new beta cells for replacement therapies by transcriptional reprogramming [214]. In many cases of monogenic inherited forms of Type I diabetes, the disease is caused by mutations in genes encoding for islet-cell transcriptional regulators.

Transcriptional control of islet cells is also relevant for Type 2 diabetes (T2D), the most prevalent form of this disease . [213] T2D results from decreased function and mass of islet beta-cells, coupled with insulin resistance in peripheral tissues [215-217]. The complex etiology of T2D includes numerous environmental and genetic factors that each contribute to disease susceptibility and pathogenesis [218]. Genome-wide association studies have identified over 20 loci that harbor common risk variants for T2D, many of which are likely to influence T2D susceptibility through impaired insulin secretion and beta cell function [63,69,201,219-229]. Many loci do not contain an associated non-synonymous coding variant, so it is likely that the functional variants at most T2D susceptibility loci are involved in regulation of gene activity. At many loci, several hundred variants are

associated with the linked genomic interval, and determining which of these is functional, either singly or in combination, is a difficult challenge. Knowledge of the exact location of the regulatory elements utilized in islets would be valuable in narrowing the search for DNA sequence variants that contribute to T2D pathogenesis.

The current state of knowledge regarding genomic regulatory sequence elements remains extremely sparse. Regulatory elements are likely to be highly specific to cell type, influenced by developmental cues, and specified in part by the local cellular environment. An added complexity is that detailed studies of individual loci suggest that regulatory elements function in concert as complex units, rather than in isolation. Comprehensive identification of regulatory DNA is needed to understand differences in transcriptional regulation across tissues, physiological conditions, and individuals.

Regulatory elements function by recruiting DNA-associated proteins to specific loci. This process of recruitment and factor binding typically results in local nucleosome eviction[230]. Nucleosome loss is therefore an evolutionarily conserved indicator of regulatory activity, and can be used as a molecular tag to isolate regions of the genome that are bound by regulatory factors in a give cell or tissue type [91]. Nucleosome loss has traditionally been detected by hypersensitivity to nuclease digestion [231-234]. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) is an alternate technique for isolating and identifying nucleosome-depleted DNA from cells. In FAIRE, cells are fixed lightly with formaldehyde. Whole cell extract is then prepared, sonicated to shear chromatin, and subjected to phenol-chloroform extraction. DNA fragments that are not

covalently linked to proteins are recovered in the aqueous phase, and can be identified by any number of common methods [235]. The mechanism by which FAIRE efficiently recovers nucleosome-depleted regions is proposed to be rooted in the extremely high efficiency of crosslinking histone proteins to DNA relative to other DNA-binding proteins such as transcription factors [236,237].

In human cells, FAIRE was first performed in a fibroblast cell line [235]. In that study, genomic segments recovered after FAIRE were identified by fluorescent labeling and hybridization to tiling DNA microarrays covering 1% of the human genome [235]. The DNA elements identified in this way exhibited concordance with established hallmarks of regulatory function such as DNaseI hypersensitivity, active promoters, and evolutionary constraint. Given the sparse annotation of the human regulatory landscape, it is not surprising that FAIRE also identified genomic regions not yet annotated by other methods [89,235]. The advent of next-generation sequencing technologies makes possible rapid genome-wide mapping and direct quantification of FAIRE fragments at high resolution. We call this method FAIRE-seq, which combines the power of deep sequencing with the simplicity and flexibility of FAIRE for the rapid identification of open chromatin regions from primary human tissue.

We performed FAIRE-seq on purified pancreatic islets. To our knowledge, this is the first detailed atlas of regulatory elements in this tissue. The unbiased maps generated by FAIRE-seq reveal new insights regarding the higher order organization of tissue-specific cis-regulatory elements, and provide a foundation for mechanistic understanding of

transcriptional regulation of genes important for pancreatic islet function and type 2 diabetes susceptibility. The data show that FAIRE-seq can identify regions of open chromatin from a small amount of primary tissue, and provide guidance for studies seeking to characterize functional SNPs in tissues relevant to human disease.

Methods

Islet sample preparation

All experiments were performed according to protocols approved by the Institutional ethical committees of the Hospital Clinic de Barcelona, Geneva University Hospitals, Istituto Scientifico Ospedale San Raffaele, and Hospital Universitari de Bellvitge. All samples were isolated from multiorgan donors without a history of glucose intolerance after informed consent from family members. Information on samples used for FAIRE-Seq are provided in **Table 5.1**. Pancreatic islets were isolated according to established procedures[238]. After isolation, islets were cultured in CMRL 1066 containing 10% FCS and shipped at room temperature in the same medium. Samples 1 and 2 were processed upon arrival, while sample 3 and subsequent samples used in locus-specific assays were recultured in RPMI 1640 containing 10% FCS for three days before performing FAIRE. Islets were rinsed with PBS three times, and crosslinked for 10 min in 1% formaldehyde at room temperature with constant shaking. After adding glycine (final concentration 125 mM), islets were rinsed with PBS containing protease inhibitor cocktail (Roche) at 4°C, snap frozen, and stored at -80°C. Islet purity was assessed by dithizone staining[239] immediately prior to fixation. The accuracy of dithizone staining was verified by immunofluorescence analysis using DAPI, anti-insulin, and anti-glucagon antibodies[240].

FAIRE

For all but 2 samples (see below), FAIRE was performed as described[235] with modifications. Frozen pellets with ~3000 crosslinked islets were thawed on ice in 1 mL lysis buffer (2% Triton X-100, 1% SDS, 100 mM NaCl, 10 mM Tris-Cl at pH 8.0, 1 mM EDTA) and disrupted with five 1-minute cycles using 0.5 mm glass beads (BioSpec). Samples were sonicated for 10-20 rounds of 30 pulses (1 second on/0.5 second off) using a Branson Sonifier 450D at 15% amplitude. After 10 rounds the efficiency of sonication was assessed, and further rounds were performed when needed to ensure that the majority of chromatin fragments were in the 200-1000 bp range. Debris was cleared by centrifugation at 15,000 g for 5 minutes at 4°C. Nucleosome-depleted DNA was extracted with phenol-chloroform followed by ethanol precipitation and RNase A (100 μg/mL) treatment[235].

For samples 1 and 2 we employed a modified procedure that yielded less consistent chromatin fragmentation. Cells were incubated with 50 mM HEPES (pH 8.0), 140 mM NaCl, 1 mM EDTA, 0.1% SDS, 0.1% sodium deoxycholate, then centrifuged at 9,000 g 10 min. The pellet was resuspended in 50 mM HEPES (pH 8.0), 140 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.1% SDS, 0.1% sodium deoxycholate, and then processed as described above.

Sequence analysis

Libraries were generated from gel-purified ~200 bp DNA fragments. After adaptor ligation and PCR-based amplification, samples were sequenced on the Illumina Genome

Analyzer II platform using standard procedures. Sequence reads were aligned to the human reference genome (hg18) using Mapping and Assembly with Qualities (MAQ) with default mapping parameters [241]. Post-alignment processing removed all reads that had an overall MAQ mapping quality <30 and artificially extended each read to a final length of 200 bp. We counted filtered reads mapping to each base in the genome to obtain a read density for each base. To facilitate display, read densities were centered on the mean read density of each chromosome.

Sites of FAIRE-seq enrichment were assessed with F-Seq[242], which uses a kernel density estimate to calculate genomic regions where the continuous probability is greater than a user-defined standard deviation threshold over the mean across a local background. We used a feature length of 1,000 and three standard deviation (s.d.) thresholds resulting in three sets of enriched regions for each sample. The most liberal threshold was set for each sample using an empirical estimation of the upper bounds on the number of nucleosome-depleted regions genome-wide (roughly 200,000). For sample 3, the thresholds used were s.d.=6 (liberal), s.d.=8 (moderate) and s.d.=10 (stringent). For samples 1 and 2, which were sequenced to a lower depth, the thresholds used were s.d.=4 (liberal), s.d.=5 (moderate) and s.d.=8 (stringent).

We estimated the mappable proportion of the reference genome in two ways, using 120 million randomly generated reads ($2.7x10^9$ mappable bases, ~85% of hg18) and using PeakSeq ($2.8x10^9$ mappable bases, ~89% of hg18)[243]. We independently calculated

genome coverage using 125 million reads obtained from islet FAIRE and found it to be 97.9% and 97.7% concordant, respectively.

Sequence reads obtained from five major ENCODE cell lines (GM12878, HeLa-S3, HUVEC, K562, and HepG2) were aligned to the reference genome (hg18) using MAQ[241] and filtered as described above. Sites of enrichment were determined using F-Seq[242], using the same parameters as islet samples 1 and 2.

Microarray analysis

Islet FAIRE preparations from samples 1 and 2 were fluorescently labeled and hybridized to a tiling DNA microarray covering 1% of the genome selected for the ENCODE pilot project[89]. Sites of enrichment were called using ChIPOTle[244]. For Receiver Operating Characteristic (ROC) curve analysis, sites of enrichment were called using a p-value threshold of $1x10^{-12}$.

Regulatory feature analysis

We recorded the percentage of bases underlying FAIRE-seq sites that overlapped 28-species conserved elements[207], predicted regulatory modules (PreMod)[110] and transcription factor binding sites (TRANSFAC[245] and MotifMap[102]). For each set of peaks we permuted positions across the mappable genome 1,000 times and re-calculated the overlap. *P* values were calculated from permutations that had a higher degree of overlap than the observed set of peaks. We used Clover to test for over-represented transcription factor binding motifs in sequences underlying intergenic FAIRE-seq enrichment[246]. Sequences were separated by chromosome and analyzed for motifs from

JASPAR[247] and TRANSFAC[245], as well as the CTCF motif[248]. Significance was calculated by comparing to the mappable intergenic portion of the tested chromosome, and motifs reaching a p-value threshold of .01 were reported.

FAIRE-Seq and expression level analysis

We used RMA-normalized signals from a previously reported experiment using HG-U133A and HG-U133B GeneChips with five non-diabetic islet samples[249], and obtained an average value for each probe. The five samples were selected by hierarchically clustering expression data from 7 non-diabetic individuals. We excluded two samples (Sydp2 and SydPI) that had poor concordance with the others and showed low expression of known islet genes. We counted the number of FAIRE-Seq reads mapping to each base in a 1 kb window surrounding each RefSeq TSS, grouped RefSeq genes by their average islet expression level, and, for each group, calculated the average mean-centered FAIRE read density at every base in the window.

Islet-selective and ubiquitous site definitions.

An islet FAIRE-seq site was considered *islet-selective* if the site did not overlap a site from any of the five additional tested cell types. Note that such sites are not expected to be necessarily unique to islets. An islet site was considered *ubiquitous* if the site overlapped a FAIRE-seq site in all five additional cell types. Moderate stringency FAIRE-seq site thresholds were used for all datasets.

Selection of genes with islet-selective open chromatin

For each RefSeq transcript we assessed the region 2 kb upstream through 2 kb downstream and calculated the percentage of bases that overlapped a moderate islet FAIRE enrichment site. We calculated the same value in the combined data from five non-islet cell lines, and selected genes that were more or equally enriched in islets compared to combined non-islet data.

Clusters of Open Regulatory Elements (COREs)

We identified 3,348 regions with at least three islet-selective sites (as defined above) located <20 kb from each other using Galaxy[250]. These criteria were based on the typical size of islet-selective clusters (**Fig. 3b)**. For comparisons we created a set of randomized COREs of identical size and mappability. To assess CTCF binding enrichment, we obtained CTCF binding sites from multiple cell types[251], calculated the frequency within COREs (0.007 sites/kb), in randomized COREs (0.013 sites/kb), and in the mappable genome (0.013 sites/kb), and tested for significance using a two-sided χ^2 test.

RefSeq, ncRNA, or Non-RefSeq Unigene transcripts were assigned to a CORE when the transcriptional start or end site was within 10 kb of either end of the CORE. When comparing CORE overlaps with one or more genes we required a minimum overlap of one bp. We used DAVID[252] to test enrichment of biological processes in CORE genes, and used all RefSeq genes as a background. We employed all biological processes from GO and PANTHER[253], removing GO terms associated with >1,000 genes from the analysis.

For gene expression comparisons, we obtained gcRMA-normalized signals from Human U133A and GNF1H microarrays from seven tissues (pancreatic islet, liver, heart, kidney, lung, skeletal muscle and whole brain)[254]. We required that probes reported a signal of >100 in at least one tissue. We used one-way ANOVA to assess expression differences between tissues for CORE genes and for an identical number of randomly selected genes that did not overlap a CORE.

Definition of T2D susceptibility variants

We identified 20 loci where SNPs have been reported to show genome-wide association ($P<5x10^{-8}$) with T2D or fasting glycemia[67,69,71,219,221,223,255,256]. For each locus we identified variants in HapMap CEU release 22 in strong linkage disequilibrium ($r^2>0.8$) with the reported reference SNP; 350 variants satisfied these criteria and were termed 'T2D-associated SNPs'. We then defined the region of association for each locus by manually identifying recombination hotspots from HapMap release 22 data flanking the associated SNPs. We identified SNPs in dbSNP v129 with an average heterozygosity >1% in these regions. SNPs that overlapped FAIRE sites in sample 3 were recorded.

Detection of allelic imbalance in open chromatin and PCR analysis of FAIRE.

We genotyped 31 human islet genomic DNA (gDNA) samples using TaqMan SNP Genotyping Assays (Applied Biosystems). Nine samples were heterozygous for rs7903146. For these samples we used TaqMan SNP Genotyping Assays to determine the allelic ratio of DNA fragments containing rs7903146 in FAIRE and input DNA. All reactions were performed in triplicate in a volume of 25 μl using 5 ng DNA quantified

with a Nanodrop 1000 (Thermo Scientific). A standard curve was generated by mixing gDNA from samples with known genotype to generate seven allelic ratios - 10:90, 20:80, 40:60, 50:50, 60:40, 80:20, and 90:10. The relative abundance of T and C alleles in each experimental sample was estimated from the standard curve, and compared to input DNA from the same samples and gDNA from unrelated heterozygous individuals. Allelic ratio was also assessed in seven samples heterozygous for rs7903146 by quantitative Sanger sequencing in FAIRE and input DNA (oligonucleotides shown in **Table 5.**). ImageJ was used to quantify the area under the curve of peaks in the chromatogram. Data are expressed as mean ± s.d. and were assessed with two-sided unpaired t-tests for gDNA vs. FAIRE, or paired t-tests for input vs. FAIRE.

To confirm islet-selective FAIRE enrichments, we employed real-time PCR with SYBR Green detection as described[257,258]. We performed triplicate measurements from 5 ng FAIRE DNA, and used a serial dilution of input DNA as the standard curve. We expressed FAIRE enrichment values relative to the enrichment values in the same sample at a local negative control region.

Luciferase reporter assays

A 240 bp fragment surrounding rs7903146 was PCR-amplified from DNA of individuals homozygous for either the T or C allele of rs7903146 (oligonucleotides shown in **Table 5.**). PCR fragments were cloned in both orientations in the multiple cloning site of the minimal promoter-containing firefly luciferase reporter vector pGL4.23 (Promega). Four independent clones for each allele for each orientation were verified by sequencing and transfected in duplicate into MIN6[259] and 832/13 (Chris Newgard, Duke University) ß-

cell lines, and into HEK293T cells. Cells were co-transfected with a phRL-TK *Renilla*-luciferase vector to control for transfection efficiency. Transfections were performed with lipofectamine 2000 (for MIN6 and HEK 293T; Invitrogen) or FUGENE-6 (for 832/13; Roche Diagnostics). Cells were assayed 48 hours after transfection using the Dual Luciferase Assay (Promega). Firefly luciferase activity was normalized to *Renilla* luciferase activity and then divided by values for a pGL4.23 minimal promoter empty vector control. A two-sided t-test was used to compare luciferase activity between alleles. Experiments in MIN6 and 832/13 cells were carried out on a second independent day and yielded comparable allele-specific results.

Results

For three samples of purified human pancreatic islets, we used FAIRE[235,236,260] to identify sites of open chromatin (**Figure 5.1a, Table 5.1, Table 5.1**). We technically validated FAIRE-seq by its high concordance to patterns determined by hybridization of the same FAIRE samples to tiling DNA microarrays (**Figure 5.1b, Figure 5.5a**). Furthermore, we found that despite differences in age, cause of death, genotype, islet isolation procedures, and level of exocrine cell contaminants, the majority of regions identified by FAIRE in any one sample were also detected in the others (**Figure 5.5b**). Thus, FAIRE-seq is a robust method for characterizing chromatin in islets.

Several lines of evidence indicate that FAIRE reliably identifies active regulatory elements in islets. Consistent with FAIRE in fibroblasts[235], the most enriched FAIRE

regions were found near known Transcription Start Sites (TSS) (**Figure 5.2a**). Overall, there was a positive relationship between FAIRE signal near TSS and transcript levels in human islets[249] (**Figure 5.2a**). Furthermore, promoters previously shown to bind RNA Polymerase II and the transcription factors HNF4A and HNF1A in islets[261] were enriched by FAIRE more frequently than other promoter regions (**Figure 5.2b)**.

To extend this observation, we identified regulatory regions that were utilized selectively in islets relative to other cell types. We compared FAIRE-seq data from islets to data from five non-islet cell lines (HeLa-S3, HUVEC, GM12878, HepG2 and K562; **Methods**) and found that 45% of islet open chromatin sites were unique to islets among this group of cell types. We refer to these sites as *islet-selective* open chromatin. We identified 340 RefSeq genes with islet-selective open chromatin in the TSS or gene body (**Table 5.3**). This relatively short list included 24 well characterized genes that are selectively expressed in islets (**Table 5.4**), including genes involved in human diabetes (*PDX1, ABCC8, SLC30A8, G6PC2, GAD2*) and islet developmental regulators (*NEUROD1, NKX6-1, PAX6, ISL1*)[221,262-265]. Therefore, islet-selective open chromatin detected by FAIRE identifies loci integral to islet-cell biology and disease.

Many sites of open chromatin detected by FAIRE are located in intergenic regions, far (>2 kb) from a known TSS. For these distal sites, evidence also points strongly toward a regulatory function. First, distal intergenic open chromatin sites were enriched in evolutionary conserved sequences, predicted transcription factor binding sites and regulatory modules, regardless of whether the open chromatin was islet-selective or

ubiquitous (shared by all six cell types) (**Figure 5.2c, Figure 5.5d** and **Table 5.5**, **Methods**). Second, ubiquitous intergenic open chromatin often coincided with binding sites of CTCF[248,251,266] (observed 16%, expected 0.39%, $P<0.001$) (**Figure 5.2c**), a transcriptional regulator and insulator protein that binds to a large number of genomic sites, many of which are shared in different cell types[266]. Open chromatin at CTCF sites was centered at the location of the CTCF binding[248], suggesting that FAIRE signal is indicative of interactions between regulatory factors and DNA (**Figure 5.2d**). Third, intergenic islet-selective (but not ubiquitous) open chromatin preferentially harbors DNA-binding motifs of pancreatic islet transcription factors, including RFX, TCF1/HNF1, HNF3B, FOXD, and MAF ($P<0.01$, **Table 5.5**). Notably, whereas ~36% of ubiquitous open chromatin was located within 2 kb upstream of a TSS or in the first exon, only 1% of islet-selective open chromatin was located in these regions (**Figure 5.2e**, **Figure 5.5c**). Collectively, these findings indicate that distal FAIRE sites harbor regulatory elements, and consistent with recent studies of histone modification patterns in enhancer regions[267] suggest that most cell-type specific open chromatin is located far from known TSS.

We next sought to link these distal islet-selective elements to specific genes. We examined whether islet-selective sites exhibit a higher-order organization that could point to the existence of functional domains. We found that open chromatin sites were not evenly distributed throughout the genome, but instead were located in physically linked clusters (**Figure 5.3a**). Clustering was also observed with islet-selective open chromatin (**Figure 5.3b**). We identified 3,348 domains containing at least three islet-selective open

chromatin sites separated by less than 20 kb, which we call islet-selective COREs (Clusters of Open Regulatory Elements) (**Figure 5.3c**, **Methods**). Islet-selective COREs had a median size of 25 kb, with the largest containing 148 FAIRE sites spanning 602 kb **(Figure 5.3d)**. Consistent with CTCF binding to insulator elements separating chromatin domains[268], the frequency of CTCF binding sites was two-fold higher outside than within COREs (P=1.3x10^{-48}). This suggested that islet-selective COREs were functional chromatin domains and provided an avenue to assigning open chromatin sites to genes.

Islet-selective COREs were located within 10 kb of at least one RefSeq gene in 69% of cases (randomized COREs=54%; P=1.5x10^{-35}, **Figure 5.3e**). Of these, 94% were associated with only one gene, and most were contained within 2 kb of gene boundaries (**Figure 5.5**) suggesting single-gene regulatory function (expected=84%; P=6.2x10^{-23}, **Figure 5.3e**). Compared to six other primary tissues, genes overlapping islet COREs had higher expression in islets and brain (one-way ANOVA, both tissues P<1x10^{-5}, **Figure 5.3f)**, consistent with the neuroendocrine nature of islet-cells[269]. Islet-selective COREs were also enriched in genes linked to islet-specific functions, including transcription factors, ion channels, and secretory apparatus components (**Table 5.6**). Thus, islet-selective COREs are typically linked to single genes that are expressed in an islet-selective manner.

A subset of islet-selective COREs spanned remarkably broad distances at loci encoding critical regulators of pancreas development and function (**Figure 5.3g, Table 5.7** and **Figure** 5.**6** and **Figure 5.7**). For example, an islet-selective CORE spanned a 46-kb

domain containing *PDX1,* a master regulator of pancreas development and β-cell function[263] (**Figure 5.3g**). At this locus, FAIRE sites coincided with previously characterized evolutionarily conserved enhancers named "Area I-IV"[270,271] and with uncharacterized putative enhancer sites (**Figure 5.3h**). Other islet-selective COREs included a 94-kb domain 3′ of *NKX6-1*, an essential regulator of β-cell differentiation[272], one located in a cluster of brain-enriched snoRNA and miRNAs[273], and another in conserved sequences >400 kb from any annotated gene (**Figure 5.7** and **Table 5.7** for additional examples). Such domains contrasted with loci devoid of open chromatin and known to be inactive in islets (**Figure 5.7q-s**). This dataset thus provides a rich resource to dissect *cis* regulation in pancreatic islets.

Recent genome-wide association studies for T2D susceptibility have implicated sequence variants at multiple loci, many of which may impair islet-cell function[69,71,221,265,274]. Many T2D susceptibility loci do not contain strongly associated variants in protein-coding regions, suggesting that the underlying functional variants regulate gene activity. Furthermore, at each locus, most associated SNPs are not expected to directly affect disease risk and are instead in linkage disequilibrium with one or more functional variant(s). We sought to use our open chromatin map to guide identification of functional regulatory SNPs. We identified known SNPs mapping to islet FAIRE sites and focused on 20 loci harboring variants associated with T2D or fasting glycemia (FG)[47,69,221,265,275] (**Figure 5.4a**, and **Table 5.8**). Of 350 SNPs in strong linkage disequilibrium with a reported SNP associated with T2D or FG (**Methods**), 38 SNPs at 10 loci overlapped islet FAIRE regions (**Figure 5.4a**, and **Table 5.8**). Notably, rs7903146 in *TCF7L2*, which

shows consistent T2D association in samples across diverse ethnic groups[276], is located in an islet-selective open chromatin site (**Figure 5.4b,** and **Figure 5.4a**).

The presence of rs7903146 in a FAIRE-enriched site allowed us to test directly whether sequence variation at this locus correlates with chromatin state in islet cells. We tested 31 human islet samples and identified nine individuals heterozygous at rs7903146. Using two independent assays, FAIRE-isolated DNA from heterozygous individuals exhibited a T:C allelic ratio that was significantly greater than observed from input genomic DNA or from genomic DNA from unrelated heterozygote individuals (*Real-time PCR*: input: 49.3±1.0% T allele, FAIRE: 57.3±2.8% T allele, P=2.1x10^{-5}, **Figure 5.4c;** *Quantitative sequencing*: input: 57.5 ± 2.7% T allele, FAIRE: 66.2 ± 4.6% T allele, P=0.004, **Figure 5.4d and Figure 5.8b**). Thus, in human islet cells, the chromatin state at rs7903146 is more open in chromosomes carrying the T allele, which is associated with increased T2D risk[47].

Next, we created allele-specific luciferase reporter constructs and measured enhancer activity in two islet β-cell lines, MIN6 and 832/13. Allelic differences in enhancer activity were observed in both cell lines. The T allele showed significantly greater enhancer activity compared to the C allele in both orientations (Forward: MIN6 P=1.6x10^{-7}, 832/13 P=0.005; Reverse: MIN6 P=3.1x10^{-7}, 832/13 P=3.1x10^{-4}, **Figure 5.4e,f, Figure 5.8c,d**). However, allele-specific differences were not observed in the human embryonic kidney cell line 293T (**Figure 5.8e**). These data suggest that sequence variation at *TCF7L2* affects T2D susceptibility by altering *cis* regulation and local

chromatin structure in islet cells. The results are consistent with a previous report of association between the T allele and increased *TCF7L2* transcripts in islets[276], although the allele-specific changes described here can potentially impact different genomic regulatory functions, including transcriptional rates, promoter usage, or splicing.

Discussion

To our knowledge, this study represents the first high-resolution atlas of regulatory elements in pancreatic islets. The unbiased maps generated by FAIRE-seq reveal new insights regarding the organization of tissue-specific *cis*-regulatory elements. Many earlier studies have shown that the genome is functionally organized in chromosomal territories[89,277-279]. Our observations extend previous findings by uncovering the existence of a large number of cell-selective regulatory domains associated with single genes, and provide a foundation for mechanistic understanding of transcriptional regulation of genes important for pancreatic islet cells. Identification of regulatory sites in a disease-relevant primary tissue also serves to dramatically reduce the genomic space when searching for functional non-coding sequence variants that influence T2D susceptibility. More generally, the current study provides a framework to move forward from the identification of large sets of disease-associated variants toward understanding the subset of functional variants that underlie disease risk.

Contributions

A version of this work has been previously published as: Kyle J Gaulton, Takao Nammo, Lorenzo Pasquali, Jeremy M Simon, Paul G Giresi, Marie P Fogarty, Tami M Panhuis, Piotr Mieczkowski, Antonio Secchi, Domenico Bosco, Tierry Berney, Eduard Montanya, Karen L Mohlke, Jason D Lieb, Jorge Ferrer. A map of open chromatin in human pancreatic islets. *Nat Genet.* 2010 Mar;42(3):255-9.

J.F. and J.D.L. conceived the study. K.J.G., T.N., L.P., J.M.S., K.L.M., J.D.L. and J.F. designed the experiments, interpreted results and wrote the manuscript. T.N. conducted FAIRE experiments, and developed and performed allelic imbalance assays. P.G.G. optimized the FAIRE protocol and performed microarray studies. K.J.G., J.M.S., and P.G.G. performed sequence analysis and K.J.G., L.P., T.N. and J.M.S. performed data analysis. L.P. conducted the analysis of COREs. M.P.F. and T.M.P. conducted reporter assays. P.M. conducted high-throughput sequencing. A.S., D.B., T.B. and E.M. provided purified human islet samples.

Chapter VI

Predicting allele-specific differences in transcription factor binding profiles

Abstract

Recent genome-wide association studies have identified more than 20 susceptibility loci for type 2 diabetes and fasting glycemia, many of which contain a large number of associated variants. Differences in transcriptional regulation are likely responsible for susceptibility at some loci, although the best candidates to directly influence a regulatory element are not often obvious. We thus developed a method to predict SNPs with differences in transcription factor binding. When tested on open chromatin data from pancreatic islets, ROC analysis confirmed that the method can use TFBS predictions to correctly identify islet open chromatin elements more often than expected (AUC=0.811). We assessed allelic differences in islet TFBS predictions using SNPs associated with type 2 diabetes. Ten SNPs had significant differences in predicted TFBS between alleles at $P < 1x10^{-4}$, including rs7903146 in *TCF7L2*, which has been previously demonstrated to have allelic differences in islet enhancer activity. More SNPs were significant at this p-value threshold than expected, suggesting that this approach may enrich for variants directly altering transcriptional regulation.

Introduction

Genome-wide association (GWA) studies have identified many loci harboring risk

variants for numerous traits and common diseases, although the ensuing process of singling out the variant or variants functionally responsible for the association signal is often not straightforward for several reasons. First, the number of variants in moderate or even high LD with an association signal can often be large and span relatively large genomic regions. For example, at a locus on chromosome 12 recently identified to be associated with levels of high-density lipoprotein cholesterol (HDL-C), there are more than 50 associated variants in high LD spanning a region of approximately 200 kb [86]. Second, relatively few loci contain an obvious functional candidate SNP, such as a deleterious non-synonymous or splice-site variant, and even for loci that do, it is possible that multiple variants on risk haplotypes contribute to trait variability [280]. Finally, while variants perturbing transcriptional regulatory elements likely contribute to trait variability at many loci, determining which variants may alter regulatory element activity is often non-trivial.

Relatively new techniques allow genome-wide identification of non-coding regulatory elements through positional identification of molecular hallmarks of regulatory activity, such as nucleosome-depletion [281 235 282], histone tail modification [202 283], and DNA-protein interaction [202 284 92]. This data has helped improve knowledge of transcriptional regulatory elements across many tissues and environmental conditions and can help prioritize non-coding variants. For example, FAIRE-seq data generated in pancreatic islets successfully limited the list of type 2 diabetes (T2D)-associated variants to the 10% most likely to be located in islet transcriptional regulatory elements [282]. One of

these variants, rs7903146, at the *TCF7L2* susceptibility locus was then shown to have allelic differences in both islet nucleosome occupancy and regulatory activity, suggesting that this approach can help identify regulatory variants. However, experimental genomic methods often identify a larger genomic region surrounding the sequence(s) directly necessary for the transcriptional regulatory element. Therefore, genomic data alone is often not sufficient to distinguish *a priori* benign SNPs from those that may influence transcriptional regulation.

Additional and complementary approaches to identify regulatory elements include computational predictions of transcription factor binding sites (TFBS). Sequences known to bind transcription factors can be stored as binding motifs, such as those in the databases JASPAR [98], TRANSFAC [99] and UniPROBE [100], which can be used to directly predict the location of where a transcription factor might bind in additional sequences [97]. One advantage of using TFBS predictions is that they provide more fine-tuned positional information over genomic data, and in several cases have been used to identify SNPs at predicted TFBS that directly alter the binding of that transcription factor [285]. Individual TFBS predictions, however, typically have low sensitivity and specificity[101,286]. More accurate method for regulatory element identification using TFBS have included combining many predictions together to identify regulatory modules often using multiple-species sequence comparisons to identify conserved sites or expression data to find TFBS potentially active in specific tissues [108-111]. These approaches are limited in that TFBS are often not conserved between species and many factors are widely expressed.

The availability of genome-wide maps of open chromatin for cell lines and primary tissue allows in-depth analysis of the regulatory landscape of these tissues. We sought to utilize open chromatin data as a means to identify *cis* regulatory modules of predicted transcription factor binding sites. The approach we developed weights transcription factor motifs based on relative enrichment in open chromatin regions compared to negative regions. Query sequences are then scanned for the presence of motifs and the sequence is scored based on the derived weights of the identified motifs. We then extended our approach to compare differences in sequence scores to help identify SNPs with potential allelic differences in regulatory activity. In principle, this approach can predict not only regulatory SNPs at a trait-associated locus, but also the nature of the functional difference at that SNP.

We tested our approach by training motifs on open chromatin data obtained using FAIRE-seq from pancreatic islets, which are a critical tissue in the pathogenesis of type 2 diabetes (T2D). Cross-validation using an independent set of islet FAIRE elements confirmed the use of weighted motifs as a successful means to identify regulatory sequences. For variant analysis, there was an excess of T2D associated variants with significant allelic differences, including positive control rs7903146, suggesting that this approach may enrich for variants directly altering regulatory element activity.

Methods

Our methodology is comprised of three main parts: motif training, sequence scoring, and allelic sequence comparison.

Motif training

A total of 361 TFBS position weight matrices (PWM, or motif) from TRANSFAC [99] and JASPAR [98] are used in training. An overview of training is presented in **Figure 6.1a**.

Motifs are trained using user-defined sets of positive and negative sequences. Each sequence in both sets is scanned for the presence of a motif using previously developed Perl modules [287], where a binding site is returned if it scores at least 80% of the highest possible motif score. The number of sequences with at least one predicted site for a given motif is counted for both the positive and negative sets, which is used to calculate the percentage of sequences in each set containing the motif. The two percentages are then divided and the log of that number is returned as the 'motif score'. The 'motif score' represents the relative enrichment of the motif in the positive training set compared to the negative training set.

Sequence scoring

Input sequences are scanned for the presence of all 361 motifs using the same threshold as in training (80% of highest possible match).

The resulting set of returned motifs each has a 'motif score' from training. The 'motif scores' for all motifs identified in the input sequence are summed together to provide an

overall ‘sequence score’. All instances of a motif found in a sequence are counted and contribute to the ‘sequence score’

Allelic sequence comparison

For variant analysis, the sequences surrounding each allele of the variant are first analyzed separately to obtain a ‘sequence score’. The score for the first allele is subtracted from the score for the second allele, resulting in the difference in sequence scores between alleles.

To help estimate the significance of an allelic difference, variants are randomly sampled from HapMap and the differences in ‘sequence scores’ are recorded to determine the expected distribution of allelic differences. Input SNP *P*-values are then estimated from the distribution of these scores using the R function pnorm().

Selection of training and validation sets for pancreatic islet open chromatin

For training, the positive set consisted of all 1,818 islet FAIRE sites (sample 3) in the ENCODE pilot project regions, and the negative training set was 1,899 contiguous 36-bp mappable regions with no islet FAIRE signal in sample 3 (**Figure 6.1a**). The negative training set regions were size-matched to the positive regions.

For validation, positive regions were 1,489 islet FAIRE sites on chromosome 10 found in all three islet samples, and true negative regions were size-matched contiguous 36-bp

mappable regions on chromosome 10 for which there was no signal in any of the three islet samples.

Selection of T2D associated SNPs

We identified 20 published variants associated with either type 2 diabetes or fasting glucose level (**Table 5.7**). Using data from the 1000 Genomes project (www.1000genomes.org; August freeze), 899 variants were in high LD (r^2>.8) with a published susceptibility variant, and were thus considered the set of associated variants.

Availability

The Perl source code can be downloaded from http://polaris.med.unc.edu/projects/TFBS/.

Results

Motif training and sequence validation using pancreatic islet open chromatin

We trained 361 TFBS motifs using open chromatin data generated in pancreatic islets (see Methods; **Figure 6.1a**) [211]. Training sets were selected from data in regions from the pilot ENCODE project. The positive training set consisted of 1,818 islet FAIRE sites, and the negative training set was 1,899 contiguous regions with no islet FAIRE signal. The motifs with the highest and lowest 'motif scores' from training are listed in **Table 6.1**.

We attempted to validate the training set results using an independent set of positive and negative sites (see Methods). For each site in the validation sets, a 'sequence score' was

obtained that represents the sum of 'motif scores' identified in that sequence (see Methods). A Receiver Operating Characteristic (ROC) was generated from the 'sequence scores', and the area under the curve (AUC) was .811 (**Figure 6.1b**). From ROC analysis, the optimal sensitivity and specificity were 73% and 86%, respectively.

Identifying type 2 diabetes associated variants with differences in islet sequence scores

We sought to determine whether our approach could help prioritize between disease-associated non-coding variants by identifying those with differences in 'sequence scores'.

We identified 899 SNPs from the 1000 Genomes project in LD with published T2D-associated variants (**Table 5.7**). For each SNP, we obtained the 'sequence score' for the 200 bp region surrounding each allele. We then calculated the difference in scores between alleles, and assessed the significance of this difference by comparing to the expected distribution (see Methods; **Figure 6.2a,b**). Of the 899 tested SNPs, 10 were significant at $p<1x10^{-4}$ (**Table 6.2**), which is greater than expected by chance (expected = .09; **Figure 6.2c**). Of the 10 SNPs significant at $p<1x10^{-4}$, eight distinct loci are represented, suggesting possible functional targets at these loci.

Characteristics of variants with significant differences in sequence scores

We next assessed the *in vivo* relevance of the genomic region surrounding SNPs with significant differences in 'sequence scores' by identifying those that overlapped FAIRE sites identified in pancreatic islets. Three of the 10 SNPs co-localized with a FAIRE site from at least one sample (**Table 6.2**), including rs7903146 in *TCF7L2,* which has been

previously demonstrated to have islet-specific differences in enhancer activity (**Figure 5.4**). We further investigated the properties of TFBS surrounding each allele (T/C) of rs7903146. The T allele, which has greater islet regulatory activity, had both a greater number of allele-specific factors (27, compared to 8 C-specific factors) as well as several high scoring factors (LHX3, FOXD3; **Table 6.1**).

We next looked for patterns in TFBS predictions across all 10 SNPs to determine if they shared features in common. A majority of SNPs overlapped predicted binding sites for FOXD3, PDX1 and/or HNF3B, transcription factors with known involvement in islet regulation[288-290]. These SNPs thus make attractive targets for future functional study.

Discussion

Our approach has two potential advantages over using *in vivo* regulatory data alone to prioritize potential regulatory variants. First, using TFBS allows the direct identification of those SNPs likely to alter transcriptional regulation while filtering out those that are benign. Second, by using sequence information, the method can be unbiased towards the incompleteness of the *in vivo* regulatory element catalog, which thus far is limited to data generated in a small number of tissues, genotypes and environmental conditions.

Still, there are several disadvantages to this approach. First, the catalog of transcription factors with known binding motifs is still a small percent of all factors, and thus the qualities of predictions are limited to those factors. Second, even when a factor has a known binding motif, the models used to build the motif often use very few known

binding sites for the protein, and the small number of factors for which genome-wide ChIP-seq data is available has almost invariably refined the original motif [291]. Using data from the UniPROBE resource, which uses an array-based approach to identify binding sites in an unbiased manner, may somewhat circumvent this problem, although the accuracy of these binding site motifs *in vivo* is unclear. Third, many transcription factors are part of large protein families that have very similar binding sites. Thus, identifying a motif for one protein might be more relevant to a closely related protein, which may have a completely different transcriptional profile. Finally, while nucleosome occupancy data is assumed to be an unbiased survey of the regulatory activity in a given tissue, biases likely exist in the types of and frequency at which transcription factor binding events are identified in every technique, and therefore likely to bias the results of TFBS profiling based on this data. For example, FAIRE and DNase HS, both techniques to identify nucleosome depleted regions, identify only a partially overlapping set of elements [235]. A less biased approach might group several datasets from different assays together. Despite these limitations, using a sequence classification approach to find SNPs with allelic differences in islet regulatory profiles successfully identified a subset of T2D-associated variants with significant differences, including rs7903146, which has been shown previously to have allelic differences in islet enhancer activity [282].

The utility of this approach may also extend beyond predicting SNPs with differential regulatory activity. For example, once a SNP has been identified and tested for regulatory activity, the TFBS output for each allele can be a useful tool to help prioritize further study to narrow in on the specific factor(s) causing the effect. Furthermore,

simply the knowledge of the relative importance of transcription factor binding between two datasets could be useful. In our example, the output of training could be used to identify potentially novel factors important for islet transcriptional regulation.

Many computational tools exist that provide information about the co-localization of a SNP with annotations from a variety of sources [292-298], and can assist manual selection of SNPs in likely functional regions. However, while many tools predict SNPs directly affecting protein products, few if any predict how a SNP might affect a non-coding regulatory element. Sorting out variants likely to alter regulatory activity from those merely co-localizing with a regulatory element will become even more critical as functional genomics projects annotate a larger percentage of the non-coding genome and millions of non-coding variants are identified in re-sequencing projects. Our approach attempts to provide this information in the form of variants with differences in predicted transcription factor binding. When coupled with other annotation tools, it could be a useful resource assisting projects aimed at identifying functional non-coding variants.

Contributions

Kyle J Gaulton and Karen L Mohlke designed the study, interpreted the results and wrote the manuscript. KJG performed the data analysis.

Chapter VII

Conclusions and future directions

In this work we have used a series of candidate-based approaches to prioritize and test variable regions of the genome for possible direct or indirect involvement in complex trait susceptibility, in particular type 2 diabetes. The scope of these approaches ranged from the full genome, by using CAESAR to narrow down a list of candidate genes from all known genes, to genes, by selecting SNPs in type 2 diabetes candidate genes for genotyping and lipid level-associated genomic regions for re-sequencing, and loci, by selecting T2D-associated variants for functional study based on pancreatic islet FAIRE-seq data, down to the individual variant level, by predicting transcription factors differentially bound to variant alleles.

Our first approach, CAESAR, represented a novel method that combined text- and data-mining to select candidate genes for complex traits based on the known biology of a trait. CAESAR was able to prioritize several published susceptibility genes for complex traits, suggesting that it was a valid metric to potentially prioritize novel susceptibility genes. We next used CAESAR to help select candidate genes for type 2 diabetes, selected tag variation to capture the majority of common variation in these genes, and genotyped selected variants in Finnish samples from the FUSION study. Using the approach we identified common variants associated with type 2 diabetes and related quantitative traits

near these candidate genes that may influence trait susceptibility. We then selected genomic regions harboring variants associated with lipid levels for re-sequencing in trait-extreme individuals, and identified additional common, less common and rare variants that may contribute to trait variability. Finally, we used several approaches to prioritize between trait-associated variants to identify functional variants influencing transcriptional regulation, including identifying T2D-associated variants in regions of pancreatic islet open chromatin identified by FAIRE-seq and identifying variants with allelic differences in predicted transcription factor binding sites.

The common goal of all of these approaches was to attempt to intelligently reduce the space in which we interrogate the genome. An obvious limitation to any candidate-based approach to finding susceptibility variants is that the success of the ensuing study relies on the quality of predictions from the candidate selection process. However, as is it is typically not cost or time effective to interrogate all variation of interest, candidate selection is often a necessary aspect of studies designed to identify variation influencing disease risk. Even as technology increases the throughput by which genetic and genomic information can be generated, candidate selection metrics will still be needed at remaining cost or time bottlenecks, and therefore will continue to be an integral part of medical research.

For example, while full genome sequencing will soon be economically feasible for large scale genetic study, gene selection methods such as CAESAR will still be needed to help identify targets for follow-up genotyping and functional study. Since the initial testing

of CAESAR, a much larger set of susceptibility loci have been identified for the complex traits used to test our algorithm. In addition, some of the published loci used for the initial tests identified in small studies have since not replicated across additional studies, and therefore may represent false-positives. Therefore, we followed up on the results of CAESAR by comparing the published gene rankings with a current list of genome-wide significant loci ($P < 5x10^{-8}$) available from the GWAs catalog. As the functional basis for the majority of loci is not known, we included all genes in cases when multiple genes are reported.

To facilitate comparison with our initial results (**Table 2.2**), successfully prioritized genes were defined as those ranked in the top 2% of all ranked genes for a given trait using the same gene lists. Of the 150 genome-wide significant loci, 27 (18%) were ranked in the top 2% for the respective trait, a 9-fold increase over the number expected by chance (exp = 3). This supports our initial conclusions that CAESAR can successfully prioritize susceptibility genes, and with continued development may be a useful tool to assist in the follow-up of next-generation genetic studies.

While genome-wide association studies have identified numerous common variants influencing trait values and disease susceptibility, the amount that the identified variants contribute to heritability is relatively low [28]. Therefore, a large focus of the next phase of human genetic research is uncovering missing heritability, which is likely in the form of less common (MAF<.05) and rare variants (MAF<.005) that have yet to be interrogated on a genome-wide scale [28]. The 1000 Genomes project is performing low-pass

sequencing (4x coverage) of a large number of samples. If data from that project can effectively identify LD patterns in less common variants (.005<MAF<.05), it would allow them to be imputed into existing GWA frameworks and tested for association. Uncovering extremely rare variants not identified in the 1000 Genomes Projects will require targeted medical re-sequencing efforts. Given rapid advances in sequencing technology and constantly lowering costs, projects of 1000 Genomes scope on medically-relevant samples should soon be able to answer many questions about the topography of the genetic landscape for complex disease.

While protein-coding variants are an important contributor to trait variability, the majority of variants are located in non-coding regions, and variants perturbing transcriptional regulation likely contribute a large amount to trait variability [280 88]. Therefore, understanding how variants shape non-coding elements is critical to understand the genetic component of complex trait variability. Until recently, however, annotation of the non-coding portion of the genome was extremely sparse. Relatively new techniques allowing genome-wide identification of non-coding regulatory elements through positional identification of molecular hallmarks of regulatory activity, such as nucleosome-depletion, histone tail modification, and transcription factor binding, have improved our knowledge of transcriptional regulatory elements, but the genome is still vastly under-annotated across tissues, environmental conditions, and disease states. Further, how variants may influence activity of these regulatory elements is also poorly understood.

Potential approaches to identifying non-coding variants with a functional effect on disease susceptibility and trait variability likely depend on the allele frequency of variants being analyzed. For variants that are common, an extension of our previous study of FAIRE-seq would be to correlate variants to a quantitative variable describing a regulatory element, for example, sequence read density from FAIRE-seq (or ChIP-seq, DNase-seq) experiments, in the same manner as expression QTL studies. This approach would require both genetic and genomic data on a large number of samples to have sufficient power to detect significant associations. Alternately, sequence reads from genomics experiments could be used directly for genotyping to identify variants with allelic imbalance in read density, though this requires extremely deep sequencing. These approaches would be most powerful using samples under standardized conditions to remove confounding non-genetic variation. One particularly enticing possibility for cell types that are hard to obtain such as pancreatic islets would be to use emerging stem cell technology to artificially create cell types of interest from more readily available cell types collected on a larger number of samples [299].

Statistical power to detect significant, single rare variant associations using QTL-based methods is low. Given medical re-sequencing of thousands of samples, tens of millions of rare variants will be discovered and methods could be developed to exploit the grouping of variants in non-coding elements in statistically meaningful ways. Our analysis of re-sequencing data from several hundred individuals at the extremes of lipid level traits suggests this will be fruitful. Similar to studies profiling differences in number of coding

changes between individuals in trait-extreme groups (high or low trait value) or different disease states (case or control), variants in annotated regulatory elements may be preferentially found in one extreme trait group. As variants perturbing the same regulatory element(s) could feasibly influence trait values in different directions, the same analysis could be performed grouping trait-extreme individuals together and compare against individuals in the middle of the trait distribution. Yet another approach could be to correlate the number of regulatory variants along a trait value gradient. Alternate approaches could function independent of genome annotation; for example, clustering methods to identify rare variants in trait-extreme individuals that group together at the sites of critical non-coding elements.

A critical component of identifying non-coding variation influencing transcriptional regulation is determining the *cis* targets of these elements. Emerging, high-throughput techniques based on chromatin conformation capture (3C) [300], such as 5C [301], 6C [302], and Hi-C [303], have started to successfully couple distal regulatory elements to their transcriptional targets, although the volume of multi-dimensional data produced by these techniques presents a major computational challenge.

As the above approaches do not necessarily identify the mechanistic basis for how variants directly influence transcriptional regulation, a supplementary component could be the continued development of computational methods for sequence analysis in the context of variants; for example, the use of transcription factor binding site motifs to

predict how a variant will change factor binding. Such methodologies could be used for *a priori* identification of interesting regulatory variants, regardless of allele frequency, as well as to place variants of already known interest in biological context. Further, combining the above-mentioned chromatin capture data with TFBS prediction and protein-protein interaction data might allow the development of *in silico* models for how and where regions of the genome interact, and facilitate both the reconstruction of transcriptional regulatory networks and the degree to which variants might affect the activity of these networks.

Ultimately, research projects of this type will help lead to a detailed understanding of how variants influence transcriptional regulation, which belongs in the larger context of the full spectrum of genomic variants with allelic differences in biological function that influence trait variability and disease susceptibility. The impending availability of genome sequencing on an individual basis makes large-scale classification of functional variation an immediately relevant problem in medical research, and answering these research questions that has the potential to broadly influence human health. The utility of such knowledge will be a critical component in the eventual development of both predictive measures for genetic disposition to complex traits and diseases as well as the field of personalized medicine.

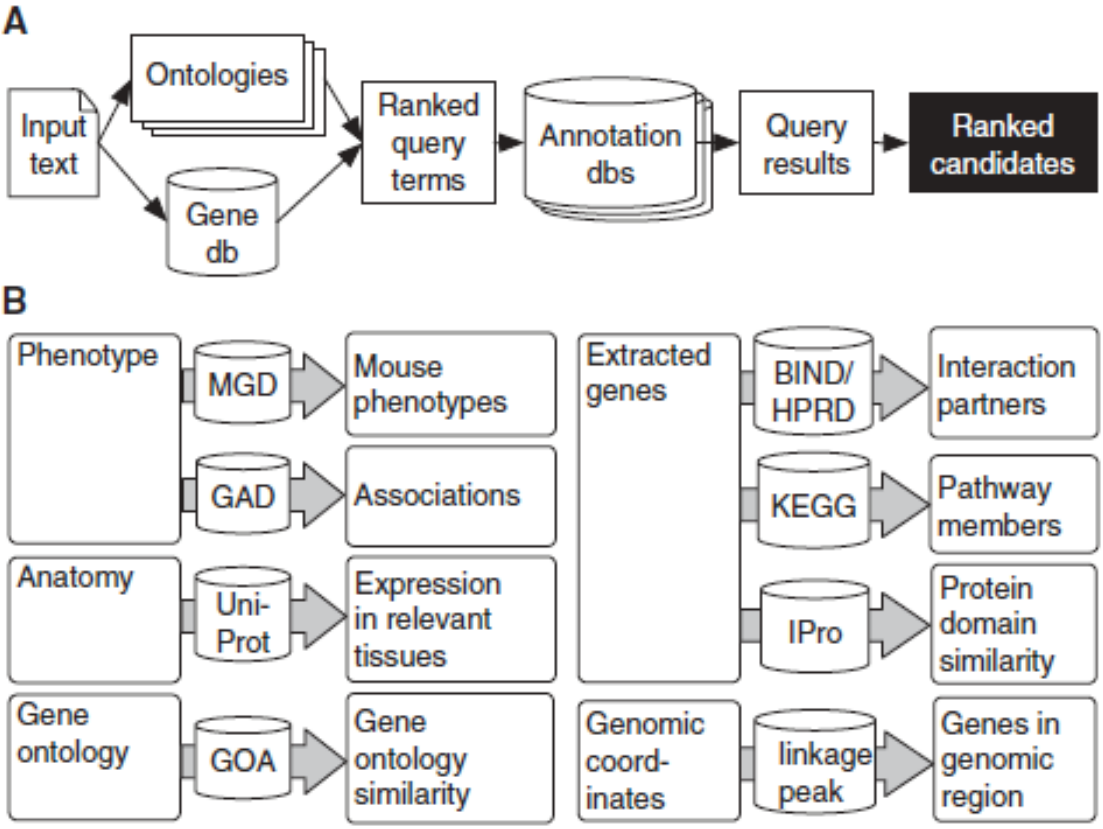


Figure 2.1. CAESAR overview. (A) Text mining is used to extract gene symbols and ontology terms from the input. In the data-mining step, genes within each gene-centric data source are ranked based on the relevance to the trait-centric terms. In the data-integration step, the results from each source are combined into a single ranked list of candidates. Db=database. (B) Eight types of functional information (GO molecular function and biological process listed together) are queried using extracted genes and anatomy, phenotype and gene ontology terms. Genomic regions of interest represent optional user input. See text for abbreviations.

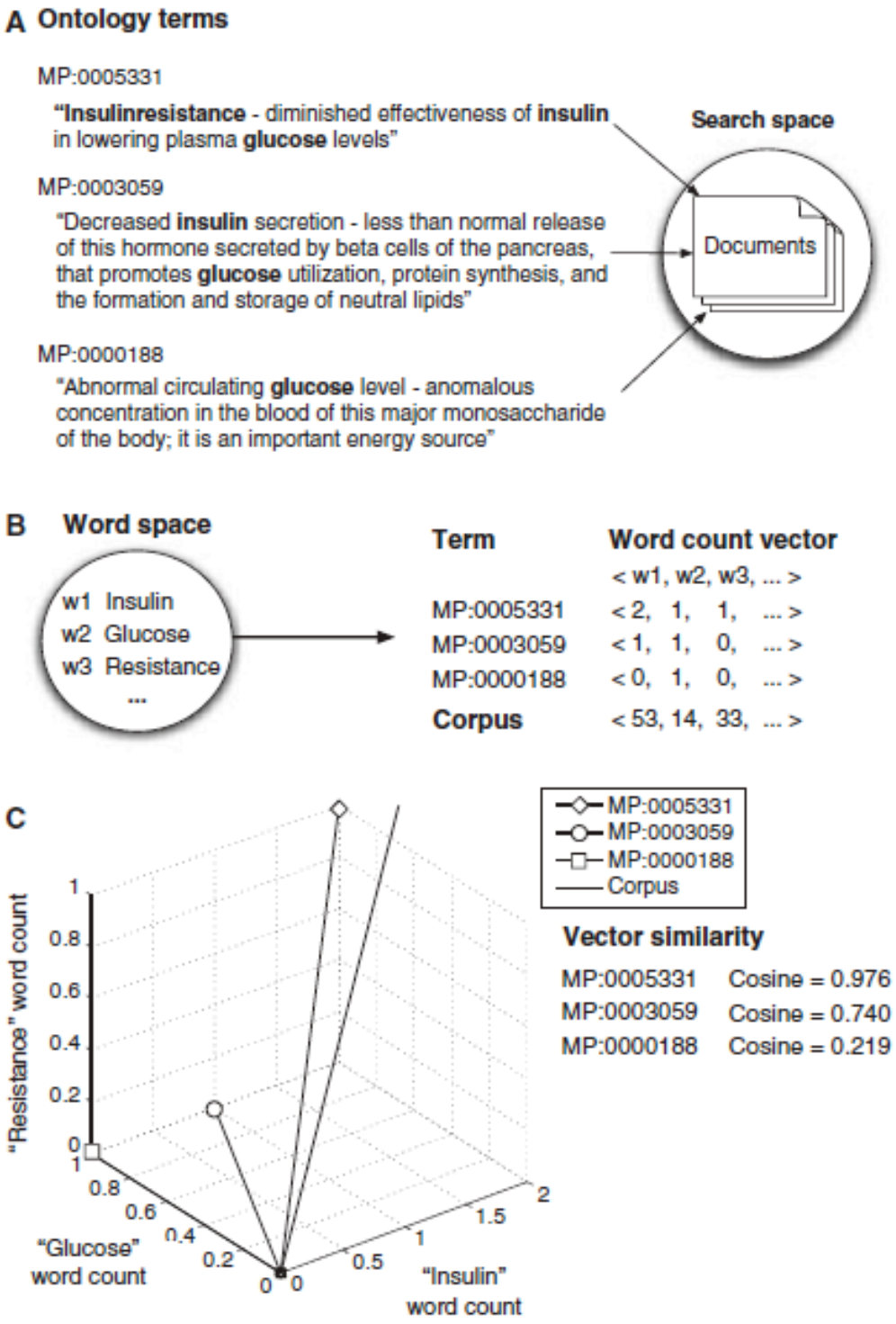


Figure 2.2. Vector-space similarity search. (A) Each ontology term and its description comprise a document, as in this example of three terms from the mammalian phenotype

ontology. (B) The word space consists of all unique words. For illustration, here the word space is (‘insulin’, ‘resistance’, ‘glucose’). Each document, including the corpus, describes a vector in word space, where the elements of the vector are weighted counts within the document of each word in the word space. (C) The similarity of each of the documents to the corpus is measured as the cosine of angle formed by the document and corpus vectors. High-ranking ontology terms have document vectors that are similar in both direction and magnitude to the corpus vector. In this example, MP:0005331 is the highest-ranking document.

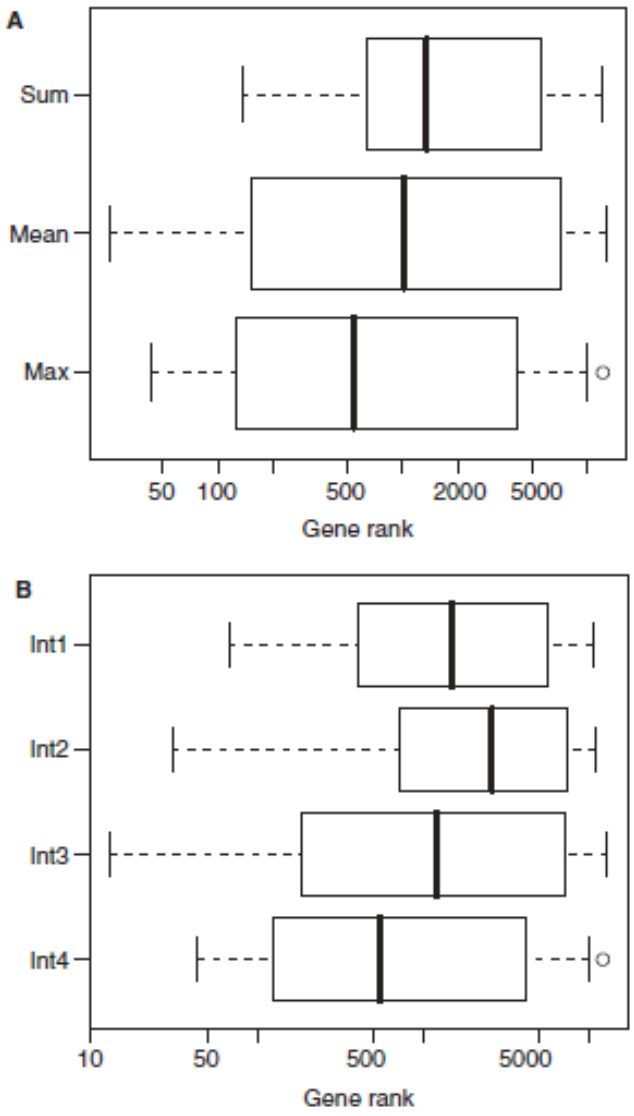


Figure 2.3. Box and whisker plot distributions of the ranks of 18 test genes in Table 2 using different CAESAR parameters. Ranks are plotted on a log scale. Plots are constructed so that the bounds of the box are the upper and lower quartile medians, the line inside the box is the median, the whiskers extend to the last value no more than 1.5 times the length of the box, and all remaining values are outliers. (A) Distribution of ranks using the max, mean and average data-mining methods (int4 method for integration). (B) Distribution of ranks using the four different integration methods (max data-mining method).

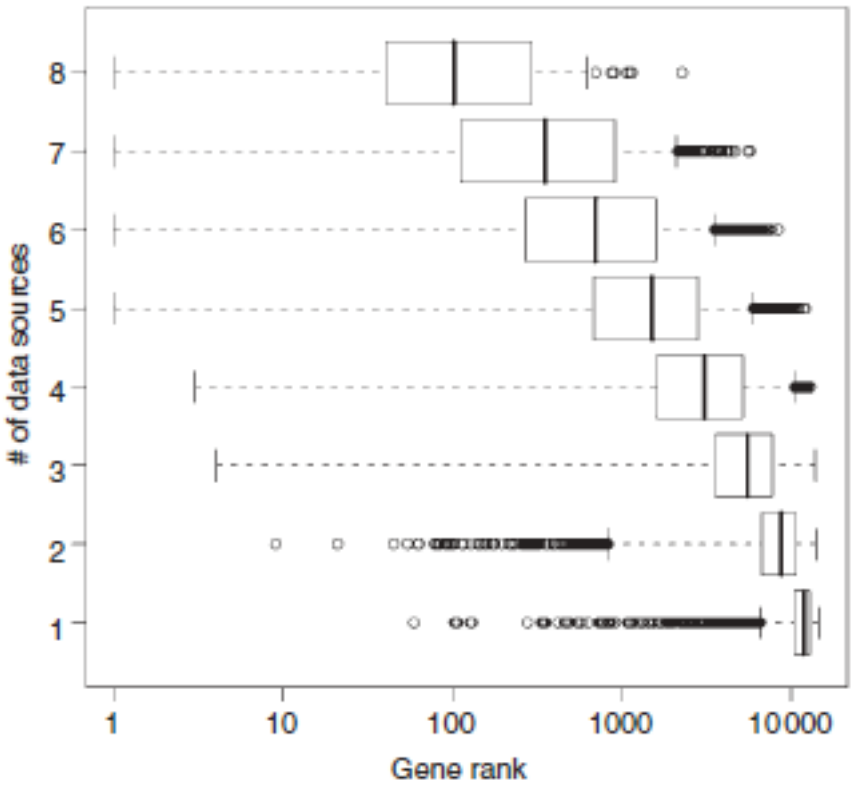
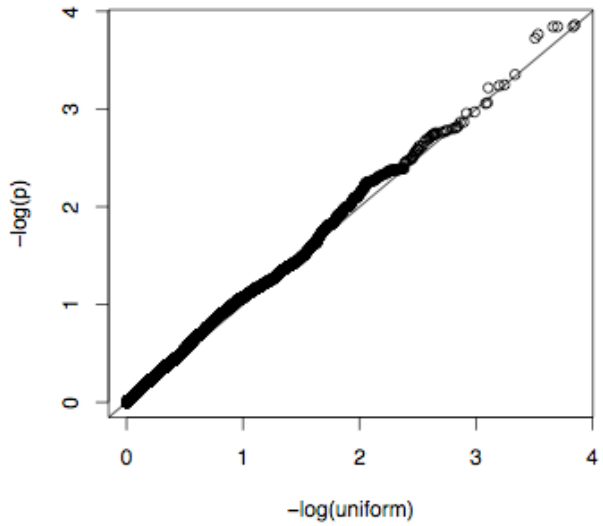


Figure 2.4. The relationship between the rank of a gene and the number of data sources in which it is annotated, using the max and int4 methods. Ranks are plotted on a log scale. Box and whisker plots were constructed as described for Figure 3.

A.



B.

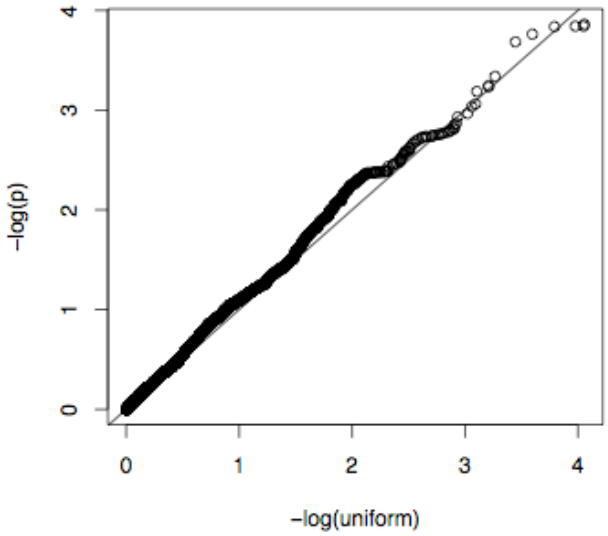


Figure 3.1. (A) Quantile-quantile plot for all genotyped and imputed SNPs comparing -$\log_{10}$ p-values (p_{add} and p_{impute}) for Stage 1 samples with p-values expected under the null distribution. (B) Quantile-quantile plot after removing SNPs corresponding to known susceptibility genes *PPARG, KCNJ11/ABCC8, HHEX,* and *WFS1*.

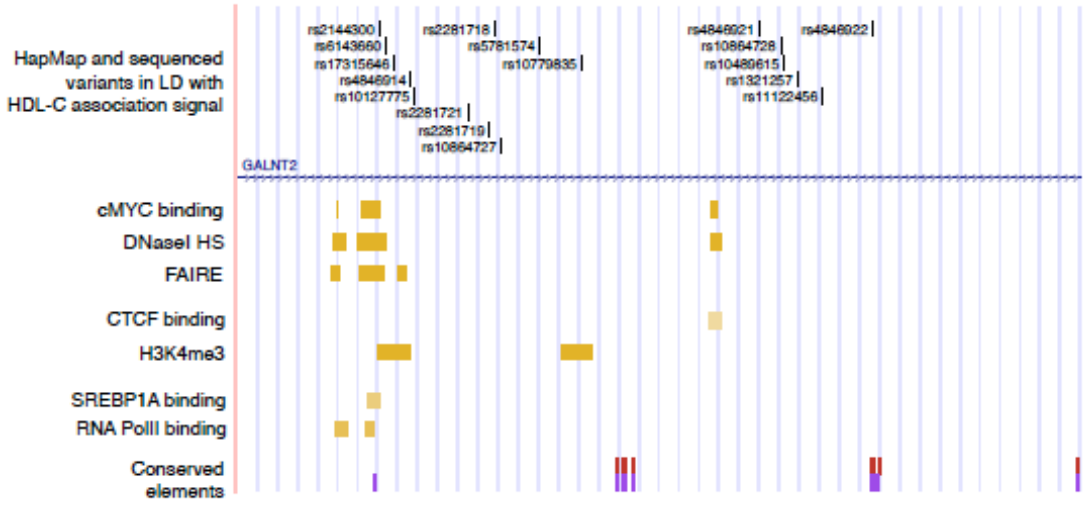


Figure 4.1. Non-coding annotation of variants in LD with HDL-C association signal in *GALNT2* intron 1. Eight HapMap variants were in $r^2 > .8$ with HDL-C associated variant rs2144300 ('HapMap associated' variants), and nine sequenced variants were in $r^2 > .8$ with a HapMap associated variant. Of these 17 variants, five are in a region containing several non-coding annotations generated in HepG2 cells (see Table 3.6 for annotation descriptions).

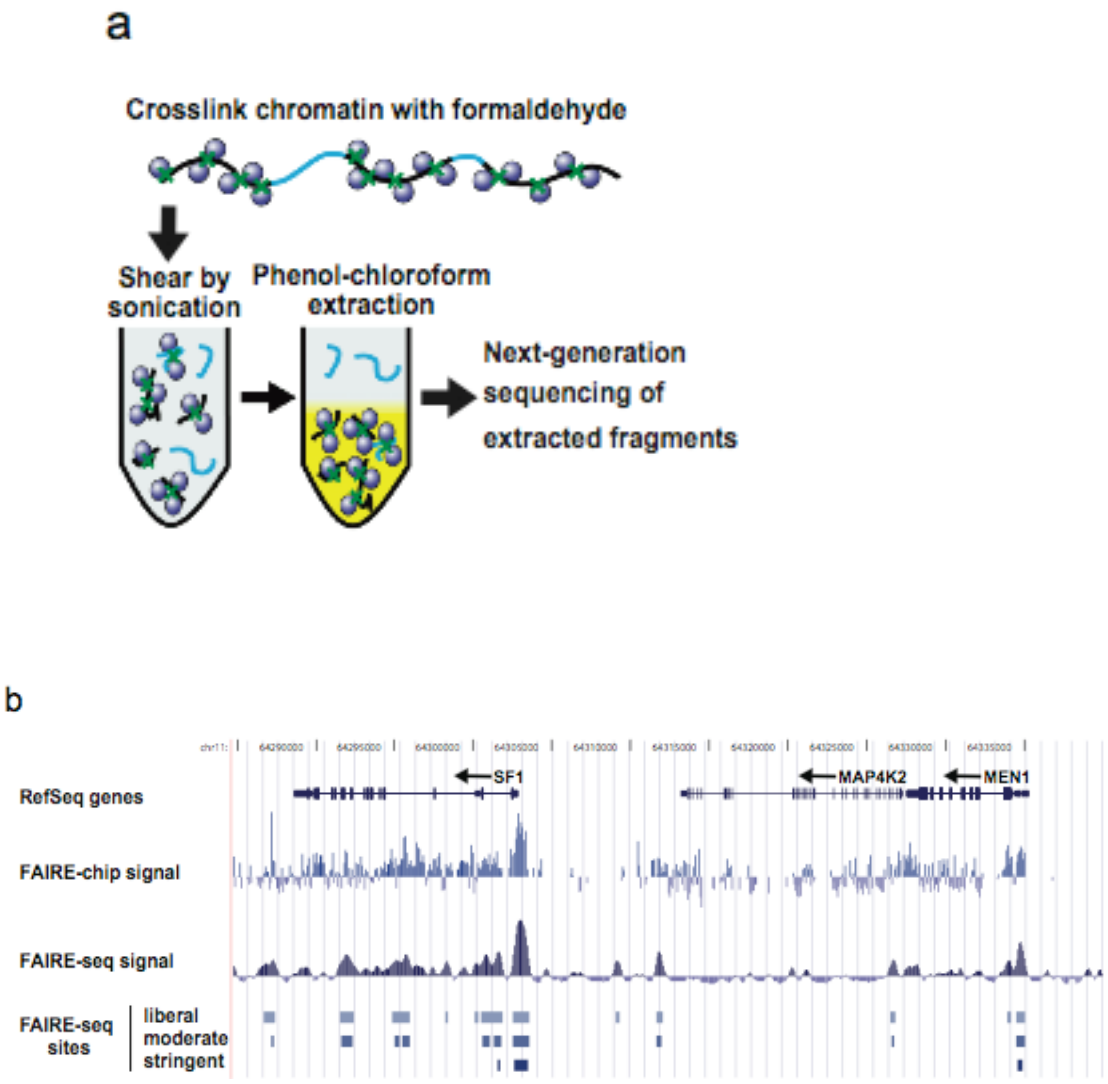


Figure 5.1. FAIRE-seq in human pancreatic islets. (a) Chromatin is cross-linked using formaldehyde, sonicated, and subjected to phenol-chloroform extraction. DNA fragments recovered in the aqueous phase are then sequenced. **(b)** Reads obtained from sequencing were highly concordant with FAIRE signal obtained from tiling microarrays covering the ENCODE pilot project regions. Arrows indicate the direction of gene transcription.

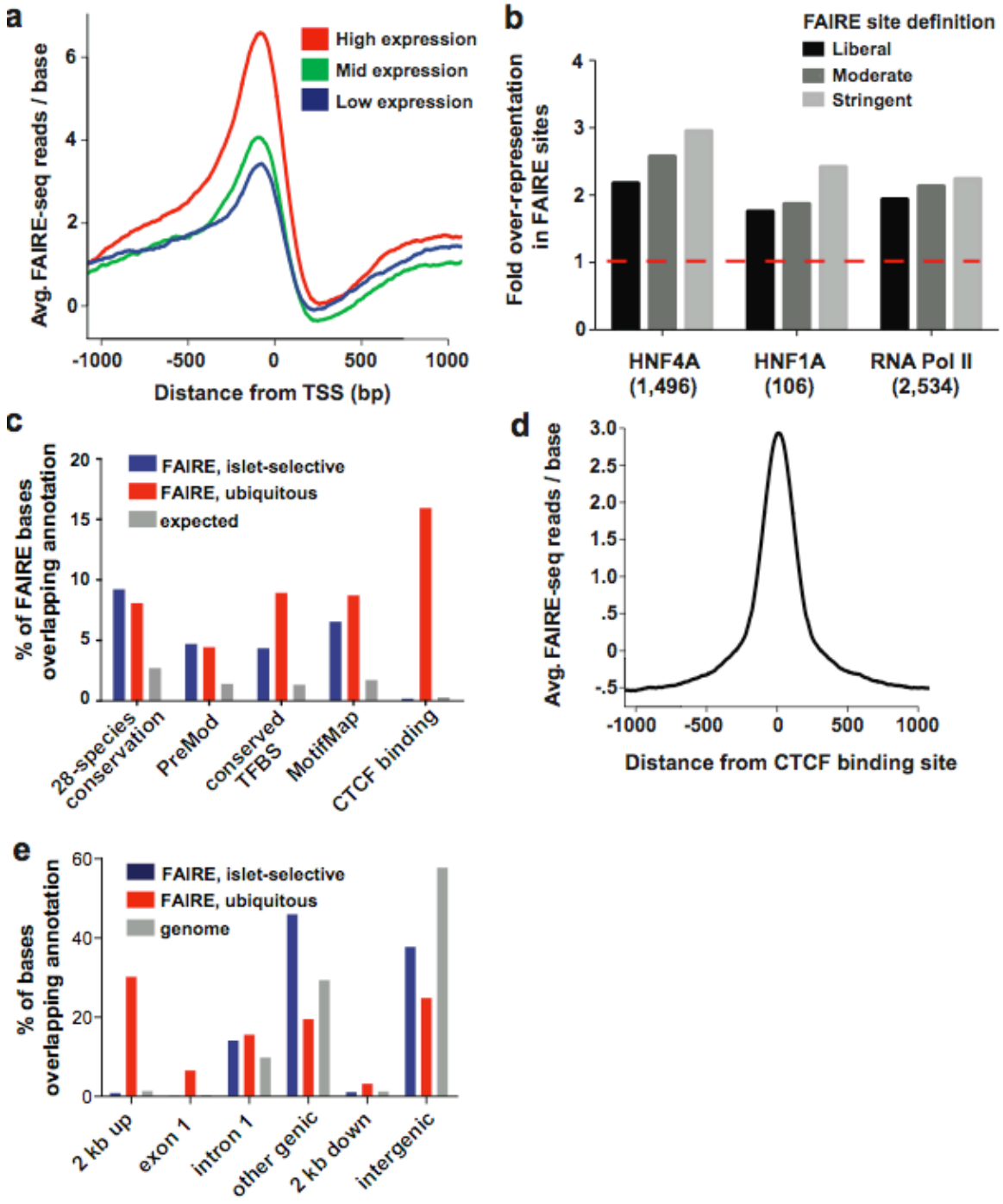


Figure 5.2. Both proximal and distant FAIRE sites harbor functional regulatory elements.

(a) Genes with high expression in islets (top 20%; red) have more FAIRE enrichment at promoters than genes with moderate (middle 20%; green) or low (bottom 20%; blue)

expression. **(b)** Promoters (-750/+250 bp) bound by RNA Pol II, HNF4A or HNF1A in human islets[261] are significantly over-represented among islet FAIRE sites (red dash indicates expected value; all bars: $P<0.001$). **(c)** Intergenic islet-selective and ubiquitous FAIRE sites that are located >2 kb from a TSS are enriched for evolutionary conserved sequences ($P<0.001$), predicted regulatory modules (PreMod, $P<0.001$), and transcription factor binding sites (conserved TFBS and MotifMap, both $P<0.001$). CTCF binding, however, is enriched in ubiquitous FAIRE sites only. Over half of intergenic open chromatin sites are coincident with an experimentally or computationally determined functional annotation (expected value for random sites: 27%). **(d)** Open chromatin is most enriched directly at sites of experimentally determined CTCF binding. **(e)** In contrast to ubiquitous FAIRE sites, islet-selective FAIRE sites are rarely located within 2 kb upstream of a TSS or in exon 1, and are instead located predominantly in more distal regions. Shown is the percentage of bases covered by each annotation category in islet-selective FAIRE sites (blue), ubiquitous FAIRE sites (red), and the mappable genome (gray).

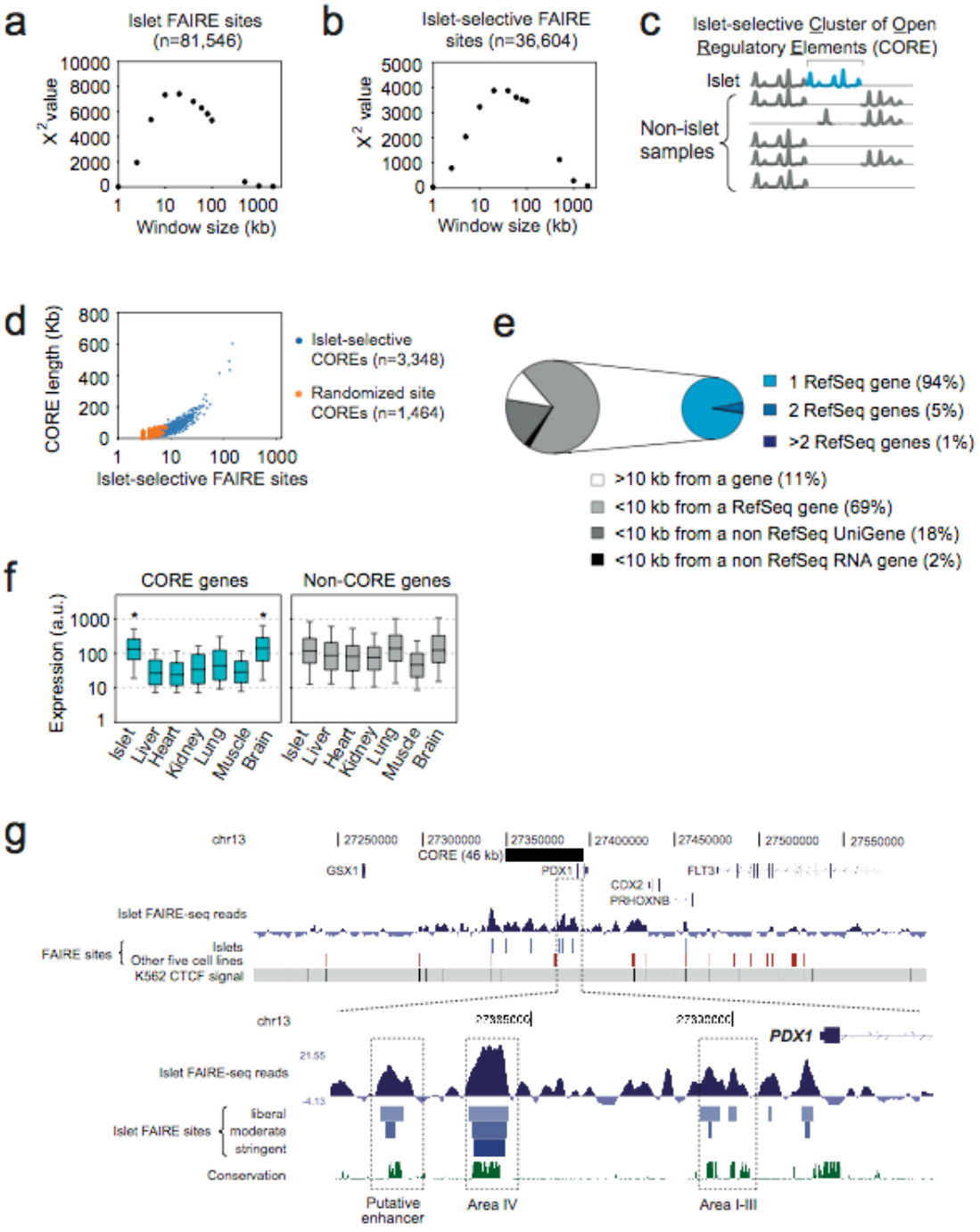


Figure 5.3. Islet-selective FAIRE sites form Clusters of Open Regulatory Elements (COREs)

(**a**) FAIRE sites are highly clustered. We divided the genome in windows of varying size (x axis), and calculated a $\chi 2$ statistic to determine if the number of windows with 0, 1 or >1 FAIRE sites differed from randomly distributed sites. The highest significance was observed in ~20 kb windows. **(b)** Same as (a) but for islet-selective sites. **(c)** We defined islet-selective clusters of open chromatin regulatory elements (COREs) as three or more islet-selective FAIRE sites separated from each other by <20 kb. **(d)** We identified 3,348 islet-selective COREs (blue points). Fewer COREs were generated using randomized FAIRE sites (orange points), and they were smaller than *in vivo* COREs **(e)** Most islet-selective COREs were associated with a single gene. **(f)** RefSeq genes associated with islet-selective COREs were on average inactive in non-islet human tissues, except for brain. Asterisks indicate $P<1x10^{-5}$ (one-way ANOVA). **(g)** Chromatin landscape of the *PDX1* locus showing an extended cluster of islet-selective FAIRE sites, contrasting with a closed conformation of the adjacent gut-specific homeodomain gene *CDX2*. The top panel depicts the density of FAIRE-Seq reads centered on the genomic average density value, the location of moderate stringency FAIRE sites in islets (blue) or in any of the 5 non-islet cells (red), and the binding sites of the CTCF insulator protein in K562 cells. CTCF sites demarcate regions that show broadly consistent FAIRE-Seq enrichment patterns. The bottom panel shows a closer view of a portion of the *PDX1* islet-selective CORE, with islet-selective open chromatin sites at previously characterized regulatory elements (Area I-III, Area IV) and in an evolutionarily conserved putative enhancer.

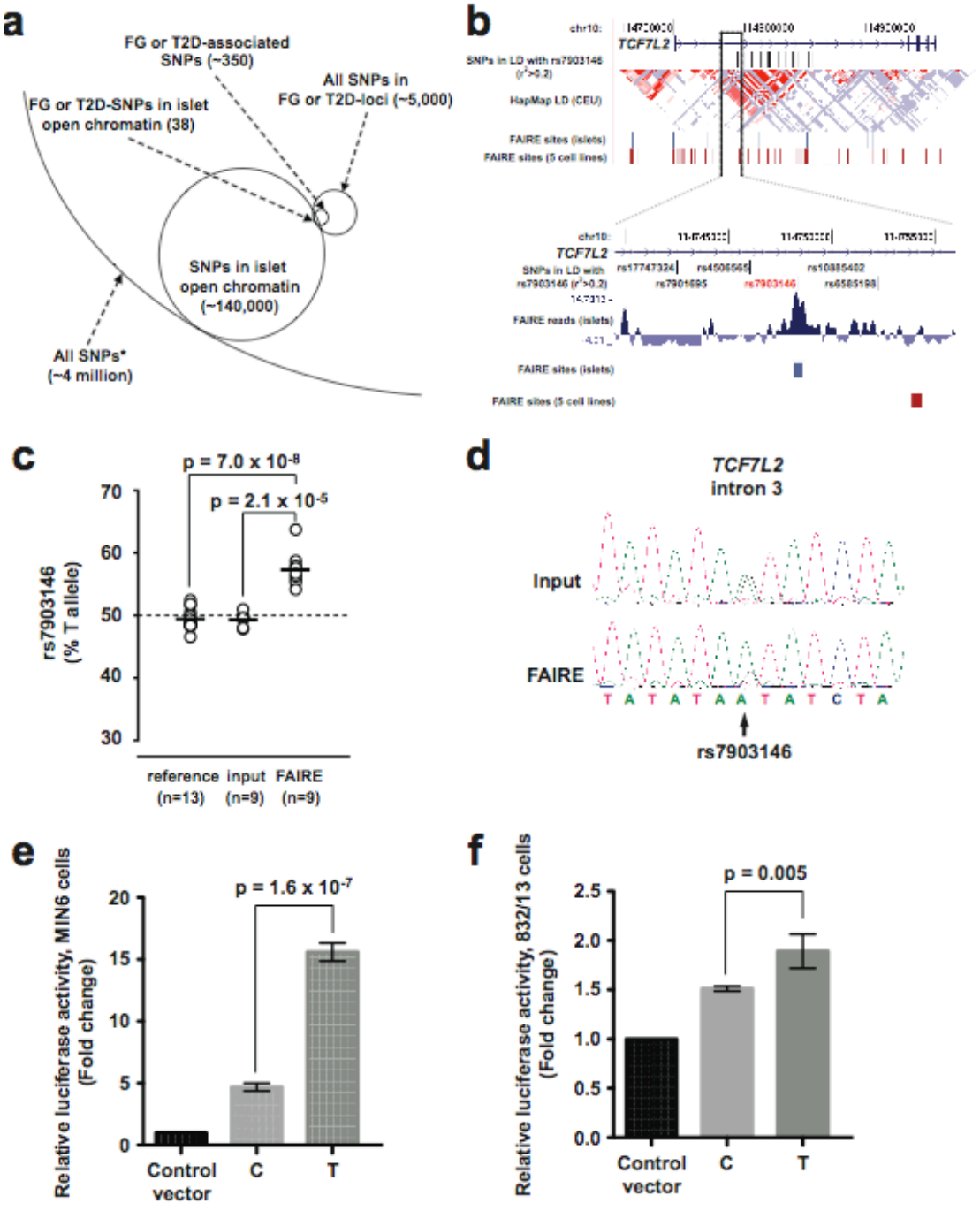


Figure 5.4. Allele-specific open chromatin and enhancer activity at the *TCF7L2* locus.

(a) Schematic representation of how FAIRE-seq enables the identification of human sequence variants located in islet open chromatin. From ~4 million SNPs present in dbSNP with average heterozygosity >1%, 38 SNPs associated with T2D or fasting glycemia mapped to islet open chromatin sites. The analysis was carried out with all SNPs in strong linkage disequilibrium (r^2>0.8) with an FG- or T2D-associated variant, which are labeled as FG or T2D SNPs, and FAIRE-seq sites identified with a liberal threshold. **(b)** Among *TCF7L2* variants in linkage disequilibrium with rs7903146 (r^2>0.2, top panel), only rs7903146 maps to an islet-selective FAIRE site. **(c)** In all 9 human islet samples that were heterozygous for rs7903146, the risk allele T was more abundant than the non-risk C allele in the open chromatin fraction, in contrast to input DNA or gDNA from unrelated heterozygous individuals. **(d)** Allelic imbalance for open chromatin at rs7903146 was verified in independent assays using quantitative Sanger sequencing (see also **Supplementary Fig. 4b**). **(e)** The risk allele T of rs7903146 exhibits greater enhancer activity than the non-risk allele C in MIN6 cells and **(f)** 832/13 cells. Standard deviations represent four independent clones for each allele. Results for inserts in the reverse direction are provided in **Supplementary Fig. 4.** *P*-values were calculated by two-sided *t*-test.

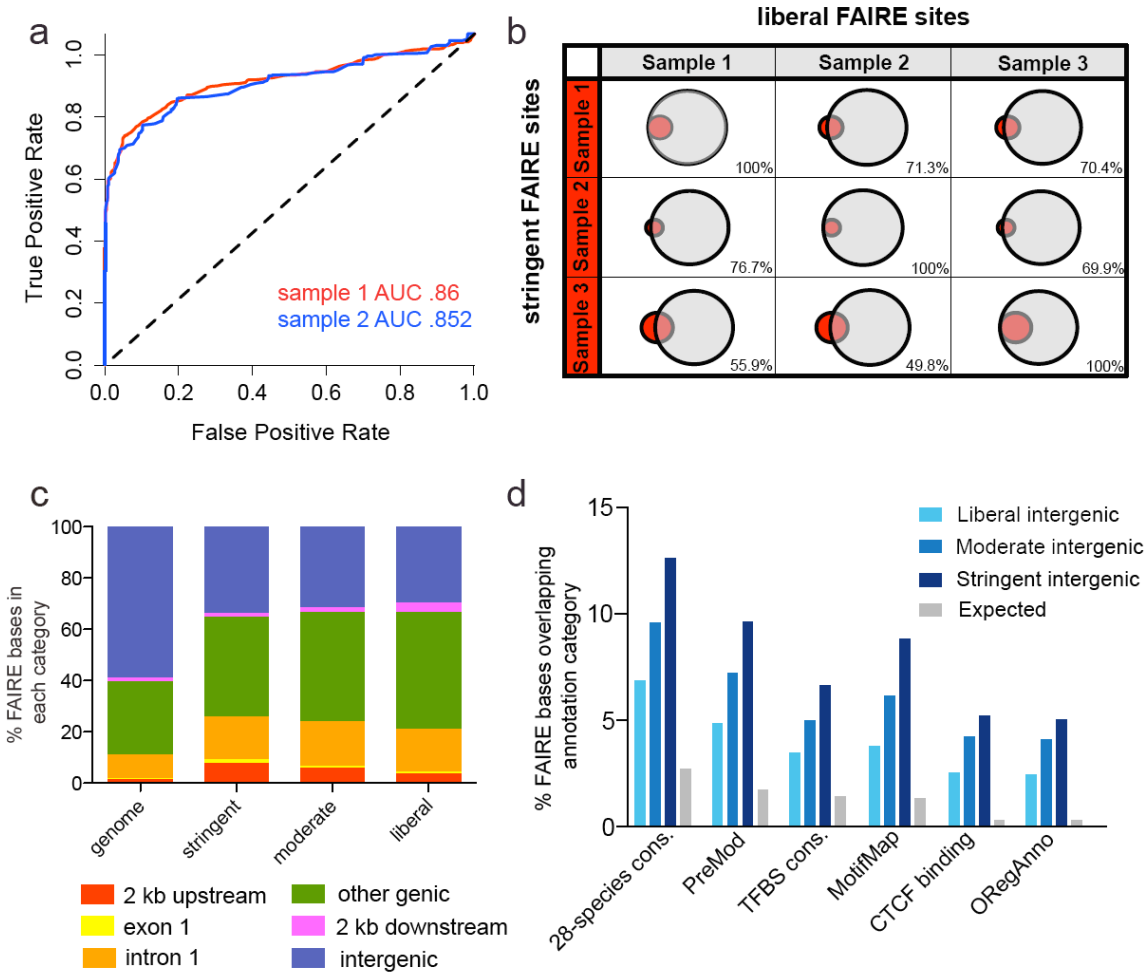


Figure 5.5. Characteristics of FAIRE-Seq in human pancreatic islets. (a) Comparison of FAIRE-chip and FAIRE-seq. Receiver Operating Characteristic (ROC) curve analysis using positive regions called from ENCODE tiling DNA microarrays at a stringent threshold ($P<1x10^{-12}$) for both samples 1 and 2, and negative regions called from contiguous regions of low intensity. The percentage of positive and negative regions captured at increasing sequence read density thresholds was then plotted. The sequence

data from both samples 1 and 2 captures positive tiling array data while rejecting negative tiling array data at a much higher rate than random. (b) Comparison of FAIRE-seq regions across three samples. Stringent islet FAIRE sites (in columns) from all three samples were compared against liberal islet FAIRE sites (in rows). Overlap is reported as the percentage of stringent sites that are captured with liberal sites from another sample. (c) Genome-wide distribution of islet FAIRE sites. Genomic regions enriched for FAIRE are preferentially located upstream and through the body of known genes. More than 25% of FAIRE regions at most stringent enrichment threshold are located proximal to known transcription start sites and in first introns. (d) Intergenic open chromatin sites are enriched for functional annotations. FAIRE sites significantly ($P<0.001$) overlap sequence conservation (28-species most conserved elements), transcription factor binding sites (TFBS) (TFBS cons. and MotifMap), predicted regulatory modules (PreMod) and the ORegAnno database of regulatory elements compared to random regions. Stringent regions are the most enriched across all functional annotations.

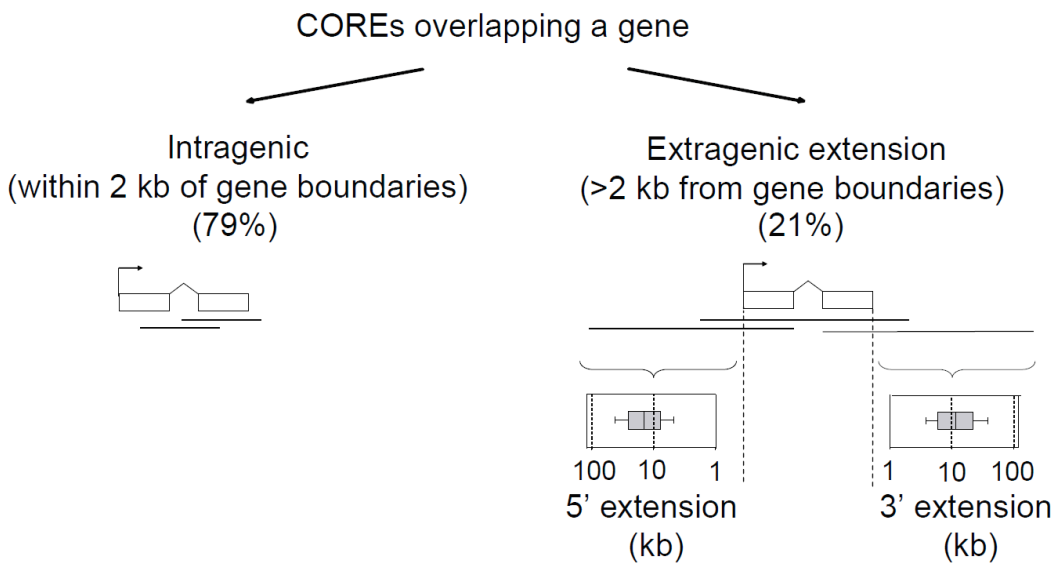
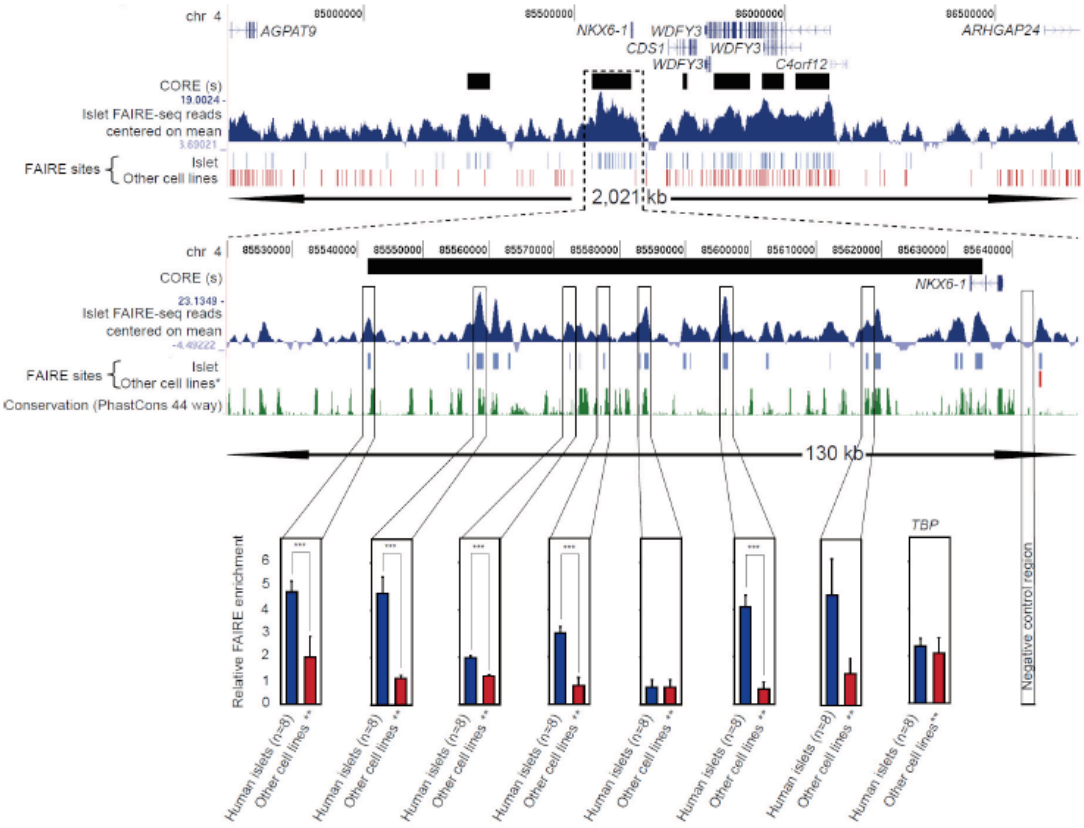
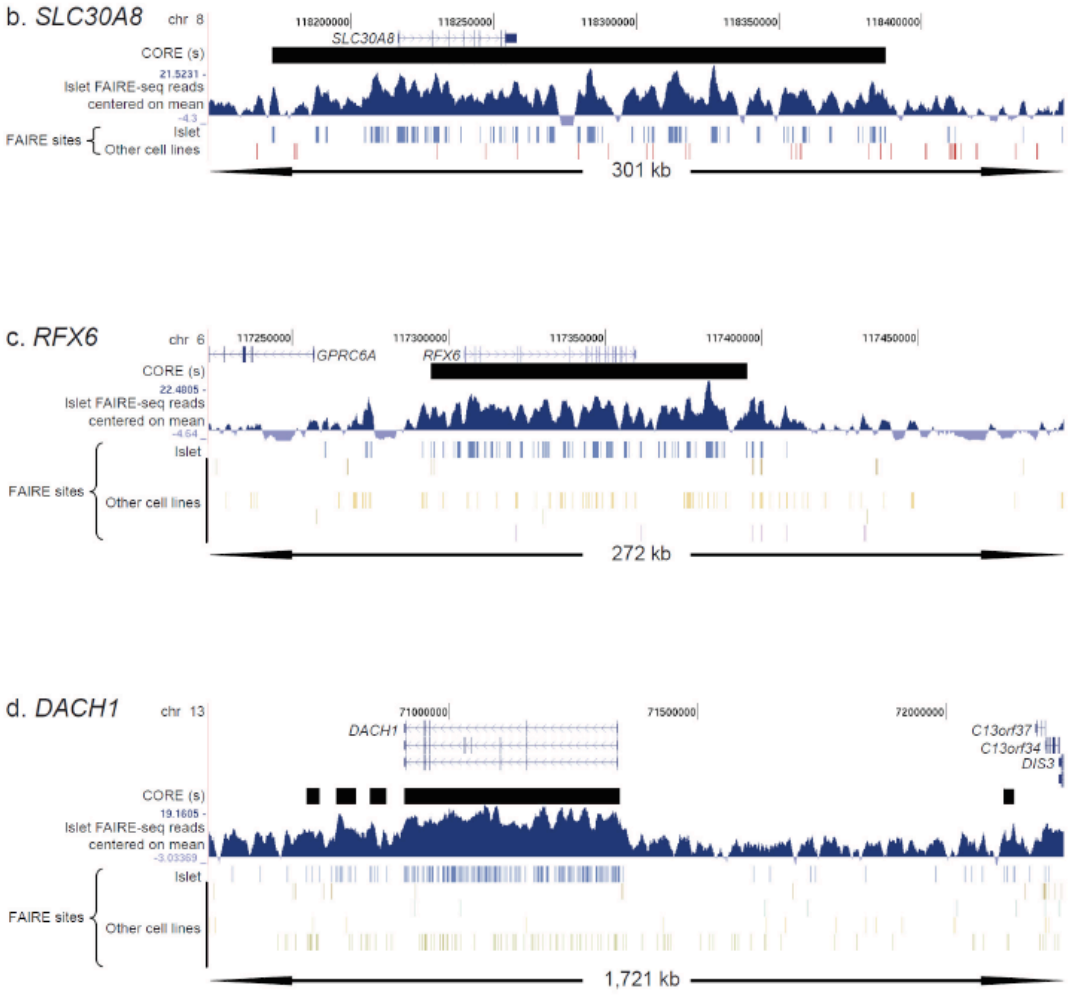
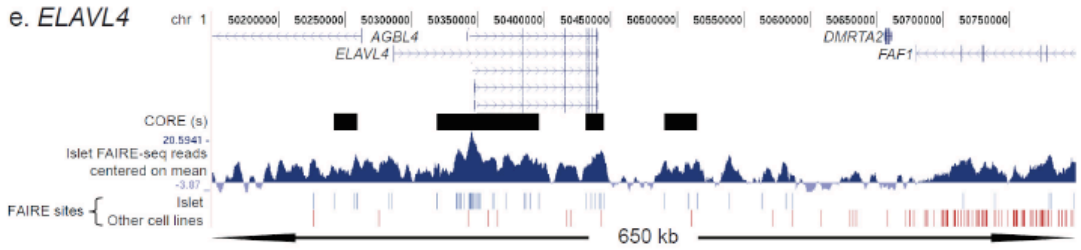


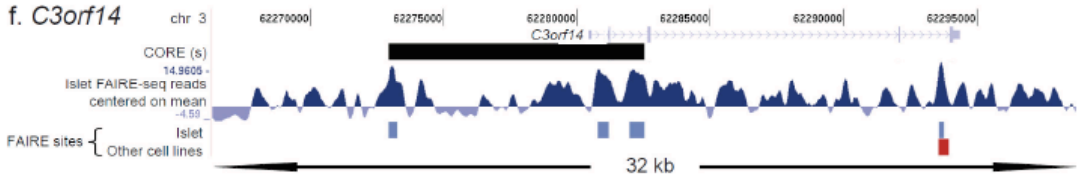
Figure 5.6. Distribution of islet-selective COREs relative to gene boundaries. In most islet-selective COREs that are associated with genes, the entire CORE is located within 2 kb of gene boundaries. However, 21% of islet-selective COREs that are associated with genes extend further than 2 kb from the transcription start site, termination site, or both. Among these COREs that extend far away from gene boundaries, in 54% of the cases the extension is greater in the 5′ rather than in the 3′ direction of the gene, whereas in 46% the extension is greater towards the 3′ end of the gene. The box plots represent the median (line), 25-75th (shaded) and 5-95th percentile (whiskers) of the sizes of COREs extending in either 5′ or 3′ directions from genes.

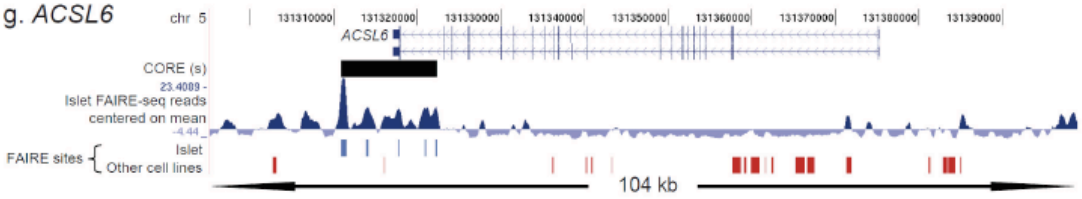
a. *NKX6-1*

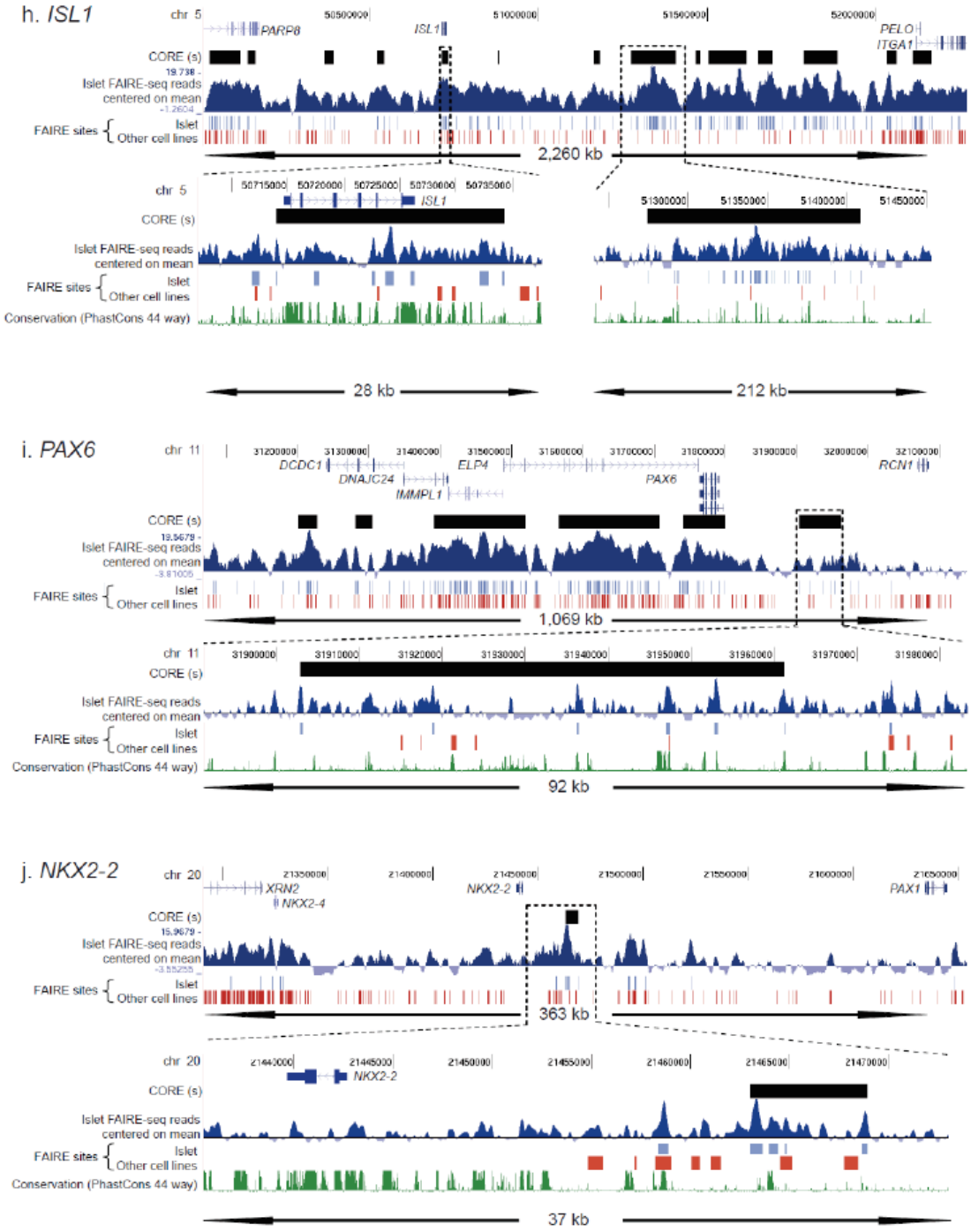

chr 4
AGPAT9
NKX6-1
WDFY3
CDS1
C4orf12
ARHGAP24
CORE (s)
Islet FAIRE-seq reads centered on mean
FAIRE sites
Islet
Other cell lines
2,021 kb
Other cell lines*
Conservation (PhastCons 44 way)
130 kb
TBP
Relative FAIRE enrichment
Human islets (n=8)
Other cell lines
Negative control region

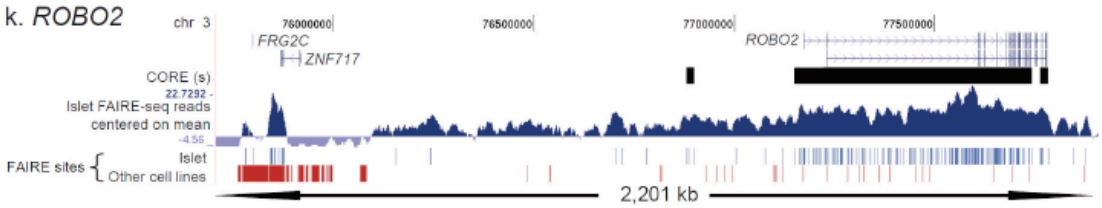

b. SLC30A8
chr 8
118200000
118250000
118300000
118350000
118400000
SLC30A8
CORE (s)
21.5231
Islet FAIRE-seq reads
centered on mean
-4.3
FAIRE sites
Islet
Other cell lines
301 kb
c. RFX6
chr 6
117250000
117300000
117350000
117400000
117450000
GPRC6A
RFX6
CORE (s)
22.4805
Islet FAIRE-seq reads
centered on mean
-4.64
Islet
FAIRE sites
Other cell lines
272 kb
d. DACH1
chr 13
71000000
71500000
72000000
DACH1
C13orf37
C13orf34
DIS3
CORE (s)
19.1605
Islet FAIRE-seq reads
centered on mean
-3.03369
Islet
FAIRE sites
Other cell lines
1,721 kb

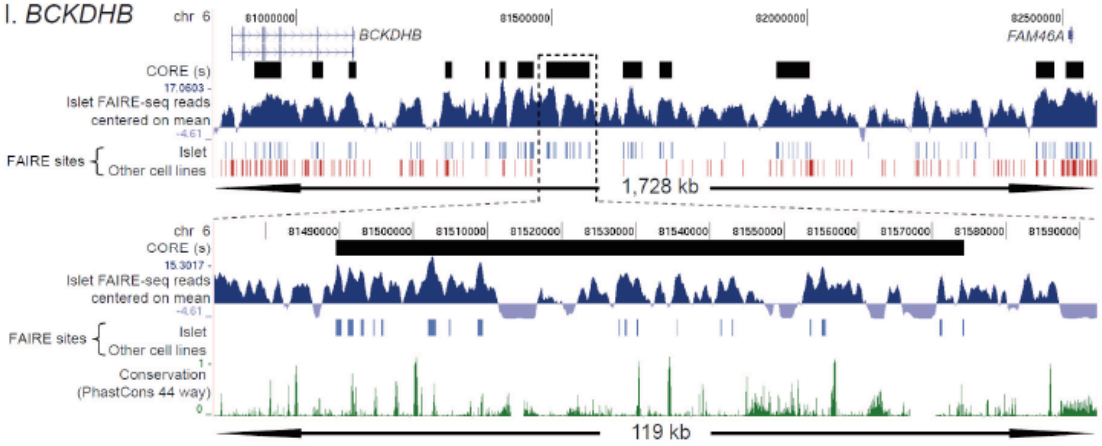

e. ELAVL4
chr 1
50200000
50250000
50300000
50350000
50400000
50450000
50500000
50550000
50600000
50650000
50700000
50750000
AGBL4
ELAVL4
DMRTA2
FAF1
CORE (s)
20.5941
Islet FAIRE-seq reads
centered on mean
-3.07
FAIRE sites
Islet
Other cell lines
650 kb

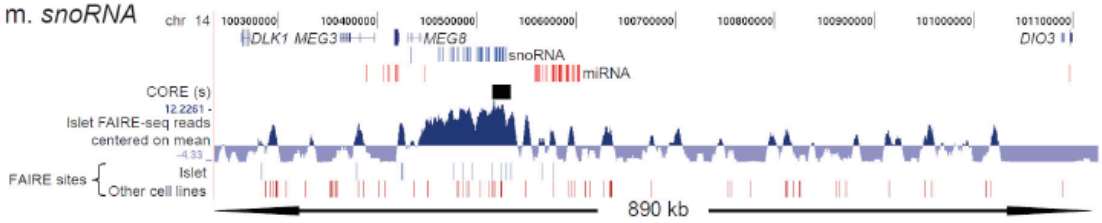

f. C3orf14
chr 3
62270000
62275000
62280000
62285000
62290000
62295000
C3orf14
CORE (s)
14.9605
Islet FAIRE-seq reads
centered on mean
-4.59
FAIRE sites
Islet
Other cell lines
32 kb


g. ACSL6
chr 5
131310000
131320000
131330000
131340000
131350000
131360000
131370000
131380000
131390000
ACSL6
CORE (s)
23.4089
Islet FAIRE-seq reads
centered on mean
-4.44
FAIRE sites
Islet
Other cell lines
104 kb

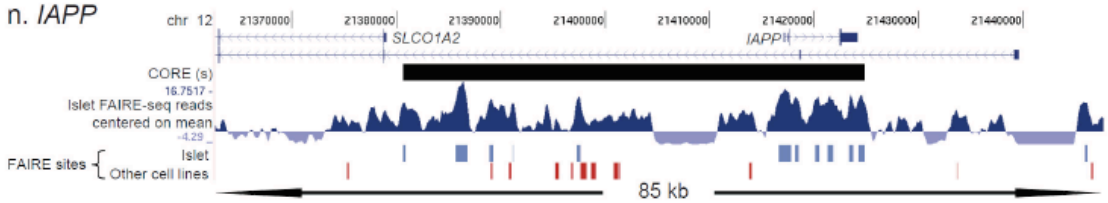

h. ISL1
chr 5
PARP8
ISL1
PELO
ITGA1
CORE (s)
Islet FAIRE-seq reads centered on mean
FAIRE sites
Islet
Other cell lines
2,260 kb
chr 5
ISL1
CORE (s)
Islet FAIRE-seq reads centered on mean
FAIRE sites
Islet
Other cell lines
Conservation (PhastCons 44 way)
28 kb
212 kb
i. PAX6
chr 11
DCDC1
DNAJC24
IMMPL1
ELP4
PAX6
RCN1
CORE (s)
Islet FAIRE-seq reads centered on mean
FAIRE sites
Islet
Other cell lines
1,069 kb
chr 11
CORE (s)
Islet FAIRE-seq reads centered on mean
FAIRE sites
Islet
Other cell lines
Conservation (PhastCons 44 way)
92 kb
j. NKX2-2
chr 20
XRN2
NKX2-4
NKX2-2
PAX1
CORE (s)
Islet FAIRE-seq reads centered on mean
FAIRE sites
Islet
Other cell lines
363 kb
chr 20
NKX2-2
CORE (s)
Islet FAIRE-seq reads centered on mean
FAIRE sites
Islet
Other cell lines
Conservation (PhastCons 44 way)
37 kb


k. ROBO2
chr 3
76000000
76500000
77000000
77500000
FRG2C
ZNF717
ROBO2
CORE (s)
22.7292
Islet FAIRE-seq reads
centered on mean
-4.56
FAIRE sites
Islet
Other cell lines
2,201 kb

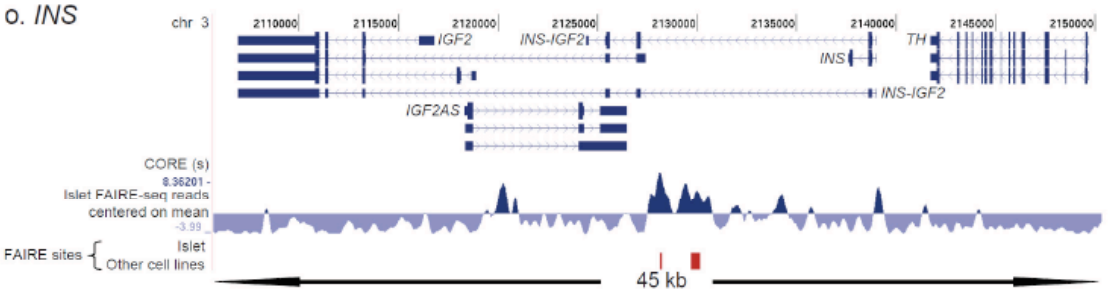

l. BCKDHB
chr 6
81000000
81500000
82000000
82500000
BCKDHB
FAM46A
CORE (s)
17.0603
Islet FAIRE-seq reads
centered on mean
-4.61
FAIRE sites
Islet
Other cell lines
1,728 kb
chr 6
81480000
81500000
81510000
81520000
81530000
81540000
81550000
81560000
81570000
81580000
81590000
CORE (s)
15.3017
Islet FAIRE-seq reads
centered on mean
-4.61
FAIRE sites
Islet
Other cell lines
1
Conservation
(PhastCons 44 way)
0
119 kb


m. snoRNA
chr 14
100300000
100400000
100500000
100600000
100700000
100800000
100900000
101000000
101100000
DLK1 MEG3
MEG8
DIO3
snoRNA
miRNA
CORE (s)
12.2261
Islet FAIRE-seq reads
centered on mean
-4.33
FAIRE sites
Islet
Other cell lines
890 kb

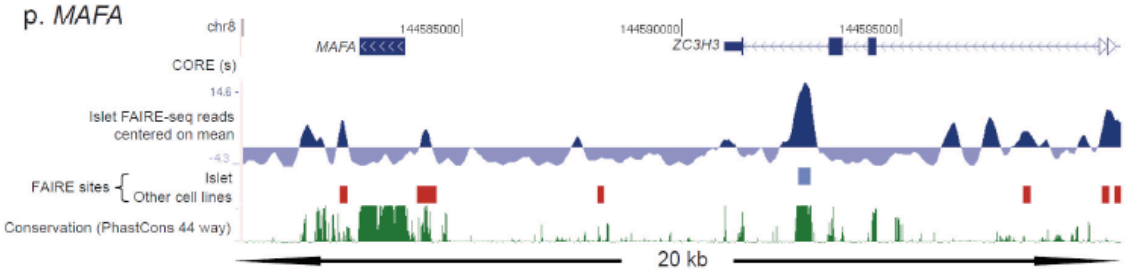
n. *IAPP*



o. *INS*



p. *MAFA*



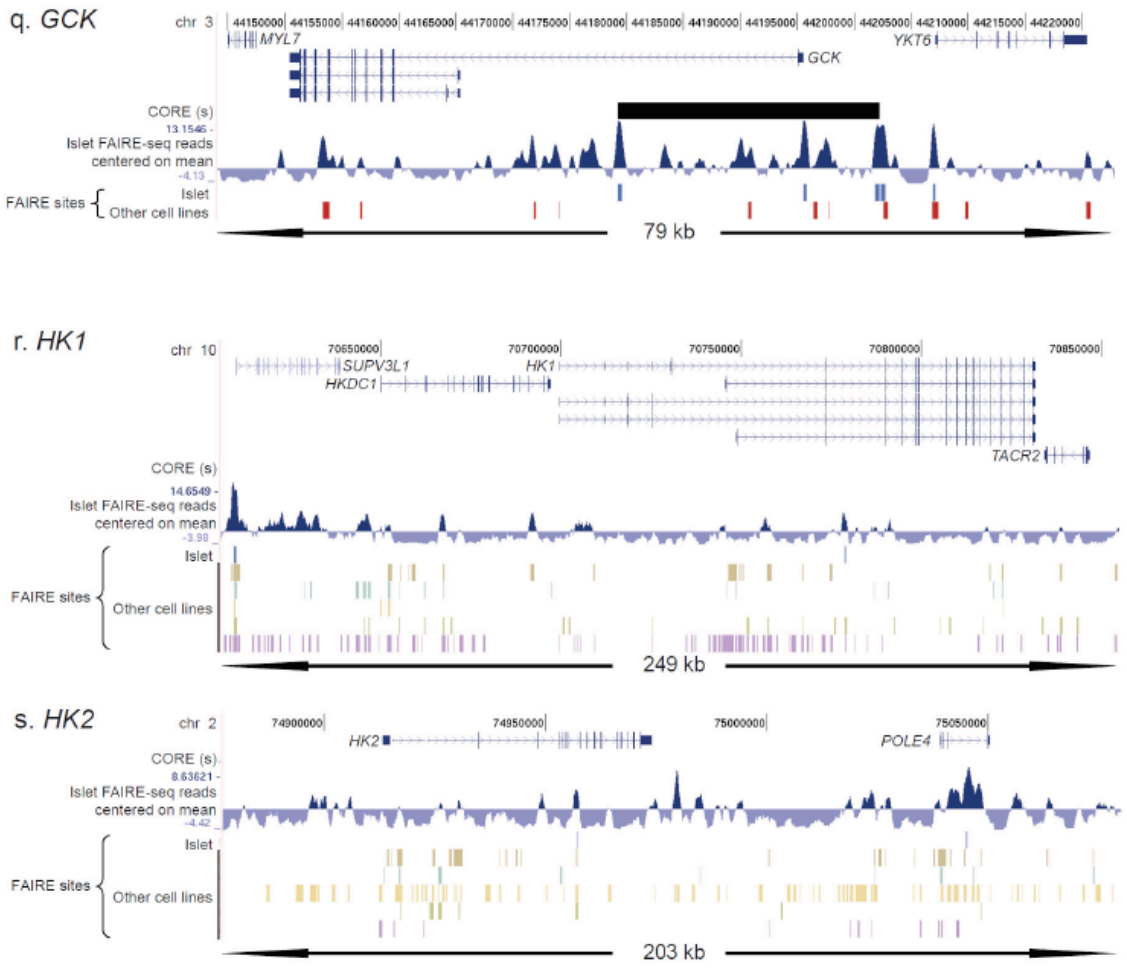


Figure 5.7. Long-range regulatory maps of selected loci. Islet-selective COREs are depicted as black horizontal stripes and labeled as COREs. Blue vertical lines are moderate stringency FAIRE sites in islets, red vertical lines are FAIRE sites present in any of the 5 non-islet cells. In loci where FAIRE sites are observed in only one non islet-cell line we separate non-islet cell line tracks to show the islet-cell selectivity of the CORE. For some loci we show zoomed images with tracks depicting evolutionary conserved sequences (PhastCons 44–way alignments). (a) A CORE spanning 94 kb located in the 3′ region of *NKX6.1*. The lower panel shows qPCR verifications of islet-selective FAIRE enrichment in 6/7 sites from this CORE in 8 additional human islet

samples and 4 additional non-islet cell lines. No differences in enrichment between the two groups of samples were found at the *TBP* promoter region. FAIRE enrichment was normalized to a local negative control region located 5′ of *NKX6.1* that lacks apparent enrichment in islet FAIRE-seq. *P*-values correspond to non-paired *t*-tests, error bars are S.E.M. (b-d) COREs overlapping *SLC30A8*, *RFX6*, and *DACH1*. (e-g) COREs extending 5′ and/or 3′ of *ELAVL4*, *C3orf14* and *ACSL6*. (h,i) COREs overlapping *ISL1* and *PAX6*. Distant 5′ or 3′ COREs are also present and shown as zoomed images. (j) A CORE in a distant region 5′ of *NKX2.2*. (k) A 602 kb CORE overlapping *ROBO2*. (l) A CORE located at a gene desert between *BCKDHB* and *FAM46A*. (m) A CORE overlapping a non-coding RNA cluster on chromosome 14 recently reported to be associated with type 1 diabetes. [304] (n) A CORE at *IAPP* with a 30-kb 5′ extension. (o) Islet FAIRE-seq map of the *INS* locus, showing a small accumulation of islet FAIRE-seq reads surrounding the *INS* promoter. (p) Islet FAIRE-seq map of *MAFA*, showing an islet-selective site at the previously reported distant upstream evolutionary conserved *MAFA* enhancer sequence located in an intronic region of *ZC3H32*. [305] (q,r,s) *GCK*, encoding for the β-cell low-affinity hexokinase, exhibits a CORE at the islet-cell promoter, whereas *HK1* and *HK2*, encoding for high-affinity hexokinases that are inactive in β-cells to enable *GCK* function, are devoid of FAIRE sites in islets but not in other cell types3. [306]

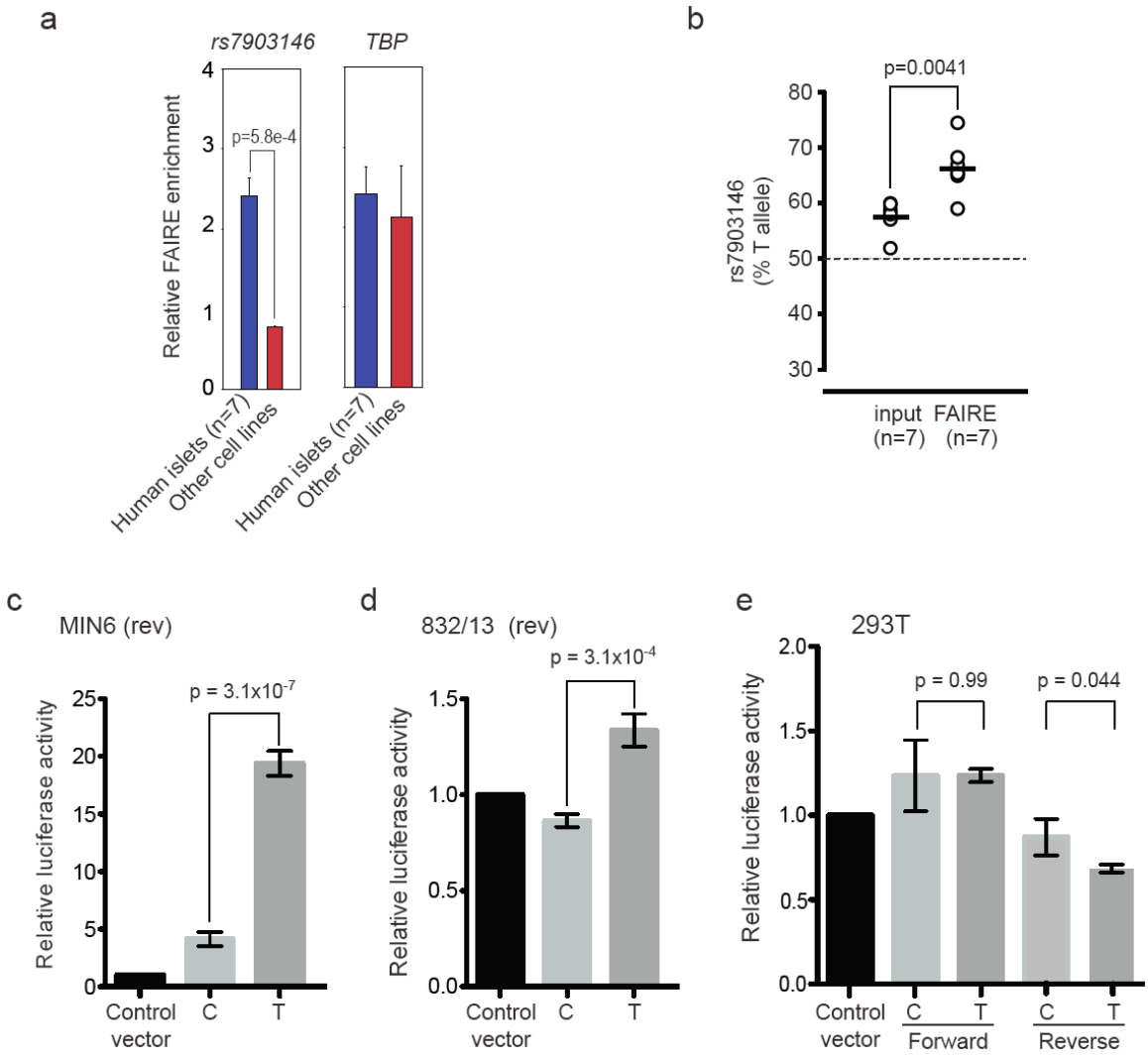


Figure 5.8. Additional functional analysis of the genomic region surrounding rs7903146 (a) Confirmation by qPCR of islet-cell selective FAIRE enrichment in the region surrounding rs7903146. FAIRE enrichment in 7 human islet samples and four non islet cell lines (SW480, C33A, HEK293T and MCF-7) was normalized to the negative control region 5′ of NKX6.1 described in Supplementary Figure 3a. Similar levels of enrichment were found at the TBP promoter region in the two groups of cells. P values correspond to non-paired t tests, error bars are S.E.M. (b) Sanger sequencing of input and FAIRE DNA surrounding rs7903146 from human islets from 7 heterozygous individuals. Area under the sequence curves of each allele was quantified using ImageJ (input: 57.5

+/- 2.7% T allele, FAIRE: 66.2 +/- 4.6% T allele). (c-e) Luciferase reporter data for (c,d) reverse orientation in MIN6 and 832/13 cell lines, and (e) both orientations for 293T cells; standard deviations represent four independent clones for each allele.

A.

B.

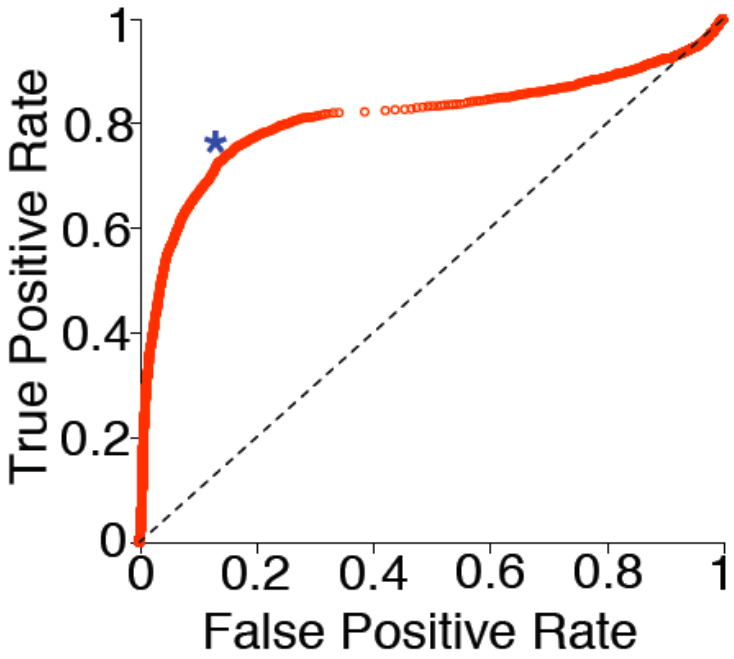


Figure 6.1. Motif training and sequence validation using pancreatic islet open chromatin. (A) Motifs are trained by comparing the percentage of positive and negative training sequences the motif was found in. Input sequences are then assigned a sequence score by adding training scores for all motifs found in the sequence. 361 motifs from

TRANSFAC and JASPAR were trained using islet FAIRE sites (positive) and contiguous regions of no FAIRE signal (negative) from pilot ENCODE regions (B) Receiver Operating Characteristic (ROC) analysis using an independent set of islet FAIRE sites and regions of no FAIRE signal on chromosome 10. The area under the curve (AUC) was .8, with an optimal sensitivity (73%) and specificity (86%) at a sequence score of 3.87 (blue star).

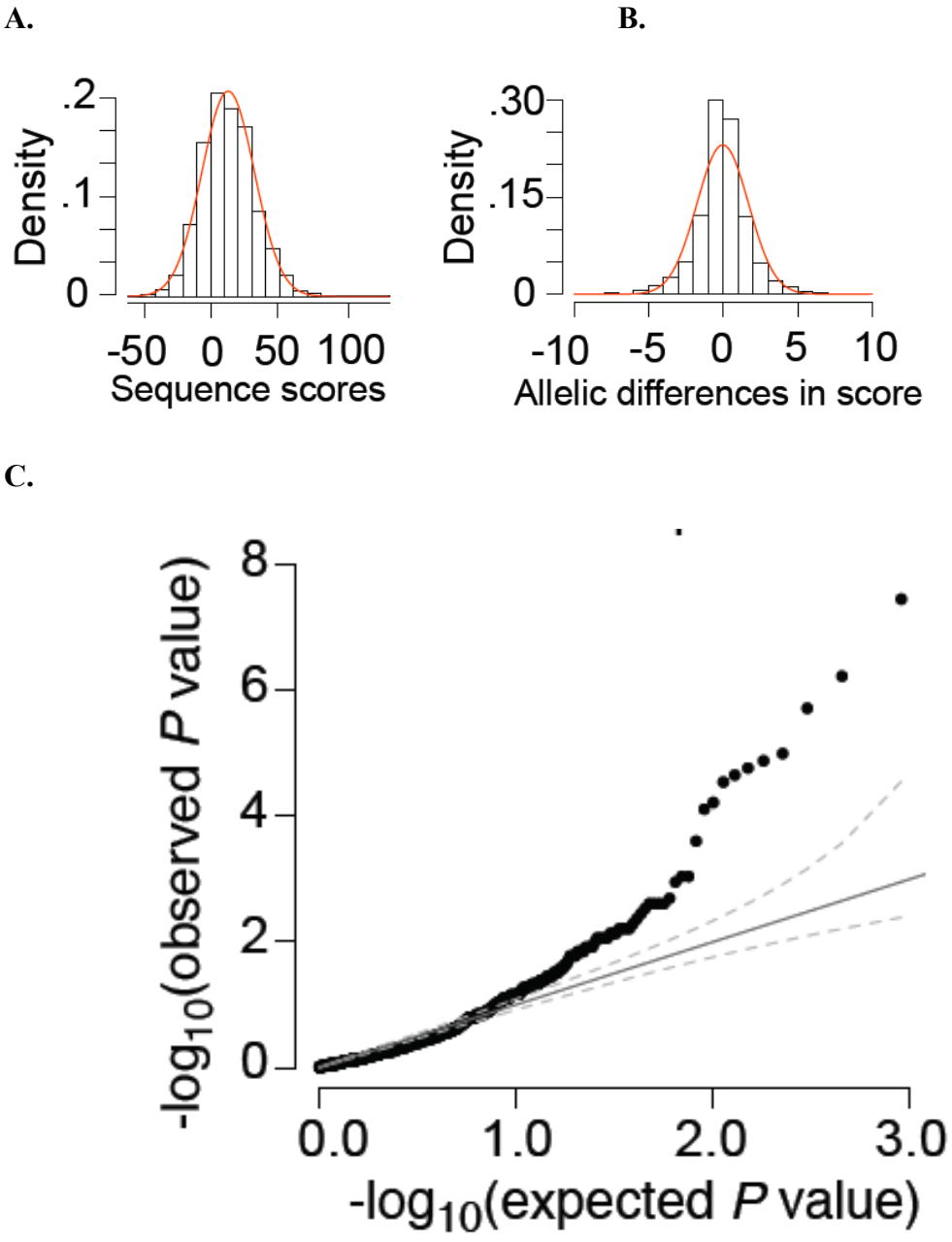


Figure 6.2. Predicting differences in TFBS between SNP alleles. (A) Distribution of sequence scores for 200 bp surrounding each allele of randomly selected HapMap SNPs. (B) Distribution of differences in sequence scores between alleles of randomly selected

HapMap SNPs. (C) Quantile-quantile plot of observed vs. expected p-values for 899 SNPs in high LD (r^2>.8) with a T2D-associated SNP.

Table 2.1. Data sources and ontologies used in CAESAR

Source [a]	Version [b]	URL	Records	Content
Ontology				
MP	1/23/06	www.informatics.jax.org/	3850	Phenotype
eVOC	2.7	www.evocontology.org/	394	Anatomy
GO bp	1/23/06	www.geneontology.org/	9687	Function
GO mf	1/23/06	www.geneontology.org/	7055	Function
Database				
OMIM	1/23/06	www.ncbi.nih.gov/	16564	Disease
Gene	10/1/05	www.ncbi.nih.gov/	32859	Gene
Ensembl	37.35j	www.ensembl.org/	20134	Gene
SwissProt	48.8	www.ebi.ac.uk/uniprot/	13434	Expression
TrEMBL	31.8	www.ebi.ac.uk/uniprot/	57551	Expression
InterPro	12	www.ebi.ac.uk/interpro/	12542	Domain
BIND	10/1/05	www.bind.ca/	35661	Interaction
HPRD	10/1/05	www.hprd.org/	33710	Interaction
KEGG	41	www.genome.jp/kegg/	209	Pathway
MGD	3.41	www.informatics.jax.org/	7705	Phenotype
GAD	1/23/06	hpcio.cit.nih.gov/gad.html	8176	Association
GOA	1/23/06	www.ebi.ac.uk/goa/	27768	Function

a. See text for abbreviations

b. Download date reported where version information not available

Table 2.2. Tests using susceptibility genes for complex human traits

				Reviews				OMIM			
Complex trait	OMIM	Review(s)[a]	Gene[b]	Rank	Total	Percent	Enrich	Rank	Total	Percent	Enrich
Age-related macular degeneration	603075	15094132; 15350892	*CFH*	7263	13771	47.3	2	10450	12608	17.1	1
			LOC387715	–	13771	–	–	–	12608	–	–
ARMD (second run)	603075	N/A[c]	*C2*	–	–	–	–	766	12875	94.1	17
			CFB	–	–	–	–	44	12875	99.7	293
Alzheimer's disease	104300	15225164	*LOC439999*	–	13550	–	–	–	13709	–	–
Asthma	600807	12810182; 14551038	*NPSR1*	1117	13881	92	12	2835	13120	78.4	5
Autism	209850	11733747; 12142938	*EN2*	98	13610	99.3	139	98	13213	99.2	135
Celiac disease	212750	12907013; 12699968; 14592529	*MYO9B*	234	13039	98.2	56	168	12703	98.7	76
Myocardial infarction	608446	15861005; 16041318	*LTA4H*	122	14043	99.1	115	–[d]	–	–	–
Parkinson's disease	168600	16026116; 16278972	*SEMA5A*	4548	13477	66.2	3	879	13329	93.4	15
Rheumatoid arthritis	180300	15478157; 12915205	*PTPN22*	333	13279	97.5	40	2156	13038	83.5	6
			FCRL3	3743	13279	71.8	3	2230	13038	82.9	6
Schizophrenia	181500	15340352; 16033310	*ENTH*	10013	14603	31.4	1	8065	13572	40.6	2
Type 1 diabetes mellitus	222100	12270944; 11921414; 11237226; 11899083	*SUMO4*	1212 3	14272	15.1	1	7675	13130	41.5	2
			PTPN22	165	14272	98.8	86	833	13130	93.7	16
			IL2RA	130	14272	99.1	110	528	13130	96	25
			CTLA4	78	14272	99.5	183	324	13130	97.5	40
Type 2 diabetes mellitus	125853	15662000; 15662001; 15662002; 15662003	*TCF7L2*	2911	13922	79.1	5	4013	13586	70.5	3

Totals	725[e]	13826[e]	94.7[e]	54[f]	879[e]	13130[e]	93.4[e]	43[f]

a. PubMed ID(s) of review articles used in corpus
b. For references see Methods. HUGO approved gene symbols used to identify genes
c. No suitable review corpus available (see Methods)
d. The OMIM record is insufficiently detailed and was not used
e. Median result
f. Mean result

Table 2.3. Independence of CAESAR data sources

	GAD	GObp	GOmf	PPI	IPro	MGD	Path	Tissue
GAD	–	-0.04	-0.04	0.08	0.06	0.1	0.11	-0.03
GObp	2e-6	–	0.43	-0.06	0.12	-0.11	-0.10	-0.06
GOmf	5e-6	2e-16	–	-0.07	0.16	-0.15	-0.08	-0.11
PPI	2e-16	2e-13	2e-16	–	0.08	0.18	0.21	-0.04
IPro	1e-10	2e-16	2e-16	2e-16	–	0.08	0.13	-0.10
MGD	2e-16	2e-16	2e-16	2e-16	2e-16	–	0.27	-0.13
Path	2e-16	2e-16	2e-16	2e-16	2e-16	2e-16	–	-0.18
Tissue	2e-4	2e-10	2e-16	1e-6	2e-16	2e-16	2e-16	–

Top: Spearman rank correlations among pairs of sources. Each value represents the maximum correlation found for a given pair across data for all 11 tested complex traits using default parameters. Bottom: Significance of each correlation. GAD = genetic association database data, GObp = GO biological process data, GOmf = GO molecular function data, PPI = protein–protein interaction data, IPro = InterPro data, MGD = mouse genome database data, path = KEGG pathway data, tissue = Swiss-prot/TrEMBL tissue data.

Table 3.1. Characteristics of the Stage 1 and Stage 2 case and control samples

	Stage 1		Stage 2	
	Cases	Controls	Cases	Controls
N	1161	1174	1215	12 58
Male	653	574	724	768
Female	508	600	491	490
Age of diagnosis (years)	53.0 (12.0)	N/A	56.0 (12.0)	N/A
Study age (years)	63.4 (11.2)	64 (11.7)	60.0 (11.5)	59.0 (10.6)
Body-mass index (kg/m2)	29.8 (6.1)	26.8 (5.0)	30.1 (6.7)	26.4 (4.9)
Fasting glucose (mmol/l)	8.4 (3.9)	5.4 (0.7)	7.2 (2.1)*	5.4 (0.6)†

* n=204 and † n=583 values converted from whole blood to plasma glucose equivalent using prediction equation from the European Diabetes Epidemiology Group, of which †n=262 fasted <8 hours

Table 3.2. Coverage of 10,762 HapMap SNPs (MAF > .05)* within -10 kb/+5 kb of 222 candidate genes

		SNPs -10 kb/+5 kb of gene	
	# SNPs analyzed†	# Captured‡	% Captured‡
SNPs genotyped on GWA panel only	2,150	8,507	79.04
All 3,531 genotyped SNPs	3,531	10,299	95.74
Genotyped and imputed SNPs from GWA panel only	10,596	10,647	98.93
All 3,531 genotyped and 7,498 imputed SNPs	11,029	10,752§	99.91

* MAF=minor allele frequency >.05 in HapMap CEU

† Genotyped SNPs are located within -50 kb/+50 kb of a gene but may not be within -10 kb/+5 kb of a gene; Imputed SNPs are all located within -10 kb/+5 kb of a gene

‡ HapMap SNPs genotyped, imputed, or tagged (r^2>.8) by a genotyped SNP

§ 10,752 includes 3,187 genotyped SNPs, 7,498 imputed SNPs, and 67 SNPs tagged (r^2>.8) by a genotyped SNP

Table 3.3. Gene regions (-10 kb/+5 kb) associated with T2D ($p_{gene} < .05$) in Stage 1 samples

Gene symbol	Chr	Start position* (bp)	End position* (bp)	Coverage (%)†	SNP‡	p_{gene}
SLC2A4	17	7,125,835	7,131,125	90.0	rs222852	.0024
FOXC1	6	1,555,680	1,557,341	100.0	rs2235718	.0028
ARID2	12	44,409,887	44,588,086	97.8	rs11183212	.0029
SOCS3	17	73,864,459	73,867,753	100.0	rs8069976	.0037
FOXC2	16	85,158,443	85,159,948	100.0	rs4843165	.012
ENPP1§,**	6	132,170,853	132,254,043	94.8	rs9402346	.014
PRKAA2	1	56,823,041	56,886,142	90.0	rs11206883	.014
JAK3	19	17,797,961	17,819,800	85.7	rs11888	.016
CBLB	3	106,859,799	107,070,577	98.8	rs17280845	.017
SLC2A2§	3	172,196,839	172,227,470	100.0	rs10513684	.023
PRKAR2B	7	106,279,129	106,396,206	97.0	rs2395836	.027
EDF1	9	137,032,408	137,036,575	100.0	rs3739942	.029
PCK2	14	23,633,323	23,643,177	100.0	rs2759407	.034
PRKAG3	2	219,512,611	219,522,017	100.0	rs6436094	.037
MECR	1	29,340,001	29,378,070	87.1	rs10915239	.038
RXRA	9	134,519,422	134,558,376	85.7	rs3118526	.040
PPARGC1A	4	23,469,914	23,567,969	94.4	rs2970871	.041
*PPARG*d	3	12,304,359	12,450,840	99.1	rs1801282	.042
NR1I3	1	158,012,528	158,021,028	100.0	rs2502807	.049

* Start and end positions of transcribed region (see Methods). Positions based on hg17.

† Percentage of common (MAF>.05) SNPs within -10 kb/+5 kb of a gene and captured at r^2 of at least .8

‡ SNP with minimum p-value in given gene used to calculate p_{gene} value (see Methods)

§ Gene has previous evidence of association in FUSION

** Selected for study only based on previous evidence of association in FUSION

Table 3.4. T2D association for SNPs genotyped in FUSION Stage 1 and 2 samples, sorted by Stage 2 p_{SNP}

						Combined Stage 1+2				
SNP	Gene symbol	Risk/Non risk allele	Risk allele freq.	Stage 1 pSNP	Stage 2 pSNP	Model	p-value	Odds ratio	95% CI	pSNP
rs4740283	*RAPGEF1*	G/A	.104	.0042	.030	REC	.000052	3.12	1.73-5.63	.00013
rs2021966[†]	*ENPP1*	A/G	.608	.00018	.27	REC	.00010	1.27	1.13-1.43	.00026
rs1042522[†,‡]	*TP53*	G/C	.263	.010	.067	MUL	.00037	1.18	1.08-1.30	.00086
rs1882095	*NRF1*	T/C	.381	.0036	.061	DOM	.00043	1.24	1.10-1.40	.00096
rs10513684	*SLC2A2*	C/T	.918	.0046	.20	MUL	.0010	1.28	1.11-1.49	.0023
rs1801282	*PPARG*	C/G	.836	.0025	.44	MUL	.0014	1.20	1.07-1.33	.0034
rs222852	*SLC2A4*	A/G	.610	.00048	.18	MUL	.0029	1.14	1.04-1.23	.0070
rs4843165	*FOXC2*	C/T	.706	.0038	.28[§]	MUL	.0033	1.15	1.05-1.25	.0078
rs5400[‡]	*SLC2A2*	G/A	.871	.0065	.46	MUL	.0045	1.19	1.06-1.35	.010
rs858341	*ENPP1*	G/A	.510	.0039	.70[§]	REC	.0052	1.21	1.06-1.39	.012
rs1349498	*RAPGEF4*	C/T	.729	.0015	.68	DOM	.0065	1.35	1.09-1.67	.015
rs8069976	*SOCS3*	C/A	.849	.0011	.90	MUL	.0070	1.17	1.04-1.31	.016
rs3769249	*RAPGEF4*	G/A	.647	.0040	.79	DOM	.0077	1.27	1.06-1.51	.018
rs17280845	*CBLB*	T/C	.238	.00083	.65	REC	.010	1.37	1.07-1.76	.027
rs5219[‡]	*KCNJ11*	T/C	.476	.0054	.45	MUL	.014	1.11	1.02-1.20	.031
rs10915239	*MECR*	C/A	.945	.0046	.60	REC	.016	1.26	1.04-1.51	.033
rs11206883	*PRKAA2*	A/G	.095	.0014	.58	MUL	.026	1.17	1.02-1.34	.054
rs11183212	*ARID2*	G/A	.200	.00036	.68	MUL	.028	1.12	1.01-1.24	.061
rs2395836	*PRKAR2B*	C/T	.519	.0022	.26	DOM	.034	1.16	1.01-1.34	.072
rs2970871	*PPARGC1A*	C/T	.424	.0012	.081	REC	.042	1.17	1.01-1.36	.088
rs11888	*JAK3*	C/T	.315	.0014	.71	MUL	.075	1.08	0.99-1.18	.15
rs2235718	*FOXC1*	T/C	.117	.00068	.28	REC	.096	1.55	0.92-2.59	.19
rs3118526	*RXRA*	C/T	.922	.0039	.60	DOM	.11	0.52	0.23-1.18	.21
rs9313	*SORBS1*	G/T	.919	.0045	.66	MUL	.11	1.13	0.97-1.32	.21
rs9402346	*ENPP1*	C/G	.646	.00062	-[**]	-	-	-	-	-
rs1830971	*ENPP1*	A/G	.648	.00072	-[**]	-	-	-	-	-
rs1409184	*ENPP1*	G/A	.646	.00072	-[**]	-	-	-	-	-

rs6802898	*PPARG*	C/T	.835	.0031	-**	-	-	-	-	-
rs7796553	*NRF1*	C/T	.172	.0039	-**	-	-	-	-	-
rs943852	*RAPGEF1*	T/C	.111	.0042	-**	-	-	-	-	-

* Positions based on hg17

† SNP was originally imputed, see Table 3.11

‡ Non-synonymous SNP selected for Stage 2 genotyping

§ Included even though Stage 2 sample success rate < 90%

** SNP was not successfully genotyped in Stage 2 or not selected for genotyping in Stage 2 based on high LD with selected SNP

Table 3.5. Quantitative trait association results for SNPs genotyped in FUSION Stage 1 and Stage 2 samples

SNP	Gene	Chr	Position (bp)	Major/ minor allele	Trait	Samples *	Stage 1 p-value †	Stage 2 p-value †	Combined p-value ‡
rs9615264	*PPARA*	22	44,953,108	G/A	HDL level	4682	1.06E-04	.13	.00013
rs10517844	*CPE*	4	166,691,996	T/C	Cholesterol to HDL ratio	4682	4.00E-05	.66	.009
					HDL level	4682	2.07E-05	.098	.065
rs4689388	*WFS1*	4	6,388,128	A/G	LDL level	4067	5.30E-05	.94	.002
rs429358	*APOE*	19	50,103,781	T/C	Cholesterol to HDL ratio	2327	1.78E-10	- §	-
					LDL level	2257	1.09E-06	- §	-
					HDL level	2327	2.36E-06	- §	-
					Cholesterol level	2327	1.51E-05	- §	-
rs4912407	*PRKAA2*	1	56,825,022	G/A	Triglyceride level	2339	3.68E-06	- §	-
					Triglyceride to HDL ratio	2339	2.77E-05	- §	-

* Number of samples corrected to an effective sample size considering the relatedness of some samples

† p-value calculated under additive model

‡ Stage 1 and Stage 2 p-values combined by meta-analysis (see Methods)

§ SNP was not successfully genotyped in Stage 2

Table 3.6. Characteristics of the Stage 1 case and control samples

	FUSION cases	FUSION controls	Finrisk 2002 controls for FUSION	Finrisk 2002 cases	Finrisk 2002 controls
N	789	523 †	276	372	375
Male	429	194	163	224	217
Female	360	329	113	148	158
Age of diagnosis (years)*	51.0 (11.0)	N/A	N/A	59.0 (12.0)	N/A
Study age (years)*	64.2 (10.1)	69.6 (7.7)	62.0 (9.0)	61.0 (12.0)	61.0 (12.0)
Body-mass index (kg/m^2)*	29.3 (6.2)	27.3 (5.5)	26.5 (4.5)	30.7 (6.0)	26.6 (4.4)
Fasting glucose (mmol/l)*	9.6 (4.7)	5.1 (0.6)	5.6 (0.5)	7.3 (1.3)	5.6 (0.5)

* Data are median (interquartile range). N/A=not applicable.

† 523 = 219 FUSION elderly controls and 304 spouse controls.

Table 3.7. Characteristics of the Stage 2 case and control samples *

	D2D cases	D2D controls	Health 2000 cases	Health 2000 controls	Action LADA cases	Action LADA controls	Finrisk 1987 cases	Finrisk 1987 controls	Savitaipale Diabetes Study cases	Savitaipale Diabetes Study controls
N	327	314	127	124	373	402†	266	300	122	118
Male	184	176	67	66	235	259	171	202	67	65
Female	143	138	60	58	138	143	95	98	55	53
Age of diagnosis (years)*	60.0 (13.0)	N/A	55.0 (13.0)	N/A	55.0 (10.0)	N/A	55.5 (13.0)	N/A	55.1 (11.7)	N/A
Study age (years)*	64.0 (11.4)	64.3 (12.0)	61.0 (15.0)	59.0 (12.0)	60.2 (10.8)	58.0 (9.0)	58.0 (11.0)	57.0 (12.0)	57.9 (13.4)	57.0 (13.0)
Body-mass index (kg/m^2)*	29.9 (7.1)	26.4 (4.9)	30.3 (5.4)	26.5 (5.6)	30.3 (6.9)	26.3 (4.7)	30.5 (6.1)	26.7 (4.8)	28.3 (7.1)	25.4 (4.5)
Fasting glucose (mmol/l)*	7.2 (2.0)	5.4 (0.5)	7.3 (2.0)	5.4 (0.5)	7.3 (2.4)	5.5 (0.6)‡	6.9 (3.0)§	5.1 (0.6)§,**	7.2 (0.9)§	5.6 (0.4)§

* Data are median (interquartile range). N/A=not applicable.

† 85 D2D, 100 Health 2000, 52 Finrisk 2002, 97 Finrisk 1987, and 68 Savitaipale Diabetes Study controls

‡ n=165 values converted from whole blood to plasma glucose equivalent using prediction equation from the European Diabetes Epidemiology Group (22), of which n=52 fasted < 8 hours

§ all values converted from whole blood to plasma glucose equivalent using prediction equation from the European Diabetes Epidemiology Group (22)

** n=210 fasted < 8 hours

Table 3.8. SNP coverage and T2D association for 222 candidate gene regions (-10 kb/+5 kb)

Table available online at http://diabetes.diabetesjournals.org/content/suppl/2008/08/19/db07-1731.DC1/Gaulton_online_appendix_tables.xls

Table 3.9. Stage 1 T2D SNP association for 3,531 genotyped SNPs, sorted by pSNP

Table available online at http://diabetes.diabetesjournals.org/content/suppl/2008/08/19/db07-1731.DC1/Gaulton_online_appendix_tables.xls

Table 3.10. Stage 1 T2D association and linkage disequilibrium for genotyped and imputed SNPs within +10kb/-5 kb of gene regions, sorted by chromosome/position

Table available online at http://diabetes.diabetesjournals.org/content/suppl/2008/08/19/db07-1731.DC1/Gaulton_online_appendix_tables.xls

Table 3.11. Imputed SNPs at least 5-fold more strongly associated with T2D than genotyped SNPs in a given gene

	Genotyped SNP			Imputed SNP			Imputed SNP after genotyping	
Gene symbol	SNP	p_{add}*	Position (kb)	SNP	Position (kb)	p_{impute}	p_{add}*	p_{SNP}†
NMU	rs11728776	.18	56,368,695	rs9999653	56,339,177	.023	.016	.035
TP53	rs8079544	.013	7,520,777	rs1042522	7,520,197	.0019	.0044	.010
ENPP1	rs1409184	.0011	132,182,184	rs2021966	132,192,132	.00019	.00026	.00018
RAPGEF1	rs4740304	.0085	131,629,563	rs10901081	131,626,229	.0015	.061	.10
CAPN10	rs7571442	.30	241,275,981	rs3792270	241,251,565	.055	-	-

* Additive model p-value used for genotyped SNPs to allow comparison to imputed p-values

† For imputed SNPs that were then genotyped, we calculated pSNP values to enable comparison to results in Table 3.9

Table 3.12. Genotyped (**bold)** and imputed (non-bold)
SNPs significant at p<.001 in Stage 1 samples before correcting for BMI

SNP	Gene symbol	Chr	Position (bp)	p_{add} or p_{impute} *	p_{add} or p_{impute} (BMI) *
rs7316454	*ARID2*	12	44,486,663	.00014	.0011
rs11183212	***ARID2***	**12**	**44,500,134**	**.00014**	**.0011**
rs7310939	*ARID2*	12	44,488,463	.00014	.0011
rs12580303	*ARID2*	12	44,468,076	.00017	.0014
rs2021966	*ENPP1*	6	132,192,132	.00019	.0012
rs8069976	***SOCS3***	**17**	**73,861,445**	**.00045**	**.00055**
rs8071356	*SOCS3*	17	73,861,591	.00057	.00071
rs11888	***JAK3***	**19**	**17,796,626**	**.00058**	**.00032**
rs11206883	***PRKAA2***	**1**	**56,815,240**	**.00061**	**.0046**
rs2395836	***PRKAR2B***	**7**	**106,381,475**	**.00086**	**.0048**
rs257384	*PRKAR2B*	7	106,399,345	.00088	.0041

* Additive model for genotyped (p_{add}) and imputed (p_{impute}) SNPs

Table 3.13. Genotyped (**bold)** and imputed (non-bold)
SNPs significant at p<.001 in Stage 1 samples after correcting for BMI

SNP	Gene symbol	Chr	Position (bp)	p_{add} or p_{impute} *	p_{add} or p_{impute} (BMI) *
rs2230204	*TRIP10*	19	6,660,848	.030	.00018
rs11888	***JAK3***	**19**	**17,796,626**	**.00058**	**.00032**
rs8069976	***SOCS3***	**17**	**73,861,445**	**.00045**	**.00055**
rs729302	***IRF5***	**7**	**128,162,911**	**.0071**	**.00058**
rs1042522	*TP53*	17	7,520,197	.0019	.00060
rs1801282	***PPARG***	**3**	**12,368,125**	**.0011**	**.00062**
rs8071356	*SOCS3*	17	73,861,591	.00057	.00071
rs11709077	*PPARG*	3	12,311,507	.0015	.00082
rs6802898	***PPARG***	**3**	**12,366,207**	**.0016**	**.00083**
rs2881654	*PPARG*	3	12,371,955	.0018	.00086
rs1899951	*PPARG*	3	12,369,840	.0018	.00090
rs2197423	*PPARG*	3	12,366,583	.0018	.00092
rs7647481	*PPARG*	3	12,366,813	.0018	.00092
rs7649970	*PPARG*	3	12,367,272	.0018	.00092
rs2241392	*C3*	19	6,636,983	.057	.00092
rs17036328	*PPARG*	3	12,365,484	.0019	.00094

* Additive model for genotyped (p_{add}) and imputed (p_{impute}) SNPs

Table 3.14. Stage 1 quantitative trait genotyped and imputed SNP association results ($p < .005$), sorted by p-value

Table available online at http://diabetes.diabetesjournals.org/content/suppl/2008/08/19/db07-1731.DC1/Gaulton_online_appendix_tables.xls

Table 4.1. Characteristics of samples selected for targeted sequencing

	high HDL (>95th PCTL)	low HDL (<5th PCTL)	high TG (>95th PCTL)	low TG (<5th PCTL)
Mean age	62.8	57	58.1	57.8
% Male	20%	80%	63%	47.9%
HDL (mmol/L)	2.30±0.25	0.87±0.13	-	-
TG (mmol/L)	-	-	2.87±0.72	0.61±0.081
BMI	25.2	28.4	29.2	24.7

Table 4.2. Sequencing success

	Targeted bp	Total bp sequenced	% bp sequenced
***GALNT2* locus**			
associated region	14,052	11,330	80.6%
GALNT2 exons	9,965	7,029	70.5%
regulatory regions	6,589	5,631	85.5%
***MMAB/MVK* locus**			
MVK gene region	27,452	22,168	80.8%
MMAB gene region	23,835	18,332	76.9%
regulatory regions	10,044	9,911	98.7%
***TRIB1* locus**			
associated region	19,051	13,108	67.2%
TRIB1 gene region	12,081	9,517	78.8%
regulatory regions	4,183	4,011	95.9%
***ANGPTL3* gene region**	11,995	8,643	72.1%
***MLXIPL* gene region**	35,348	20,786	58.8%
Total	174,595	130,466	74.7%

Table 4.3. Variants identified by sequencing 188 individuals

		known (dbSNP v129)		novel		
region	**total**	total	common (MAF>.05)	total	common (MAF>.05)	rare (MAF<.01)
GALNT2	211	98	78	113	15	70
MMAB/MVK	334	122	102	212	53	120
TRIB1	204	94	79	110	15	72
ANGPTL3	16	5	4	11	1	8
MLXIPL	109	29	25	80	14	43
All	**874**	**348 (40%)**	288 (83%)	**526 (60%)**	98 (19%)	313 (60%)

Table 4.4. HapMap, sequenced and 1000 Genomes Projects variants in LD with HDL-C or TG associated SNPs

			Sequencing (targeted regions only) [a]			1000 Genomes Project (all data)[b]		
	Index SNP [a]	HapMap SNPs in LD (r^2>.8)	r^2>.9	r^2>.5	r^2>.2	r^2>.9	r^2>.5	r^2>.2
MVK/MMAB	rs2338104	34	8 (1)	11 (2)	19 (4)	16	39	250
MLXIPL	rs17145738	10	0 (0)	5 (0)	13 (1)	31	93	124
GALNT2	rs2144300	8	7 (0)	9 (1)	20 (1)	14	25	124
TRIB1	rs2954029	17	3 (0)	3 (0)	14 (3)	7	11	37
ANGPTL3	rs1748195	103	2 (1)	2 (1)	2 (1)	134	151	275

a. # of total variants sequenced in 188 individuals in LD with any previously associated sequenced variant at specified threshold (# of indels)

b. # of total 1000 Genomes (60 CEU samples) variants in LD with any previously associated variant

Table 4.5. Non-HapMap variants in LD (r^2>.2) with trait associated SNPs, sorted by locus and r^2

rs id	locus	alleles	minor allele	MAF	best r^2 with HapMap associated SNP	HapMap associated SNP	Stage 1 p-value	Annotation [a]
rs34483103	*ANGPTL3*	-/TAATGTGGT	-	0.25	1	rs11207997	0.061	*ANGPTL3* 3'UTR
rs10789117	*ANGPTL3*	A/C	C	0.24	1	rs11207997	0.063	*ANGPTL3* 3'UTR
rs2281721	*GALNT2*	C/T	C	0.46	1	rs2281719	0.0030	
rs10864727	*GALNT2*	A/G	A	0.45	1	rs2281719	0.0030	
rs10864728	*GALNT2*	A/G	A	0.46	1	rs2281719	0.0027	
rs11122456	*GALNT2*	A/G	A	0.46	1	rs2281719	0.0027	
rs4846921	*GALNT2*	A/G	G	0.46	0.989	rs2281719	0.0032	
rs2281718	*GALNT2*	A/T	A	0.47	0.966	rs2281719	0.012	
rs4846922	*GALNT2*	C/T	T	0.43	0.915	rs1321257	0.00070	
rs5781574	*GALNT2*	-/CAA	CAA	0.41	0.821	rs1321257	0.00031	
rs4846923	*GALNT2*	G/T	T	0.34	0.616	rs10489615	0.0030	
rs2144301	*GALNT2*	C/T	T	0.27	0.445	rs10779835	0.022	
rs4846917	*GALNT2*	C/T	T	0.24	0.407	rs2281719	0.39	
rs4846841	*GALNT2*	A/G	A	0.24	0.393	rs2281719	0.39	
rs4846840	*GALNT2*	G/T	T	0.24	0.392	rs2281719	0.39	
rs6672758	*GALNT2*	C/T	C	0.24	0.387	rs2281719	0.39	
rs966333	*GALNT2*	C/T	T	0.19	0.322	rs2281719	0.18	
rs6666884	*GALNT2*	G/T	G	0.21	0.322	rs1321257	0.0038	H3K4me3
rs2103827	*GALNT2*	A/T	T	0.19	0.283	rs10489615	0.17	
rs598203	*GALNT2*	C/G	G	0.19	0.281	rs2281719	0.068	DnaseHS
rs59153235	*GALNT2*	G/T	T	0.28	0.234	rs2281719	0.05	
rs35198744	*GALNT2*	C/T	T	0.2	0.215	rs10489615	0.064	

rs13240065	*MLXIPL*	A/G	A	0.1	0.87	rs17145750	0.00018	
rs3812316	*MLXIPL*	C/G	G	0.1	0.87	rs17145750	0.00018	*MLXIPL* Q241H
rs13246993	*MLXIPL*	A/G	A	0.1	0.87	rs17145750	0.00017	
rs34062580	*MLXIPL*	A/G	A	0.1	0.868	rs17145750	0.00023	
rs13235543	*MLXIPL*	C/T	T	0.1	0.867	rs17145750	0.00018	RNA Pol II
rs34060476	*MLXIPL*	A/G	G	0.1	0.844	rs17145750	0.00091	H3K4me3,RNA Pol II,H3ac,HNF4a,DnaseHS
rs13247874	*MLXIPL*	C/T	T	0.15	0.737	rs17145750	0.0029	
rs13225660	*MLXIPL*	C/T	T	0.16	0.732	rs17145750	0.0023	
rs55747707	*MLXIPL*	A/G	A	0.15	0.722	rs17145750	0.0037	H3K4me3,RNA Pol II,SREBP1A,SREBP2,HNF4A,cMYC,FAIRE,DnaseHS
rs35493868	*MLXIPL*	C/G	G	0.15	0.722	rs17145750	0.0090	H3K4me3
rs35368205	*MLXIPL*	C/T	T	0.15	0.711	rs17145750	0.0081	DnaseHS
rs61010704	*MLXIPL*	A/G	G	0.22	0.449	rs17145750	0.0056	
rs35512732	*MLXIPL*	-/C	-	0.22	0.449	rs17145750	0.0056	
rs10850358	*MMAB/MVK*	A/G	G	0.49	1	rs11067231	0.0013	RNA pol II,DNaseHS
rs2058805	*MMAB/MVK*	C/T	C	0.48	1	rs7134594	0.0015	
rs2058806	*MMAB/MVK*	A/C	A	0.49	1	rs11067231	0.0013	
rs6606734	*MMAB/MVK*	G/T	G	0.48	1	rs7134594	0.0012	
rs736344	*MMAB/MVK*	A/G	G	0.48	1	rs7134594	0.0015	
rs7953014	*MMAB/MVK*	A/G	G	0.48	1	rs7134594	0.0015	
rs3782894	*MMAB/MVK*	G/T	T	0.48	0.989	rs7134594	0.00096	
rs60036171	*MMAB/MVK*	-/T	-	0.49	0.989	rs7134594	0.00092	

rs59227481	*MMAB/MVK*	A/G	G	0.37	0.632	rs7134594	0.0071	
rs57044180	*MMAB/MVK*	-/CACT	-	0.32	0.522	rs10850435	0.017	DNaseHS
rs12322541	*MMAB/MVK*	A/G	G	0.3	0.503	rs7134594	0.013	DNaseHS
rs3782897	*MMAB/MVK*	C/T	C	0.32	0.496	rs7134594	0.019	
rs10774774	*MMAB/MVK*	G/T	T	0.31	0.482	rs7134594	0.015	H3K4me3,RNA pol II,SREBP1A,SREBP2, cMYC,DNaseHS
rs10774775	*MMAB/MVK*	C/T	T	0.31	0.482	rs7134594	0.013	*MMAB* R18H, H3K4me3,RNA pol II,SREBP1A,SREBP2, cMYC,DNaseHS
rs11364376	*MMAB/MVK*	-/G	-	0.3	0.48	rs7134594	0.013	
rs59652081	*MMAB/MVK*	-/CTT	-	0.31	0.47	rs7134594	0.014	
rs877709	*MMAB/MVK*	G/T	T	0.29	0.459	rs10850435	0.019	3' UTR
rs11067359	*MMAB/MVK*	A/G	A	0.21	0.316	rs10850435	0.12	
rs11067271	*MMAB/MVK*	C/T	C	0.2	0.273	rs7134594	0.048	
rs2980886	*TRIB1*	A/G	G	0.49	1	rs2980853	0.00040	
rs2001846	*TRIB1*	C/T	T	0.42	0.978	rs2954021	0.015	
rs2954017	*TRIB1*	C/T	T	0.42	0.967	rs2954021	0.015	
rs2954023	*TRIB1*	-/A	T	0.25	0.422	rs2954021	0.11	
rs12674939	*TRIB1*	A/T	A	0.21	0.341	rs2954021	0.060	
rs7015677	*TRIB1*	A/G	G	0.18	0.297	rs2954021	0.091	
rs72655675	*TRIB1*	-/T	T	0.18	0.284	rs2954021	0.090	
rs12679184	*TRIB1*	C/T	T	0.18	0.28	rs2954021	0.072	
rs7828194	*TRIB1*	A/G	A	0.14	0.274	rs2954021	0.088	

rs34604874	*TRIB1*	A/G	A	0.18	0.274	rs2954021	0.071	
rs62521034	*TRIB1*	C/T	T	0.29	0.267	rs2980856	0.77	3' UTR
rs7828701	*TRIB1*	C/G	G	0.15	0.249	rs2954021	0.072	
rs4419828	*TRIB1*	G/T	T	0.27	0.246	rs6982636	0.013	
rs66488903	*TRIB1*	-/TTGTT	-	0.19	0.203	rs2954021	0.13	RNA pol II

Table 4.6. Datasets used to annotate non-coding variants

HepG2 experimental data

Dataset	Type	Source
DNaseI hypersensitivity	Nucleosome occupancy	ENCODE; Crawford
FAIRE	Nucleosome occupancy	ENCODE; Lieb
FOXA2 binding	DNA-binding protein	Wallerman O, Nucleic Acids Res. 2009 Dec;37(22):7498-508.
HNF4A binding	DNA-binding protein	Wallerman O, Nucleic Acids Res. 2009 Dec;37(22):7498-508.
GABP binding	DNA-binding protein	Wallerman O, Nucleic Acids Res. 2009 Dec;37(22):7498-508.
SREBP1A binding	DNA-binding protein	ENCODE; Snyder
SREBP2	DNA-binding protein	ENCODE; Snyder
RNA Pol II binding	DNA-binding protein	ENCODE; Snyder
USF1 binding	DNA-binding protein	Rada-iglesias, Genome Res. 2008 Mar;18(3):380-92.
USF2 binding	DNA-binding protein	Rada-iglesias, Genome Res. 2008 Mar;18(3):380-92.
CTCF binding	DNA-binding protein	ENCODE; Bernstein/Iyer
cMYC binding	DNA-binding protein	ENCODE; Bernstein/Iyer
H3 acetylation	Histone modification	Rada-iglesias, Genome Res. 2008 Mar;18(3):380-92.
H3K4me3	Histone modification	Guenther MG, Cell. 2007 Jul 13;130(1):77-88.

Non-coding element predictions

Dataset	Reference
28-species most conserved elements	Miller W. 2007 Dec;17(12):1797-808.
Predicted transcription factor binding sites	Xie X. 2009 Jan 15;25(2):167-74.
Predicted regulatory modules	Blanchette M. Genome Res. 2006 May;16(5):656-68
Predicted tissue-specific enhancers	Pennacchio LA. Genome Res. 2007 Feb;17(2):201-11.

Table 4.7. Stage 1 quantitative trait association with variants in low LD (r^2<.2) with previously associated HapMap variants

SNP id	locus	alleles	MAF	annotation	r^2	pvalue
rs72649520	*MMAB/MVK*	A/G	0.005		0.98	0.0018
rs72650176	*MMAB/MVK*	C/T	0.005	*MMAB* 3' UTR	0.97	0.0018
rs72650173	*MMAB/MVK*	AT/C	0.005	RNA PolII	0.93	0.0018
rs72648004	*MMAB/MVK*	C/T	0.005	Dnase HS	0.61	0.0021
rs56200521	*GALNT2*	C/G	0.12		0.92	0.005
rs56217501	*GALNT2*	C/G	0.12		0.92	0.005
rs60122995	*GALNT2*	A/G	0.12		0.94	0.014
rs1051943	*MLXIPL*	C/T	0.02	*MLXIPL* 3' UTR RNA PolII	0.72	0.016
rs72647336	*TRIB1*	A/G	0.02	CTCF_IMR90	0.59	0.018
rs67866345	*MMAB/MVK*	C/T	0.01		0.94	0.02
rs66815418	*GALNT2*	C/T	0.13		0.95	0.022
rs72655702	*TRIB1*	A/T	0.008		0.94	0.031
rs72649530	*GALNT2*	-/G	0.21		0.96	0.031
rs12038714	*GALNT2*	A/G	0.20		0.96	0.031
rs72649011	*MLXIPL*	A/G	0.008	RNA PolII	0.65	0.035
rs72649026	*MLXIPL*	C/T	0.008	CTCF_IMR90	0.65	0.035
rs72646994	*MMAB/MVK*	G/T	0.05		0.67	0.037
rs72647325	*TRIB1*	C/T	0.008		0.54	0.037
rs55882275	*TRIB1*	A/T	0.089		0.99	0.04
rs72648003	*MMAB/MVK*	A/G	0.02		0.72	0.041
rs72651720	*MMAB/MVK*	C/T	0.008		0.7	0.045

Table 4.8. Re-sequenced SNPs Stage 1+2 quantitative trait association, sorted by combined p-value

				Stage 1			Stage 2			Combined		
SNP id	gene region	major/minor allele	trait	MAF	Effect size	*P*-value	MAF	Effect size	*P*-value	MAF	Effect size	*P*-value
rs72647336	*TRIB1*	G/A	Triglyceride level	.02	.31	.023	.035	.11	.00083	.033	.12	.00011
rs72650176	*MMAB*	T/C	HDL-C level	.0006	-2.12	.0018	.001	-.14	.14	.0009	-.42	.15
rs56217501	*GALNT2*	C/G	HDL-C level	.12	-.15	.0051	.12	.061	.20	.12	.02	.27
rs28934897	*MVK*	A/G	HDL-C level	.0013	.43	.33	.0016	.08	.15	.0016	.12	.42
rs72649012	*MLXIPL*	T/C	Triglyceride level	.012	.066	.66	.015	.044	.39	.015	.038	.44
rs72650181	*MMAB*	T/-	HDL-C level	.0004	-1.57	.026	.0001	2.3	.022	.0002	-.29	.62

Table 4.9. Aming acid changing variants (MAF<.05) identified by sequencing in 188 samples

SNP	Gene	Amino acid change	# high trait value	#low trait value	Predicted effect*
rs72650181	*MMAB*	Y238*	0	1	-
rs72649574	*ANGPTL3*	N151D	0	1	TOLERATED
rs66489924	*MLXIPL*	R841W	1	0	DAMAGING
rs72649573	*ANGPTL3*	L127F	2	1	DAMAGING
rs28934897	*MVK*	V377I	0	1	TOLERATED
rs72649012	*MLXIPL*	V758I	3	1	DAMAGING

* Predictions using SIFT (Kumar *et al*. Nat Protoc. 2009;4(7):1073-81)

Table 4.10. Excess of high or low trait value individuals with variants in genomic windows of 200 bp, 400 bp, 1 kb and 2 kb

Window size [a]	Coordinates	# high trait value [b]	# low trait value [b]	Uncorrected p-value [c]
2 kb	chr1:228482029-228484029	1	11	.0055
	chr1:228367973-228369973	1	9	.016
1 kb	chr1:228483591-22848591	0	6	.030
	chr1:228483119-228484119	1	7	.063
400 bp	chr12:108497292-108497692	0	5	.057
	chr12:108496572-108496972	5	0	.055
200 bp	chr12:108497292-108497492	0	5	.05
	chr12:108496572-108496772	5	0	.056

a. Two most significant results shown for each window size
b. Number of high or low individuals with a rare allele in window
c. P-values calculated using 10,000 permutations of high / low trait status. Experiment-wide significance = 4.8x10-5

Table 5.1. FAIRE-seq sequence depth and enrichment sites in three human islet samples

		FAIRE-seq sites		
	Sequence reads	Liberal	Moderate	Stringent
sample 1	39,359,429	205,922	99,361	18,189
sample 2	25,176,624	213,972	91,455	9,601
sample 3	60,515,180	202,783	81,546	33,305

FAIRE was performed on three human pancreatic islet sample (Methods). Sample 3 had the highest purity (Table 5.2) and was thus sequenced at greater depth and used for subsequent analysis. Aligned reads were used to call FAIRE sites at three thresholds

Table 5.2. Donor profiles of islet samples

	sample 1	sample 2	sample 3
Ethnicity	Caucasian	Caucasian	N/A
Sex	Male	Male	Male
Age (years)	26	53	53
BMI (kg/m2)	27.3	32.65	23.5
Cause of death	Head trauma	Cerebral hemmorhage	Cerebral hemmorhage
Cold ischemia (hours)	9	18	6.5
Islet purity (%)	55	55	85
Culture duration before shipment (days)	4	0	3
3-days reculture after arrival	No	No	Yes

Table 5.3. RefSeq transcripts with preferential islet FAIRE enrichment

			% FAIRE enriched [a]		
Transcript id	Gene symbol	position [b]	islets	5 cell lines [c]	fold
NM_000415	IAPP	chr12:21415084-21425683	33.63	0.00	-
NM_006168	NKX6-1	chr4:85631459-85640411	19.71	0.00	-
NM_004472	FOXD1	chr5:72775840-72782108	16.00	0.00	-
NM_014469	RBMXL2	chr11:7064740-7070955	9.04	0.00	-
NM_019062	RNF186	chr1:20011108-20016358	8.67	0.00	-
NM_004138	KRT33A	chr17:36753896-36762582	8.61	0.00	-
NM_004982	KCNJ8	chr12:21807155-21821014	8.25	0.00	-
NM_145175	FAM84A	chr2:14688306-14695898	7.84	0.00	-
NM_012183	FOXD3	chr1:63559317-63565385	7.50	0.00	-
NM_001001343	C5orf40	chr5:156699184-156707307	6.59	0.00	-
NM_020633	VN1R1	chr19:62656353-62661666	6.06	0.00	-
NM_006057	B3GALT5	chr21:39949123-39958685	5.95	0.00	-
NM_033170	B3GALT5	chr21:39949123-39958685	5.95	0.00	-
NM_033171	B3GALT5	chr21:39949123-39958685	5.95	0.00	-
NM_033172	B3GALT5	chr21:39949123-39958685	5.95	0.00	-
NM_033173	B3GALT5	chr21:39949123-39958685	5.95	0.00	-
NM_178445	CCRL1	chr3:133799670-133806072	5.84	0.00	-
NM_022843	PCDH20	chr13:60879819-60889656	5.41	0.00	-
NM_033512	TSPYL5	chr8:98352889-98361352	5.13	0.00	-
NM_012403	ANP32C	chr4:165335608-165340313	4.68	0.00	-
NM_018971	GPR27	chr3:71883890-71889018	4.50	0.00	-
NM_001077710	FAM110C	chr2:29607-38385	4.48	0.00	-
NM_032099	PCDHGB5	chr5:140755878-140762335	3.41	0.00	-
NM_032089	PCDHGA9	chr5:140760703-140767190	3.39	0.00	-

NM_032115	KCNK16	chr6:39388459-39400294	3.39	0.00	-
NM_001004310	FCRL6	chr1:158036796-158054671	3.33	0.00	-
NM_182532	TMEM61	chr1:55217052-55232554	3.30	0.00	-
NM_012226	P2RX2	chr12:131703475-131711045	3.20	0.00	-
NM_016318	P2RX2	chr12:131703475-131711045	3.20	0.00	-
NM_174872	P2RX2	chr12:131703475-131711045	3.20	0.00	-
NM_170683	P2RX2	chr12:131703475-131711045	3.20	0.00	-
NM_174873	P2RX2	chr12:131703475-131711045	3.20	0.00	-
NM_170682	P2RX2	chr12:131703475-131711045	3.20	0.00	-
NM_001105195	FAM123C	chr2:131228333-131244177	3.17	0.00	-
NM_001105194	FAM123C	chr2:131228333-131244177	3.17	0.00	-
NM_031883	PCDHAC2	chr5:140324535-140331190	3.07	0.00	-
NM_001105193	FAM123C	chr2:131227693-131244177	3.05	0.00	-
NM_152698	FAM123C	chr2:131227546-131244177	3.02	0.00	-
NM_001004734	OR14I1	chr1:246909292-246914228	2.80	0.00	-
NM_006308	HSPB3	chr5:53785201-53789964	2.27	0.00	-
NM_017594	DIRAS2	chr9:92409933-92446928	2.21	0.00	-
NM_001005611	EDA	chrX:68750635-68755868	2.18	0.00	-
NM_001024215	FBLIM1	chr1:15961580-15976302	2.07	0.00	-
NM_144605	SEPT12	chr16:4765674-4780348	2.02	0.00	-
NM_130786	A1BG	chr19:63547983-63558677	1.88	0.00	-
NM_006439	MAB21L2	chr4:151720526-151727293	1.82	0.00	-
NM_018228	C14orf115	chr14:73882918-73898464	1.74	0.00	-
NM_020160	MEIS3	chr19:52596194-52616597	1.40	0.00	-
NM_001009813	MEIS3	chr19:52596192-52616597	1.40	0.00	-
NM_030883	OR2H1	chr6:29532208-29542078	1.37	0.00	-
NM_207379	TMEM179	chr14:104129464-	1.24	0.00	-

		104144142			
NM_000209	PDX1	chr13:27390167-27400451	1.23	0.00	-
NM_181709	FAM101A	chr12:123337662-123368521	1.22	0.00	-
NM_000814	GABRB3	chr15:24337786-24571344	1.15	0.00	-
NM_021912	GABRB3	chr15:24337786-24572020	1.15	0.00	-
NM_031882	PCDHAC1	chr5:140284485-140291569	0.99	0.00	-
NM_001039360	ZBTB7C	chr18:43805742-43823492	0.97	0.00	-
NM_023926	ZSCAN18	chr19:63285017-63303389	0.85	0.00	-
NM_001128618	C9orf57	chr9:73854116-73867341	0.83	0.00	-
NM_016615	SLC6A13	chr12:198051-244263	0.82	0.00	-
NM_001127648	GABRA1	chr5:161208280-161261543	0.73	0.00	-
NM_001127647	GABRA1	chr5:161207969-161261543	0.73	0.00	-
NM_001127646	GABRA1	chr5:161206475-161261543	0.71	0.00	-
NM_001127645	GABRA1	chr5:161206119-161261543	0.71	0.00	-
NM_001127644	GABRA1	chr5:161205517-161261543	0.70	0.00	-
NM_001127643	GABRA1	chr5:161205262-161261543	0.69	0.00	-
NM_000806	GABRA1	chr5:161204774-161261543	0.69	0.00	-
NM_021098	CACNA1H	chr16:1141241-1213773	0.52	0.00	-
NM_001005407	CACNA1H	chr16:1141241-1213773	0.52	0.00	-
NM_002259	KLRC1	chr12:10487903-10499196	0.51	0.00	-
NM_007328	KLRC1	chr12:10487903-10499196	0.51	0.00	-
NM_213657	KLRC1	chr12:10487903-10500251	0.47	0.00	-
NM_213658	KLRC1	chr12:10487903-10500251	0.47	0.00	-
NM_014351	SULT4A1	chr22:42549719-42591711	0.43	0.00	-
NM_182594	ZNF454	chr5:178298829-178328040	0.42	0.00	-
NM_014392	D4S234E	chr4:4436883-4473685	0.42	0.00	-
NM_001040101	D4S234E	chr4:4436883-4473686	0.42	0.00	-
NM_198904	GABRG2	chr5:161425225-161517123	0.28	0.00	-

NM_198903	GABRG2	chr5:161425225-161517123	0.28	0.00	-
NM_000816	GABRG2	chr5:161425225-161517123	0.28	0.00	-
NM_015347	RIMBP2	chr12:129444633-129570363	0.27	0.00	-
NM_021599	ADAMTS2	chr5:178508735-178706935	0.25	0.00	-
NM_014244	ADAMTS2	chr5:178471455-178706935	0.24	0.00	-
NM_024690	MUC16	chr19:8818519-8955018	0.12	0.00	-
NM_000555	DCX	chrX:110421662-110543030	0.09	0.00	-
NM_178151	DCX	chrX:110421662-110543962	0.09	0.00	-
NM_178152	DCX	chrX:110421662-110544062	0.09	0.00	-
NM_178153	DCX	chrX:110421662-110544062	0.09	0.00	-
NM_145793	GFRA1	chr10:117810942-118023784	0.06	0.00	-
NM_005264	GFRA1	chr10:117810942-118024966	0.06	0.00	-
NM_001005615	EDA	chrX:68750635-68999614	0.05	0.00	-
NM_002545	OPCML	chr11:131788084-132320247	1.09	0.01	181.6
NM_001080455	SLC35F4	chr14:57098392-57135368	3.23	0.06	49.7
NM_001130682	GUCY1A3	chr4:156805311-156874951	15.93	0.33	47.8
NM_000856	GUCY1A3	chr4:156805311-156874951	15.93	0.33	47.8
NM_001130683	GUCY1A3	chr4:156805598-156874951	15.99	0.33	47.8
NM_001130685	GUCY1A3	chr4:156805598-156874951	15.99	0.33	47.8
NM_001130684	GUCY1A3	chr4:156806264-156874951	16.15	0.34	47.8
NM_001130686	GUCY1A3	chr4:156805311-156853983	20.48	0.48	43.0
NM_001130687	GUCY1A3	chr4:156805311-156864835	16.74	0.39	43.0
NM_001389	DSCAM	chr21:40304212-41142909	1.55	0.04	36.8
NM_018940	PCDHB7	chr5:140530426-140538141	22.85	0.74	30.9
NM_173851	SLC30A8	chr8:118214517-118260134	27.04	0.89	30.5
NM_001128929	ROBO2	chr3:77227852-77781353	12.94	0.51	25.3
NM_002942	ROBO2	chr3:77169983-	12.53	0.53	23.4

		77781353			
NM_001004492	OR2B11	chr1:245678953-245683907	1.39	0.06	23.0
NM_001083619	GRIA2	chr4:158359185-158508676	9.80	0.44	22.5
NM_000826	GRIA2	chr4:158359185-158508676	9.80	0.44	22.5
NM_001083620	GRIA2	chr4:158359269-158508676	9.75	0.44	22.3
NM_017433	MYO3A	chr10:26261007-26543471	6.25	0.31	20.0
NM_003469	SCG2	chr2:224167901-224177365	45.95	3.09	14.9
NM_003787	NOL4	chr18:29683061-30059444	13.58	0.91	14.9
NM_020225	STOX2	chr4:185061502-185177869	2.43	0.19	12.9
NM_153456	HS6ST3	chr13:95539093-96291813	4.23	0.33	12.8
NM_020297	ABCC9	chr12:21839590-21982895	3.12	0.26	12.0
NM_005691	ABCC9	chr12:21847374-21982895	3.16	0.27	11.5
NM_020298	ABCC9	chr12:21847374-21982895	3.16	0.27	11.5
NM_178177	NMNAT3	chr3:140759722-140881530	1.58	0.15	10.5
NM_001040429	PCDH17	chr13:57101789-57203066	20.98	2.00	10.5
NM_020872	CNTN3	chr3:74392411-74655033	0.38	0.04	10.4
NM_021952	ELAVL4	chr1:50345224-50441643	5.89	0.60	9.8
NM_024944	CHODL	chr21:18537020-18563558	4.53	0.46	9.8
NM_014227	SLC5A4	chr22:30942462-30983319	2.89	0.31	9.4
NM_194300	CCDC129	chr7:31521502-31661828	1.04	0.11	9.3
NM_001001850	STX19	chr3:95213904-95232144	17.74	2.25	7.9
NM_004770	KCNB2	chr8:73610179-74015138	5.78	0.77	7.5
NM_000818	GAD2	chr10:26543599-26635493	4.69	0.66	7.1
NM_001008539	SLC7A2	chr8:17443205-17474352	12.11	1.70	7.1
NM_152573	RASEF	chr9:84785136-84869863	1.35	0.20	6.9
NM_002500	NEUROD1	chr2:182247438-182255626	37.76	5.69	6.6
NM_206857	RTN1	chr14:59130446-59266184	4.00	0.66	6.1

NM_020203	MEPE	chr4:88971163-88988968	23.56	3.91	6.0
NM_017419	ACCN5	chr4:156968330-157008875	0.86	0.14	6.0
NM_152774	TMEM196	chr7:19723462-19781541	13.40	2.27	5.9
NM_000857	GUCY1B3	chr4:156897664-156949506	13.95	2.36	5.9
NM_001438	ESRRG	chr1:214741210-214965430	7.64	1.30	5.9
NM_002202	ISL1	chr5:50712714-50728320	12.03	2.06	5.9
NM_207303	ATRNL1	chr10:116841113-117700486	3.40	0.60	5.7
NM_198515	C10orf96	chr10:118071929-118131531	1.73	0.31	5.6
NM_015912	FAM135B	chr8:139209447-139580247	2.43	0.45	5.4
NM_198353	KCTD8	chr4:43869029-44147581	2.79	0.52	5.4
NM_020783	SYT4	chr18:39099854-39113342	28.56	5.38	5.3
NM_014510	PCLO	chr7:82285731-82632133	9.45	1.87	5.0
NM_080760	DACH1	chr13:70908098-71341331	22.05	4.44	5.0
NM_080759	DACH1	chr13:70908098-71341331	22.05	4.44	5.0
NM_004392	DACH1	chr13:70908098-71341331	22.05	4.44	5.0
NM_001005463	EBF3	chr10:131521536-131654081	1.24	0.25	4.9
NM_019065	NECAB2	chr16:82557737-82595880	2.14	0.46	4.6
NM_003749	IRS2	chr13:109202184-109238915	3.59	0.81	4.4
NM_014677	RIMS2	chr8:104898591-105336627	5.12	1.17	4.4
NM_002971	SATB1	chr3:18362437-18442344	20.93	4.87	4.3
NM_003360	UGT8	chr4:115760971-115819651	5.31	1.25	4.2
NM_019120	PCDHB8	chr5:140535613-140542205	17.75	4.20	4.2
NM_004921	CLCA3	chr1:86870546-86895647	5.43	1.29	4.2
NM_033026	PCLO	chr7:82219256-82632133	8.63	2.08	4.1
NM_014790	JAKMIP2	chr5:146948898-147144445	3.86	0.96	4.0
NM_005651	TDO2	chr4:157042296-157063000	2.90	0.72	4.0
NM_001013659	ZNF793	chr19:42687680-	2.10	0.55	3.8

		42728079			
NM_020973	GBA3	chr4:22301645-22432290	0.47	0.12	3.8
NM_001128432	GBA3	chr4:22301645-22432290	0.47	0.12	3.8
NM_030965	ST6GALNAC5	chr1:77103773-77304325	3.45	0.93	3.7
NM_001080544	LOC653314	chr5:177412995-177417888	7.26	1.96	3.7
NM_001085490	LOC285501	chr4:178884900-179150663	2.25	0.61	3.7
NM_201591	GPM6A	chr4:176789081-176973176	1.49	0.41	3.6
NM_198281	GPRIN3	chr4:90382451-90450184	17.14	4.74	3.6
NM_206594	ESRRG	chr1:214741210-215331599	3.30	0.95	3.5
NM_206595	ESRRG	chr1:214741210-215331599	3.30	0.95	3.5
NM_178011	LRRTM3	chr10:68353797-68532873	5.74	1.65	3.5
NM_139211	HOPX	chr4:57206920-57219408	8.06	2.32	3.5
NM_175929	FGF14	chr13:101169205-101854125	3.75	1.10	3.4
NM_004115	FGF14	chr13:101169205-101368996	5.37	1.58	3.4
NM_021136	RTN1	chr14:59130446-59409310	5.00	1.47	3.4
NM_014682	ST18	chr8:53183951-53486856	9.45	2.94	3.2
NM_001100117	RIMS2	chr8:104580151-105336627	4.55	1.42	3.2
NM_020685	C3orf14	chr3:62278435-62296360	6.58	2.10	3.1
NM_031501	PCDHA5	chr5:140179544-140185995	6.23	2.02	3.1
NM_001104629	C4orf19	chr4:37129946-37273527	4.37	1.41	3.1
NM_001076682	NCAM1	chr11:112335204-112643129	2.80	0.91	3.1
NM_004570	PIK3C2G	chr12:18303740-18694619	1.08	0.35	3.1
NM_201572	CACNB2	chr10:18467815-18872694	4.15	1.39	3.0
NM_201571	CACNB2	chr10:18467815-18872694	4.15	1.39	3.0
NM_201597	CACNB2	chr10:18467611-18872694	4.15	1.39	3.0
NM_201596	CACNB2	chr10:18467611-18872694	4.15	1.39	3.0
NM_201593	CACNB2	chr10:18467611-18872694	4.15	1.39	3.0

NM_004822	NTN1	chr17:8863583-9090042	0.57	0.19	3.0
NM_020346	SLC17A6	chr11:22314242-22359619	2.45	0.82	3.0
NM_000248	MITF	chr3:70066442-70102177	1.40	0.47	3.0
NM_198158	MITF	chr3:70066442-70102177	1.40	0.47	3.0
NM_206852	RTN1	chr14:59130446-59169273	2.01	0.69	2.9
NM_003936	CDK5R2	chr2:219530641-219537121	13.07	4.60	2.8
NM_144596	TTC8	chr14:88358730-88416088	22.12	7.96	2.8
NM_198310	TTC8	chr14:88358730-88416088	22.12	7.96	2.8
NM_198309	TTC8	chr14:88358730-88416088	22.12	7.96	2.8
NM_181351	NCAM1	chr11:112335204-112656368	2.68	0.97	2.8
NM_000615	NCAM1	chr11:112335204-112656368	2.68	0.97	2.8
NM_015678	NBEA	chr13:34412455-35146873	7.12	2.56	2.8
NM_000343	SLC5A1	chr22:30767258-30838645	1.31	0.48	2.7
NM_198178	MITF	chr3:70008945-70102177	1.59	0.59	2.7
NM_175768	GRIK2	chr6:101951625-102626651	3.76	1.42	2.6
NM_021956	GRIK2	chr6:101951625-102626651	3.76	1.42	2.6
NM_133448	TMEM132D	chr12:128120223-128956165	0.81	0.32	2.6
NM_174937	TCERG1L	chr10:132778644-133001974	0.47	0.18	2.6
NM_153331	KCTD6	chr3:58457131-58465127	6.60	2.59	2.6
NM_080872	UNC5D	chr8:35519451-35773722	0.32	0.13	2.5
NM_000724	CACNB2	chr10:18587589-18872694	3.28	1.32	2.5
NM_018168	C14orf105	chr14:57004347-57032329	7.69	3.12	2.5
NM_138818	PRUNE2	chr9:78626000-78712823	4.46	1.81	2.5
NM_153604	MYOCD	chr17:12508230-12609686	0.88	0.37	2.4
NM_002772	PRSS7	chr21:18561560-18699844	0.99	0.41	2.4
NM_001843	CNTN1	chr12:39370624-39752361	5.01	2.09	2.4
NM_175038	CNTN1	chr12:39370624-	5.01	2.09	2.4

		39752361			
NM_173560	RFXDC1	chr6:117303068-117362007	28.29	11.96	2.4
NM_001285	CLCA1	chr1:86705113-86740562	6.78	2.89	2.3
NM_015480	PVRL3	chr3:112271554-112337752	11.47	4.91	2.3
NM_000855	GUCY1A2	chr11:106061119-106396381	0.51	0.22	2.3
NM_005068	SIM1	chr6:100941470-101020272	8.30	3.60	2.3
NM_015888	HOOK1	chr1:60051120-60116638	20.56	9.04	2.3
NM_078625	VNN3	chr6:133083618-133099596	16.04	7.09	2.3
NM_018399	VNN3	chr6:133083618-133099596	16.04	7.09	2.3
NM_001024460	VNN3	chr6:133083618-133099596	16.04	7.09	2.3
NM_001008781	FAT3	chr11:91722909-92271283	1.03	0.46	2.3
NM_003046	SLC7A2	chr8:17438664-17474352	15.36	6.83	2.3
NM_178532	RNF180	chr5:63495426-63551095	8.66	3.86	2.2
NM_018349	MCTP2	chr15:92640498-92825267	4.83	2.16	2.2
NM_020856	TSHZ3	chr19:36455690-36534030	0.86	0.39	2.2
NM_130902	COX7B2	chr4:46429603-46608009	0.17	0.08	2.2
NM_000347	SPTB	chr14:64300901-64361619	1.90	0.86	2.2
NM_030632	ASXL3	chr18:29410538-29583397	6.22	2.84	2.2
NM_201590	CACNB2	chr10:18667619-18872694	2.36	1.08	2.2
NM_002753	MAPK10	chr4:87154655-87595307	4.28	1.96	2.2
NM_003054	SLC18A2	chr10:118988705-119029085	2.43	1.12	2.2
NM_033196	ZNF682	chr19:19974226-20013277	0.98	0.45	2.2
NM_001077349	ZNF682	chr19:19974226-20013064	0.98	0.45	2.2
NM_003658	BARX2	chr11:128749090-128829384	1.25	0.58	2.2
NM_152744	SDK1	chr7:3305605-4277157	2.89	1.36	2.1
NM_145001	STK32A	chr5:146592771-146710585	1.58	0.76	2.1
NM_015879	ST8SIA3	chr18:53168718-53189159	6.77	3.29	2.1
NM_001112719	LIMCH1	chr4:41307675-	4.56	2.22	2.1

		41398818			
NM_001112720	LIMCH1	chr4:41307675-41398818	4.56	2.22	2.1
NM_003122	SPINK1	chr5:147182335-147193453	4.19	2.05	2.0
NM_020864	KIAA1486	chr2:225971845-226228978	4.40	2.16	2.0
NM_020724	RNF150	chr4:142004174-142276066	1.49	0.73	2.0
NM_001017970	TMEM30B	chr14:60811841-60820283	19.51	9.93	2.0
NM_001127612	PAX6	chr11:31760915-31798085	8.39	4.29	2.0
NM_001604	PAX6	chr11:31760915-31791455	10.21	5.23	2.0
NM_000280	PAX6	chr11:31760915-31791455	10.21	5.23	2.0
NM_145913	SLC5A8	chr12:100072124-100130147	0.46	0.24	1.9
NM_000352	ABCC8	chr11:17369007-17457025	2.91	1.51	1.9
NM_014814	PSMD6	chr3:63969270-63986160	13.06	6.81	1.9
NM_001128085	ASPA	chr17:3322153-3351450	3.68	1.94	1.9
NM_020866	KLHL1	chr13:69170725-69582460	2.08	1.10	1.9
NM_001113561	RNF180	chr5:63495426-63706452	4.56	2.41	1.9
NM_019851	FGF20	chr8:16892704-16906045	1.17	0.62	1.9
NM_138980	MAPK10	chr4:87154655-87502240	4.55	2.43	1.9
NM_002374	MAP2	chr2:210150647-210309079	10.10	5.45	1.9
NM_031845	MAP2	chr2:210150647-210309079	10.10	5.45	1.9
NM_031847	MAP2	chr2:210150647-210309079	10.10	5.45	1.9
NM_000049	ASPA	chr17:3324045-3351450	3.78	2.07	1.8
NM_213599	TMEM16E	chr11:22169297-22259975	3.34	1.83	1.8
NM_006198	PCP4	chr21:40159216-40225192	3.06	1.68	1.8
NM_001128174	UGT8	chr4:115737059-115819651	7.68	4.24	1.8
NM_005925	MEP1B	chr18:28021984-28056364	4.05	2.23	1.8
NM_001100391	RALYL	chr8:85257654-85998633	0.92	0.51	1.8
NM_173848	RALYL	chr8:85256140-85998633	0.92	0.51	1.8

NM_001100392	RALYL	chr8:85256007-85998633	0.92	0.51	1.8
NM_001100393	RALYL	chr8:85256007-85998633	0.92	0.51	1.8
NM_001080476	GRXCR1	chr4:42588040-42729432	0.56	0.31	1.8
NM_138982	MAPK10	chr4:87154655-87496767	4.09	2.30	1.8
NM_182948	PRKACB	chr1:84380539-84478769	20.31	11.44	1.8
NM_001080463	DYNC2H1	chr11:102483369-102857801	3.78	2.13	1.8
NM_002407	SCGB2A1	chr11:61730715-61739987	4.55	2.58	1.8
NM_001039538	MAP2	chr2:209995015-210309079	14.17	8.08	1.8
NM_080818	OXGR1	chr13:96433973-96446605	6.72	3.84	1.8
NM_001112724	STK32A	chr5:146592771-146745961	1.57	0.90	1.7
NM_012128	CLCA4	chr1:86783346-86821020	1.03	0.59	1.7
NM_004248	PRLHR	chr10:120340905-120347150	4.32	2.48	1.7
NM_001080477	ODZ3	chr4:183480130-183963171	4.40	2.53	1.7
NM_001099	ACPP	chr3:133516901-133562379	11.24	6.54	1.7
NM_001005527	FAM19A4	chr3:68861606-69066401	0.50	0.29	1.7
NM_182522	FAM19A4	chr3:68861606-69066401	0.50	0.29	1.7
NM_022901	LRRC19	chr9:26981585-26997670	8.39	4.91	1.7
NM_175607	CNTN4	chr3:2115246-3076645	3.25	1.92	1.7
NM_033126	PSKH2	chr8:87127806-87152967	2.03	1.21	1.7
NM_002062	GLP1R	chr6:39122534-39165498	1.82	1.09	1.7
NM_000901	NR3C2	chr4:149217364-149585093	5.35	3.20	1.7
NM_007123	USH2A	chr1:214411914-214665361	2.57	1.54	1.7
NM_022571	GPR135	chr14:58997992-59003812	9.79	5.88	1.7
NM_015206	KIAA1024	chr15:77509912-77553697	1.62	0.97	1.7
NM_002612	PDK4	chr7:95048744-95065861	19.82	11.91	1.7
NM_001873	CPE	chr4:166517546-166640932	11.71	7.06	1.7
NM_000722	CACNA2D1	chr7:81415353-81912967	8.48	5.16	1.6

NM_024603	C1orf165	chr1:48964126-49017134	1.34	0.82	1.6
NM_001100916	MBOAT4	chr8:30106728-30123742	11.04	6.78	1.6
NM_016557	CCRL1	chr3:133796783-133806072	11.42	7.06	1.6
NM_201592	GPM6A	chr4:176789081-177162642	1.06	0.65	1.6
NM_005277	GPM6A	chr4:176789081-177162642	1.06	0.65	1.6
NM_032785	AGBL4	chr1:48769113-50264213	1.06	0.66	1.6
NM_138981	MAPK10	chr4:87154655-87249830	5.85	3.65	1.6
NM_001012428	ASB11	chrX:15209547-15244590	1.64	1.03	1.6
NM_080873	ASB11	chrX:15209547-15245648	1.59	1.00	1.6
NM_001007470	TRPM3	chr9:72586597-72675794	1.33	0.83	1.6
NM_206948	TRPM3	chr9:72586597-72675794	1.33	0.83	1.6
NM_134431	SLCO1A2	chr12:21309650-21441638	4.32	2.72	1.6
NM_001083592	ROR1	chr1:64010277-64383640	2.39	1.53	1.6
NM_002338	LSAMP	chr3:117009831-117649068	4.28	2.76	1.5
NM_001042406	HMGCLL1	chr6:55405129-55553971	11.67	7.55	1.5
NM_019036	HMGCLL1	chr6:55405129-55553971	11.67	7.55	1.5
NM_021153	CDH19	chr18:62320300-62424196	3.46	2.24	1.5
NM_032606	CAPS2	chr12:73954025-74012103	5.47	3.56	1.5
NM_001112812	GRIA4	chr11:104984620-105291441	2.32	1.52	1.5
NM_001077244	GRIA4	chr11:104984009-105291441	2.32	1.52	1.5
NM_201570	CACNB2	chr10:18727518-18872694	2.29	1.50	1.5
NM_022351	NECAB1	chr8:91870953-92042806	2.33	1.53	1.5
NM_001012393	OPCML	chr11:131788084-132909613	0.60	0.40	1.5
NM_139212	HOPX	chr4:57206920-57244322	5.14	3.46	1.5
NM_032495	HOPX	chr4:57206920-57244322	5.14	3.46	1.5
NM_001101320	LOC647174	chr13:50811168-50836240	2.02	1.37	1.5
NM_007038	ADAMTS5	chr21:27210102-	6.53	4.42	1.5

		27263310			
NM_032435	KIAA1804	chr1:231528136-231589517	7.65	5.19	1.5
NM_001039580	MAP9	chr4:156481261-156519572	9.23	6.31	1.5
NM_201548	CERKL	chr2:182107649-182231996	15.57	10.69	1.5
NM_153377	LRIG3	chr12:57550203-57602529	18.32	12.58	1.5
NM_001030313	CERKL	chr2:182109155-182231978	15.17	10.49	1.4
NM_001030312	CERKL	chr2:182109155-182231978	15.17	10.49	1.4
NM_001030311	CERKL	chr2:182109155-182231978	15.17	10.49	1.4
NM_000329	RPE65	chr1:68665094-68690230	2.74	1.90	1.4
NM_148898	FOXP2	chr7:113840287-114120328	16.43	11.54	1.4
NM_148899	FOXP2	chr7:113840287-114120328	16.43	11.54	1.4
NM_014491	FOXP2	chr7:113840287-114120328	16.43	11.54	1.4
NM_207491	MGC48628	chr4:91373204-91924174	2.81	2.00	1.4
NM_022824	FBXL17	chr5:107221347-107747010	8.39	5.99	1.4
NM_004734	DCLK1	chr13:35241477-35605443	0.94	0.68	1.4
NM_198177	MITF	chr3:69996131-70102177	1.76	1.28	1.4
NM_020741	KIAA1257	chr3:130170471-130197676	3.67	2.67	1.4
NM_001040428	SPATA7	chr14:87919764-87976557	7.41	5.42	1.4
NM_018418	SPATA7	chr14:87919764-87976557	7.41	5.42	1.4
NM_000232	SGCB	chr4:52579628-52601203	7.67	5.63	1.4
NM_182644	EPHA3	chr3:89237363-89534185	0.60	0.44	1.4
NM_207578	PRKACB	chr1:84314332-84445567	20.02	14.73	1.4
NM_000919	PAM	chr5:102227425-102395316	13.05	9.62	1.4
NM_138766	PAM	chr5:102227425-102395316	13.05	9.62	1.4
NM_138821	PAM	chr5:102227425-102395316	13.05	9.62	1.4
NM_138822	PAM	chr5:102227425-102395316	13.05	9.62	1.4
NM_152788	ANKS1B	chr12:97651201-98904563	1.63	1.21	1.3

NM_031889	ENAM	chr4:71711324-71733400	6.66	4.98	1.3
NM_014802	KIAA0528	chr12:22490784-22590719	12.32	9.23	1.3
NM_006203	PDE4D	chr5:58298622-58920081	8.11	6.09	1.3
NM_006914	RORB	chr9:76300071-76493937	1.71	1.28	1.3
NM_004795	KL	chr13:32486570-32540279	6.38	4.80	1.3
NM_014648	DZIP3	chr3:109789273-109898383	9.65	7.27	1.3
NM_152489	UBE2U	chr1:64440077-64484615	5.72	4.34	1.3
NM_001024611	LRRC66	chr4:52552622-52580543	5.76	4.37	1.3
NM_144591	C10orf32	chr10:104602008-104615960	13.12	9.99	1.3
NM_031457	MS4A8B	chr11:60221622-60241861	3.57	2.72	1.3
NM_001453	FOXC1	chr6:1553679-1561128	10.97	8.38	1.3
NM_152290	C1orf158	chr1:12726749-12745689	2.54	1.94	1.3
NM_145243	OMA1	chr1:58716978-58787034	12.48	9.57	1.3
NM_203403	C9orf150	chr9:12763011-12815059	22.00	16.87	1.3
NM_021073	BMP5	chr6:55726195-55850334	22.54	17.30	1.3
NM_005233	EPHA3	chr3:89237363-89615974	0.47	0.36	1.3
NM_006022	TSC22D1	chr13:43903658-43910979	37.41	28.88	1.3
NM_206937	LIG4	chr13:107655792-107667883	9.01	6.97	1.3
NM_002312	LIG4	chr13:107655792-107667131	9.61	7.43	1.3
NM_002731	PRKACB	chr1:84314332-84478769	16.62	12.89	1.3
NM_018394	ABHD10	chr3:113178517-113196900	21.70	16.91	1.3
NM_207361	FREM2	chr13:38157172-38361267	0.31	0.24	1.3
NM_032859	ABHD13	chr13:107666763-107686604	24.48	19.10	1.3
NM_014576	A1CF	chr10:52234330-52317441	12.85	10.03	1.3
NM_138932	A1CF	chr10:52234330-52317441	12.85	10.03	1.3
NM_138933	A1CF	chr10:52234330-52317441	12.85	10.03	1.3
NM_000810	GABRA5	chr15:24661150-24778749	0.51	0.40	1.3

NM_206933	USH2A	chr1:213860858-214665361	2.09	1.63	1.3
NM_177398	LMX1A	chr1:163435728-163593641	0.89	0.69	1.3
NM_001014797	KCNMA1	chr10:78297367-79069583	3.28	2.58	1.3
NM_052953	LRRC3B	chr3:26637303-26729269	0.33	0.26	1.3
NM_001103184	FMN1	chr15:30851636-31149377	6.41	5.06	1.3
NM_018315	FBXW7	chr4:153459859-153495560	26.95	21.27	1.3
NM_019090	KIAA1383	chr1:231005260-231014715	22.17	17.61	1.3
NM_002247	KCNMA1	chr10:78312640-79069583	3.31	2.63	1.3
NM_207322	FAM148A	chr15:60144467-60152408	24.96	19.98	1.2
NM_021936	PAPPA2	chr1:174696929-174928964	3.16	2.56	1.2
NM_175056	ZPLD1	chr3:103634548-103683375	3.56	2.88	1.2
NM_001113380	RGS4	chr1:161306318-161315216	22.04	17.89	1.2
NM_003429	ZNF85	chr19:20895919-20927343	2.32	1.89	1.2
NM_001031804	MAF	chr16:78183731-78194112	11.40	9.31	1.2
NM_005360	MAF	chr16:78183731-78194112	11.40	9.31	1.2
NM_173642	FAM80A	chr1:42617054-42664487	1.45	1.19	1.2
NM_015978	TNNI3K	chr1:74471672-74784696	1.21	0.99	1.2
NM_021255	PELI2	chr14:55652845-55839784	6.02	4.93	1.2
NM_001111061	NHLH2	chr1:116178521-116186856	3.79	3.13	1.2
NM_005599	NHLH2	chr1:116178521-116187270	3.61	2.98	1.2
NM_183422	TSC22D1	chr13:43903654-44050701	10.00	8.31	1.2
NM_172315	MEIS2	chr15:34968523-35179795	11.71	9.73	1.2
NM_001004303	C1orf168	chr1:56955064-57059957	5.97	4.97	1.2
NM_001083907	BANK1	chr4:102952005-103216992	2.46	2.04	1.2
NM_005941	MMP16	chr8:89116575-89410833	4.70	3.92	1.2
NM_015009	PDZRN3	chr3:73512341-73758762	4.59	3.85	1.2
NM_032138	KBTBD7	chr13:40661710-	20.29	17.12	1.2

		40668702			
NM_152606	ZNF540	chr19:42732147-42798836	1.82	1.54	1.2
NM_052907	TMEM132B	chr12:124375114-124711542	1.36	1.16	1.2
NM_004801	NRXN1	chr2:49998991-51115178	4.81	4.12	1.2
NM_139167	SGCZ	chr8:13989743-15142163	0.31	0.26	1.2
NM_006217	SERPINI2	chr3:168640416-168676512	1.16	0.99	1.2
NM_033326	SOX6	chr11:15946370-16456494	14.67	12.68	1.2
NM_005964	MYH10	chr17:8316254-8476761	6.36	5.51	1.2
NM_025075	THOC7	chr3:63792585-63826637	6.67	5.78	1.2
NM_173808	NEGR1	chr1:71639212-72522865	2.61	2.27	1.2
NM_152321	ERP27	chr12:14956244-14984722	2.55	2.22	1.2
NM_001101669	INPP4B	chr4:143166631-143989054	3.38	2.95	1.1
NM_003866	INPP4B	chr4:143166631-143989054	3.38	2.95	1.1
NM_003716	CADPS	chr3:62357060-62838094	2.29	2.00	1.1
NM_183393	CADPS	chr3:62357060-62838094	2.29	2.00	1.1
NM_183394	CADPS	chr3:62357060-62838094	2.29	2.00	1.1
NM_172316	MEIS2	chr15:34968523-35180889	11.65	10.20	1.1
NM_144996	ARL13B	chr3:95179671-95258813	8.98	7.87	1.1
NM_182896	ARL13B	chr3:95179671-95258813	8.98	7.87	1.1
NM_144981	IMMP1L	chr11:31408524-31489745	21.92	19.23	1.1
NM_001042784	FLJ25770	chr4:77451215-77549482	5.28	4.64	1.1
NM_017508	SOX6	chr11:15946370-16382968	15.76	13.85	1.1
NM_017970	C14orf102	chr14:89812150-89870032	4.04	3.57	1.1
NM_199043	C14orf102	chr14:89812150-89870032	4.04	3.57	1.1
NM_024114	TRIM48	chr11:54784233-54797171	0.80	0.70	1.1
NM_003453	ZMYM2	chr13:19428809-19560939	8.61	7.64	1.1
NM_197968	ZMYM2	chr13:19428809-19560939	8.61	7.64	1.1

NM_001127384	CTNNA3	chr10:67347730-69097422	1.39	1.23	1.1
NM_080874	ASB5	chr4:177369821-177429269	1.73	1.53	1.1
NM_020431	TMEM63C	chr14:76715854-76797591	1.22	1.09	1.1
NM_052959	PANX3	chr11:123984662-123997461	2.46	2.20	1.1
NM_001015887	IGSF11	chr3:120100168-120238366	2.67	2.39	1.1
NM_000926	PGR	chr11:100403564-100507754	4.73	4.25	1.1
NM_025114	CEP290	chr12:86964920-87062124	9.19	8.31	1.1
NM_002399	MEIS2	chr15:34968523-35182792	11.67	10.57	1.1
NM_153262	SYT14	chr1:208176160-208406256	4.22	3.83	1.1
NM_019040	ELP4	chr11:31485872-31763905	14.38	13.06	1.1
NM_032679	ZNF577	chr19:57064361-57085009	2.23	2.02	1.1
NM_170675	MEIS2	chr15:34968523-35181996	11.59	10.61	1.1
NM_170674	MEIS2	chr15:34968523-35181996	11.59	10.61	1.1
NM_170676	MEIS2	chr15:34968523-35181996	11.59	10.61	1.1
NM_170677	MEIS2	chr15:34968523-35181996	11.59	10.61	1.1
NM_021161	KCNK10	chr14:87718998-87865004	1.09	1.00	1.1
NM_138318	KCNK10	chr14:87718998-87809008	1.77	1.62	1.1
NM_138317	KCNK10	chr14:87718998-87861100	1.12	1.03	1.1
NM_152538	IGSF11	chr3:120100168-120349588	1.86	1.70	1.1
NM_001105531	FAM135A	chr6:71177827-71329596	12.90	11.91	1.1
NM_020819	FAM135A	chr6:71177827-71329596	12.90	11.91	1.1
NM_001098517	CADM1	chr11:114547554-114882451	4.41	4.08	1.1
NM_014333	CADM1	chr11:114547554-114882451	4.41	4.08	1.1
NM_173559	C6orf224	chr6:109918756-109936386	4.00	3.70	1.1
NM_032036	FAM14A	chr14:93661870-93667710	7.77	7.19	1.1
NM_018972	GDAP1	chr8:75423172-75443890	10.60	9.81	1.1
NM_001040875	GDAP1	chr8:75423207-	10.62	9.82	1.1

		75443890			
NM_033437	PDE5A	chr4:120632997-120770650	7.22	6.69	1.1
NM_033430	PDE5A	chr4:120632997-120769890	7.26	6.73	1.1
NM_001083	PDE5A	chr4:120632997-120771429	7.18	6.66	1.1
NM_022564	MMP16	chr8:89148576-89410833	4.52	4.20	1.1
NM_014980	STXBP5L	chr3:122107739-122628298	1.47	1.37	1.1
NM_001102445	RGS4	chr1:161303019-161315216	27.15	25.30	1.1
NM_001113381	RGS4	chr1:161303667-161315216	28.67	26.72	1.1
NM_005613	RGS4	chr1:161303387-161315216	27.99	26.09	1.1
NM_213606	SLC16A12	chr10:91178035-91287293	6.97	6.51	1.1
NM_183387	EML5	chr14:88148955-88330910	3.66	3.44	1.1
NM_018717	MAML3	chr4:140854995-141296683	9.99	9.40	1.1
NM_000829	GRIA4	chr11:104984009-105360029	2.13	2.00	1.1
NM_001077243	GRIA4	chr11:104984009-105360029	2.13	2.00	1.1
NM_152279	ZNF585B	chr19:42365561-42395291	2.65	2.51	1.1
NM_018179	ATF7IP	chr12:14407877-14544964	18.47	17.50	1.1
NM_001112717	LIMCH1	chr4:41055560-41398818	3.00	2.84	1.1
NM_001112718	LIMCH1	chr4:41055560-41398818	3.00	2.84	1.1
NM_014988	LIMCH1	chr4:41055560-41398818	3.00	2.84	1.1
NM_001111031	ACVR1C	chr2:158089524-158164318	11.05	10.49	1.1
NM_014841	SNAP91	chr6:84317331-84477495	2.18	2.07	1.1
NM_000306	POU1F1	chr3:87389472-87410427	6.33	6.03	1.0
NM_001122757	POU1F1	chr3:87389472-87410427	6.33	6.03	1.0
NM_024829	FLJ22662	chr12:14545863-14614058	0.97	0.93	1.0
NM_153184	CADM2	chr3:85856321-86202640	0.58	0.56	1.0
NM_003383	VLDLR	chr9:2609792-2646485	5.54	5.34	1.0
NM_001018056	VLDLR	chr9:2609792-2646485	5.54	5.34	1.0
NM_004253	PLAA	chr9:26892517-26927207	3.90	3.77	1.0

NM_012301	MAGI2	chr7:77482309-78922826	3.12	3.01	1.0
NM_005012	ROR1	chr1:64010277-64419295	2.29	2.21	1.0
NM_203282	ZNF254	chr19:24059815-24106494	1.28	1.24	1.0
NM_001093734	LOC645441	chr1:77364632-77369995	8.24	8.09	1.0
NM_013361	ZNF223	chr19:49246003-49265982	3.09	3.04	1.0
NM_001080440	OTOL1	chr3:162695289-162706424	3.25	3.21	1.0
NM_001033602	KIAA0774	chr13:28494747-28980084	1.15	1.14	1.0
NM_021635	PBOV1	chr6:138576819-138583320	9.11	9.00	1.0
NM_012419	RGS17	chr6:153371724-153496082	8.78	8.69	1.0
NM_052867	NALCN	chr13:100502130-100868814	2.27	2.25	1.0
NM_002069	GNAI1	chr7:79600075-79688661	12.51	12.38	1.0
NM_021176	G6PC2	chr2:169463995-169476756	34.72	34.43	1.0
NM_001081686	G6PC2	chr2:169463995-169476756	34.72	34.43	1.0
NM_144979	RBM46	chr4:155919876-155971414	3.32	3.31	1.0
NM_138453	RAB3C	chr5:57912695-58185163	4.49	4.48	1.0
NM_145695	DGKB	chr7:14180130-14849600	0.99	0.99	1.0
NM_001013356	OR8U8	chr11:55897675-56267136	0.27	0.27	1.0

A. Calculated as the percentage of bases across transcript that overlap moderate FAIRE-enriched site

B. Region 2 kb upstream through 2 kb downstream of transcript (hg18)

C. Moderate peaks for HeLa-S3, HUVEC, GM12878, HepG2 and K562 were merged into one peak set

Table 5.4. Referenced list of genes with preferential islet FAIRE enrichment that are known to be expressed in islet-cells in a selective manner

Gene symbol	Reference
IAPP	Nishi, M., Sanke, T., Nagamatsu, S., Bell, G.I. & Steiner, D.F. Islet amyloid polypeptide. A new beta cell secretory product related to islet amyloid deposits. *J Biol Chem* **265**, 4173-6 (1990).
NKX6-1	Sander, M. et al. Homeobox gene Nkx6.1 lies downstream of Nkx2.2 in the major pathway of beta-cell formation in the pancreas. Development 127, 5533-40 (2000).
FOXD3	Perera, H.K. et al. Expression and shifting subcellular localization of the transcription factor, Foxd3, in embryonic and adult pancreas. *Gene Expr Patterns* **6**, 971-7 (2006).
PDX1	Ohlsson, H., Karlsson, K. & Edlund, T. IPF1, a homeodomain-containing transactivator of the insulin gene. *EMBO J* **12**, 4251-9 (1993).
SULT4A1	Falany, C.N., Xie, X., Wang, J., Ferrer, J. & Falany, J.L. Molecular cloning and expression of novel sulphotransferase-like cDNAs from human and rat brain. Biochem J 346 Pt 3, 857-64 (2000).
SCG2	Karlsson, E. The role of pancreatic chromogranins in islet physiology. Curr Mol Med 1, 727-32 (2001).
GAD2	Kaufman, D.L. et al. Autoimmunity to two forms of glutamate decarboxylase in insulin-dependent diabetes mellitus. J Clin Invest 89, 283-92 (1992).
NEUROD1	Naya, F.J. et al. Diabetes, defective pancreatic morphogenesis, and abnormal enteroendocrine differentiation in BETA2/neuroD-deficient mice. Genes Dev 11, 2323-34 (1997).
ISL1	Ahlgren, U., Pfaff, S.L., Jessell, T.M., Edlund, T. & Edlund, H. Independent requirement for ISL1 in formation of pancreatic mesenchyme and islet cells. Nature 385, 257-60 (1997).
DACH1	Miyatsuka, T., Li, Z. & German, M.S. Chronology of islet differentiation revealed by temporal cell labeling. Diabetes 58, 1863-8 (2009).

ST18	Wang, S. et al. Loss of Myt1 function partially compromises endocrine islet cell differentiation and pancreatic physiological function in the mouse. Mech Dev 124, 898-910 (2007).
NCAM1	Esni, F. et al. Neural cell adhesion molecule (N-CAM) is required for cell type segregation and normal ultrastructure in pancreatic islets. J Cell Biol 144, 325-37 (1999).
CDK5R2	Lilja, L. et al. Cyclin-dependent kinase 5 associated with p39 promotes Munc18-1 phosphorylation and Ca(2+)-dependent exocytosis. J Biol Chem 279, 29534-41 (2004).
RFXDC1	Miyatsuka, T., Li, Z. & German, M.S. Chronology of islet differentiation revealed by temporal cell labeling. Diabetes 58, 1863-8 (2009).
PAX6	Sander, M. et al. Genetic analysis reveals that PAX6 is required for normal transcription of pancreatic hormone genes and islet development. Genes Dev 11, 1662-73 (1997).
ABCC8	Aguilar-Bryan, L. et al. Cloning of the beta cell high-affinity sulfonylurea receptor: a regulator of insulin secretion. Science 268, 423-6 (1995).
GLP1R	Thorens, B. Expression cloning of the pancreatic beta cell receptor for the gluco-incretin hormone glucagon-like peptide 1. Proc Natl Acad Sci U S A 89, 8641-5 (1992).
CPE	Naggert, J.K. et al. Hyperproinsulinaemia in obese fat/fat mice associated with a carboxypeptidase E mutation which reduces enzyme activity. Nat Genet 10, 135-42 (1995).
LMX1A	Iannotti, C.A. et al. Identification of a human LMX1 (LMX1.1)-related gene, LMX1.2: tissue-specific expression and linkage mapping on chromosome 9. Genomics 46, 520-4 (1997).
KCNK10	Kang, D., Choe, C. & Kim, D. Functional expression of TREK-2 in insulin-secreting MIN6 cells. Biochem Biophys Res Commun 323, 323-31 (2004).
G6PC2	Hutton, J.C. & Eisenbarth, G.S. A pancreatic beta-cell-specific homolog of glucose-6-phosphatase emerges as a major target of cell-mediated autoimmunity in diabetes. Proc Natl Acad Sci U S A 100, 8626-8 (2003).

KCNB2	Wolf-Goldberg, T. et al. Target soluble N-ethylmaleimide-sensitive factor attachment protein receptors (t-SNAREs) differently regulate activation and inactivation gating of Kv2.2 and Kv2.1: Implications on pancreatic islet cell Kv channels. Mol Pharmacol 70, 818-28 (2006).
SYT4	Gauthier, B.R. et al. Synaptotagmin VII splice variants alpha, beta, and delta are expressed in pancreatic beta-cells and regulate insulin exocytosis. FASEB J 22, 194-206 (2008).
SLC30A8	Sladek, R. et al. A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature 445, 881-5 (2007).

Table 5.5. Over- and under-represented transcription factor binding motifs

Motif	# chr (p<.01)[d]	# chr (p>.99)[e]	Motif	# chr (p<.01)	# chr (p>.99)	Motif	# chr (p<.01)	# chr (p>.99)
CTCF_main	23	-	MA0047 Foxa2 FORKHEAD	16	-	CTCF_main	22	-
V$SRY_01	22	-	MA0031 FOXD1 FORKHEAD	15	-	MA0060 NF-Y CAAT-BOX	7	-
V$TAL1ALPHAE47_01	21	-	CTCF_main	13	-	V$NFY_01	7	1
MA0020 Dof2 ZN-FINGER, DOF	19	-	MA0094 Ubx HOMEO	13	-	V$AP1_Q6	6	1
MA0031 FOXD1 FORKHEAD	19	-	V$SRY_01	12	-	MA0026 E74A ETS	5	-
V$AP1_C	19	-	V$RFX1_01	9	-	MA0080 SPI1 ETS	5	-
V$AP1_Q6	19	-	MA0053 MNB1A ZN-FINGER, DOF	9	-	MA0098 c-ETS ETS	5	-
MA0047 Foxa2 FORKHEAD	18	-	MA0089 TCF11-MafG bZIP	8	-	V$AP1FJ_Q2	5	2
MA0053 MNB1A ZN-FINGER, DOF	18	-	MA0020 Dof2 ZN-FINGER, DOF	8	-	MA0028 ELK1 ETS	4	-
MA0064 PBF ZN-FINGER, DOF	18	-	MA0064 PBF ZN-FINGER, DOF	8	-	MA0062 GABPA ETS	4	-
V$NFE2_01	18	-	MA0091 TAL1-TCF3 bHLH	7	-	MA0076 ELK4 ETS	4	-
V$NFAT_Q6	17	-	V$TAL1ALPHAE47_01	7	-	MA0099 Fos bZIP	4	1
V$TAL1BETAE47_01	16	-	V$TAL1BETAITF2_01	6	-	V$AP1_Q4	4	2
MA0091 TAL1-TCF3 bHLH	15	-	MA0072 RORA1 NUCLEAR RECEPTOR	5	-	V$ATF_01	4	-
MA0099 Fos bZIP	15	-	V$RORA2_01	5	-	V$CREB_Q2	4	-
V$AP1_Q2	15	-	MA0046 TCF1 HOMEO	4	-	V$CREB_Q4	4	-
V$AP1_Q4	15	-	V$FREAC4_01	4	-	MA0009 T T-BOX	3	-
V$TAL1BETAITF2_01	15	-	V$MIF1_01	4	-	MA0039 Klf4 ZN-FINGER, C2H2	3	-
MA0089 TCF11-MafG bZIP	13	-	V$NFAT_Q6	4	-	MA0120 ID1 ZN-FINGER, C2H2	3	1
MA0046 TCF1 HOMEO	12	-	V$RFX1_02	4	-	V$E47_01	3	-

V$AP1FJ_Q2	12	-	V$HNF1_01	3	-	V$HLF_01	3	2
V$RFX1_01	11	-	V$TAL1BETAE47_01	3	-	V$HSF1_01	3	-
MA0094 Ubx HOMEO	10	-	MA0059 MYC-MAX bHLH-ZIP	2	3	V$HSF2_01	3	-
V$HNF1_01	9	-	MA0075 Prrx2 HOMEO	2	-	V$NFE2_01	3	1
MA0084 SRY HMG	8	-	MA0084 SRY HMG	2	-	V$SP1_Q6	3	1
V$AREB6_04	8	-	MA0100 Myb TRP-CLUSTER	2	-	V$STAT_01	3	-
V$RFX1_02	7	-	V$AP1_C	2	-	MA0003 TFAP2A AP2	2	1
V$TGIF_01	7	-	V$CREBP1CJUN_01	2	1	MA0004 Arnt bHLH	2	-
MA0026 E74A ETS	6	-	V$EGR1_01	2	4	MA0007 Ar NUCLEAR RECEPTOR	2	-
MA0055 Myf bHLH	6	-	V$EGR2_01	2	3	MA0010 Broad-complex_1 ZN-FINGER, C2H2	2	4
V$RORA2_01	6	-	V$EGR3_01	2	4	MA0020 Dof2 ZN-FINGER, DOF	2	-
MA0072 RORA1 NUCLEAR RECEPTOR	5	-	V$RSRFC4_01	2	-	MA0025 NFIL3 bZIP	2	-
MA0052 MEF2A MADS	4	-	V$TGIF_01	2	-	MA0053 MNB1A ZN-FINGER, DOF	2	-
V$FREAC4_01	4	-	MA0005 Agamous MADS	1	2	MA0055 Myf bHLH	2	-
V$HNF1_C	4	-	MA0026 E74A ETS	1	1	MA0058 MAX bHLH-ZIP	2	-
V$ISRE_01	4	-	MA0040 Foxq1 FORKHEAD	1	-	MA0059 MYC-MAX bHLH-ZIP	2	-
V$MIF1_01	4	-	MA0048 NHLH1 bHLH	1	2	MA0093 USF1 bHLH-ZIP	2	1
V$RSRFC4_01	4	-	MA0052 MEF2A MADS	1	1	MA0100 Myb TRP-CLUSTER	2	-

MA0060 NF-Y CAAT-BOX	3	1	MA0054 MYB.ph3 TRP-CLUSTER	1	-	MA0101 REL REL	2	-
MA0076 ELK4 ETS	3	-	MA0055 Myf bHLH	1	1	MA0104 Mycn bHLH-ZIP	2	-
V$E47_01	3	-	MA0062 GABPA ETS	1	4	MA0123 ABI4 AP2	2	-
V$NFY_01	3	1	MA0087 Sox5 HMG	1	-	V$AP1_C	2	1
V$RORA1_01	3	-	MA0113 NR3C1 NUCLEAR	1	2	V$AP1_Q2	2	-
MA0021 Dof3 ZN-FINGER, DOF	2	-	MA0115 NR1H2-RXR NUCLEAR RECEPTOR	1	1	V$AP2_Q6	2	1
MA0050 IRF1 TRP-CLUSTER	2	-	V$AHRARNT_02	1	9	V$CREB_01	2	-
MA0071 RORA NUCLEAR RECEPTOR	2	-	V$AP1_Q6	1	-	V$CREBP1_01	2	-
MA0080 SPI1 ETS	2	5	V$AP4_Q5	1	7	V$E4BP4_01	2	-
MA0081 SPIB ETS	2	1	V$E4BP4_01	1	-	V$ELK1_02	2	-
MA0096 bZIP910 bZIP	2	-	V$FREAC3_01	1	-	V$NFAT_Q6	2	-
MA0097 bZIP911 bZIP	2	-	V$GATA1_04	1	-	V$NFY_C	2	-
MA0117 MafB bZIP, MAF	2	-	V$HFH3_01	1	-	V$NFY_Q6	2	-
MA0119 Hox11-CTF1 HOMEO/CAAT	2	-	V$HNF1_C	1	-	V$TAL1ALPHAE47_01	2	-
V$MEF2_03	2	-	V$MEF2_01	1	-	V$USF_Q6	2	1
V$MYCMAX_01	2	-	V$MYCMAX_01	1	1	MA0006 Arnt-Ahr bHLH	1	-
V$MYOD_Q6	2	1	V$NF1_Q6	1	2	MA0021 Dof3 ZN-FINGER, DOF	1	-
V$NFY_C	2	-	V$NFE2_01	1	-	MA0022 Dorsal_1 REL	1	-
V$NFY_Q6	2	3	V$NFY_C	1	-	MA0023 Dorsal_2 REL	1	-
MA0005 Agamous MADS	1	-	V$STAT1_01	1	-	MA0024 E2F1 Unknown	1	-
MA0012 Broad-complex_3 ZN-FINGER, C2H2	1	-	V$STAT3_01	1	-	MA0029 Evi1 ZN-FINGER, C2H2	1	3
MA0018 CREB1 bZIP	1	-	MA0007 Ar NUCLEAR RECEPTOR	-	1	MA0043 HLF bZIP	1	1
MA0024 E2F1 Unknown	1	-	MA0021 Dof3 ZN-FINGER,	-	1	MA0049	1	1

			DOF			Hunchback ZN-FINGER, C2H2		
						MA0056 ZNF42_1-4 ZN-FINGER,		
MA0028 ELK1 ETS	1	-	MA0032 FOXC1 FORKHEAD	-	1	C2H2	1	2
MA0030 FOXF2 FORKHEAD	1	-	MA0038 Gfi ZN-FINGER, C2H2	-	1	MA0064 PBF ZN-FINGER, DOF	1	-
MA0038 Gfi ZN-FINGER, C2H2	1	-	MA0078 Sox17 HMG	-	1	MA0067 Pax2 PAIRED	1	1
MA0040 Foxq1 FORKHEAD	1	-	MA0093 USF1 bHLH-ZIP	-	1	MA0081 SPIB ETS	1	-
MA0042 FOXI1 FORKHEAD	1	-	MA0097 bZIP911 bZIP	-	1	MA0088 Staf ZN-FINGER, C2H2	1	2
MA0059 MYC-MAX bHLH-ZIP	1	-	V$CDP_01	-	1	MA0091 TAL1-TCF3 bHLH	1	-
MA0062 GABPA ETS	1	-	V$CREB_Q4	-	1	MA0096 bZIP910 bZIP	1	-
MA0083 SRF MADS	1	2	V$CREBP1_Q2	-	2	MA0097 bZIP911 bZIP	1	-
MA0087 Sox5 HMG	1	-	V$FOXJ2_01	-	1	MA0117 MafB bZIP, MAF	1	2
MA0098 c-ETS ETS	1	3	V$GATA_C	-	1	MA0119 Hox11-CTF1 HOMEO/CAA T	1	2
MA0100 Myb TRP-CLUSTER	1	-	V$MYOGNF1_01	-	1	V$AP4_01	1	-
MA0113 NR3C1 NUCLEAR	1	1	V$OCT1_06	-	1	V$AP4_Q5	1	-
MA0116 Roaz ZN-FINGER, C2H2	1	1	V$OCT1_Q6	-	1	V$AP4_Q6	1	-
V$AP4_Q5	1	2	V$SRF_01	-	1	V$AREB6_04	1	-
V$AP4_Q6	1	4	MA0004 Arnt bHLH	-	2	V$ARNT_01	1	-
V$CREB_01	1	-	MA0017 NR2F1 NUCLEAR RECEPTOR	-	2	V$CHOP_01	1	3

V$CREBP1_01	1	-	MA0028 ELK1 ETS	-	2	V$CREB_02	1	1
V$CREBP1CJUN_01	1	-	MA0043 HLF bZIP	-	3	V$CREBP1_Q2	1	1
V$E2F_02	1	-	MA0044 HMG-1 HMG	-	2	V$CREBP1CJUN_01	1	-
V$E4BP4_01	1	1	MA0060 NF-Y CAAT-BOX	-	2	V$CREL_01	1	-
V$EGR1_01	1	2	MA0067 Pax2 PAIRED	-	2	V$E2F_02	1	-
V$EGR2_01	1	2	MA0070 Pbx HOMEO	-	2	V$ELK1_01	1	-
V$EGR3_01	1	2	MA0083 SRF MADS	-	2	V$FREAC3_01	1	-
V$FOXD3_01	1	2	V$ARNT_01	-	2	V$HNF4_01	1	-
V$FOXJ2_01	1	-	V$COUP_01	-	2	V$LMO2COM_01	1	2
V$GATA1_04	1	-	V$CREB_01	-	2	V$MAX_01	1	-
V$HFH3_01	1	-	V$CREB_Q2	-	2	V$MYCMAX_02	1	-
V$HSF2_01	1	4	V$E2F_01	-	3	V$NF1_Q6	1	1
V$MEF2_01	1	-	V$MEIS1_01	-	2	V$NFKAPPAB65_01	1	-
V$MEF2_02	1	-	V$NFKAPPAB65_01	-	2	V$PAX6_01	1	1
V$SRF_01	1	1	V$NFY_01	-	2	V$RFX1_02	1	-
V$STAT_01	1	-	V$SOX9_B1	-	2	V$SP1_01	1	1
V$STAT1_01	1	-	V$SRF_Q6	-	2	V$SRF_Q6	1	-
V$USF_Q6	1	2	V$TCF11MAFG_01	-	2	V$STAT1_01	1	-
MA0043 HLF bZIP	-	1	V$USF_01	-	2	V$STAT3_01	1	-
MA0048 NHLH1 bHLH	-	1	MA0018 CREB1 bZIP	-	3	V$TAL1BETAE47_01	1	-
MA0058 MAX bHLH-ZIP	-	1	MA0058 MAX bHLH-ZIP	-	3	V$TAL1BETAITF2_01	1	-
MA0065 PPARG-RXRA NUCLEAR RECEPTOR	-	1	MA0074 RXR-VDR NUCLEAR RECEPTOR	-	3	V$TAXCREB_01	1	2
MA0066 PPARG NUCLEAR RECEPTOR	-	1	MA0081 SPIB ETS	-	3	V$USF_01	1	1
MA0067 Pax2 PAIRED	-	1	MA0092 HAND1-TCF3 bHLH	-	3	MA0014 Pax5 PAIRED	-	1
MA0104 Mycn bHLH-ZIP	-	1	MA0104 Mycn bHLH-ZIP	-	3	MA0017 NR2F1 NUCLEAR	-	1

						RECEPTOR		
						MA0019		
MA0108 TBP TATA-box	-	1	MA0107 RELA REL	-	3	Chop-cEBP	-	1
						bZIP		
						MA0032		
V$ARNT_01	-	1	V$CEBP_Q2	-	4	FOXC1	-	1
						FORKHEAD		
						MA0035 Gata1		
V$CART1_01	-	1	V$CREB_02	-	3	ZN-FINGER,	-	1
						GATA		
						MA0041		
V$CDP_01	-	1	V$ELK1_02	-	3	Foxd3	-	1
						FORKHEAD		
						MA0044		
V$CREB_02	-	1	V$LMO2COM_02	-	3	HMG-1 HMG	-	1
						MA0045		
						HMG-IY		
V$ELK1_01	-	1	V$NFKB_C	-	3	HMG	-	1
						MA0051 IRF2		
						TRP-		
V$MYCMAX_02	-	1	V$P53_02	-	3	CLUSTER	-	1
						MA0068 Pax4		
						PAIRED-		
V$MYOGNF1_01	-	1	V$SRF_C	-	3	HOMEO	-	1
			MA0066 PPARG NUCLEAR			MA0070 Pbx		
V$NFKAPPAB65_01	-	1	RECEPTOR	-	4	HOMEO	-	1
						MA0073		
						RREB1 ZN-		
						FINGER,		
V$OCT1_03	-	1	V$CEBP_C	-	4	C2H2	-	1
						MA0074 RXR-		
						VDR		
						NUCLEAR		
V$OCT1_Q6	-	1	V$CEBPA_01	-	5	RECEPTOR	-	1
						MA0078		
V$PBX1_01	-	1	V$GATA1_03	-	4	Sox17 HMG	-	1
						MA0082		
V$SRF_Q6	-	1	V$HAND1E47_01	-	4	SQUA MADS	-	1
						MA0083 SRF		
V$TATA_01	-	1	V$NFY_Q6	-	4	MADS	-	1
V$TCF11MAFG_01	-	1	V$P53_01	-	4	MA0095 YY1	-	1

MA0007 Ar NUCLEAR RECEPTOR	-	2	V$TAXCREB_02	-	4	ZN-FINGER, C2H2 MA0106 TP53 P53	-	1
MA0017 NR2F1 NUCLEAR RECEPTOR	-	2	V$USF_02	-	5	MA0110 ATHB5 HOMEO-ZIP	-	1
MA0032 FOXC1 FORKHEAD	-	2	V$USF_Q6	-	4	MA0111 Spz1 bHLH-ZIP	-	1
MA0033 FOXL1 FORKHEAD	-	2	MA0015 CF2-II ZN-FINGER, C2H2	-	6	MA0112 ESR1 NUCLEAR	-	1
MA0041 Foxd3 FORKHEAD	-	2	MA0049 Hunchback ZN-FINGER, C2H2	-	6	MA0114 HNF4 NUCLEAR	-	1
MA0061 NF-kappaB REL	-	2	MA0082 SQUA MADS	-	5	V$AHRARNT _02	-	1
MA0078 Sox17 HMG	-	2	MA0116 Roaz ZN-FINGER, C2H2	-	6	V$ARP1_01	-	1
MA0107 RELA REL	-	2	V$CDPCR1_01	-	6	V$CEBPA_01	-	1
V$COUP_01	-	2	V$CDPCR3_01	-	5	V$E47_02	-	1
V$E2F_01	-	2	V$CEBPB_01	-	6	V$FREAC7_0 1	-	1
V$FREAC7_01	-	2	V$E47_01	-	6	V$GATA1_03	-	1
V$GATA_C	-	2	V$GRE_C	-	5	V$GATA1_04	-	1
V$GRE_C	-	2	V$MYOD_Q6	-	6	V$GATA3_01	-	1
V$NFKAPPAB_01	-	2	V$OCT1_04	-	5	V$HNF1_C	-	1
V$NFKB_C	-	2	MA0022 Dorsal_1 REL	-	6	V$IRF1_01	-	1
V$NRSF_01	-	2	MA0085 SU_h IPT/TIG domain	-	6	V$MEF2_02	-	1
V$PBX1_02	-	2	MA0106 TP53 P53	-	7	V$MYOD_Q6	-	1
V$SOX9_B1	-	2	V$AP4_Q6	-	7	V$MZF1_01	-	1
V$SRF_C	-	2	V$CDPCR3HD_01	-	7	V$MZF1_02	-	1
V$TAXCREB_02	-	2	V$USF_C	-	7	V$NKX61_01	-	1
MA0044 HMG-1 HMG	-	3	MA0061 NF-kappaB REL	-	7	V$OCT_C	-	1
MA0074 RXR-VDR NUCLEAR RECEPTOR	-	3	MA0090 TEAD TEA	-	7	V$OCT1_01	-	1
MA0085 SU_h IPT/TIG domain	-	3	MA0123 ABI4 AP2	-	7	V$OCT1_02	-	1

MA0092 HAND1-TCF3 bHLH	-	3	V$AP4_01	-	7	V$OCT1_03	-	1
MA0106 TP53 P53	-	3	V$NFKAPPAB_01	-	7	V$OCT1_07	-	1
V$AHRARNT_02	-	3	V$TST1_01	-	8	V$PAX5_01	-	1
V$AP4_01	-	3	MA0014 Pax5 PAIRED	-	9	V$RORA1_01	-	1
V$CDPCR3_01	-	3	MA0023 Dorsal_2 REL	-	8	V$TATA_C	-	1
V$HAND1E47_01	-	3	MA0080 SPI1 ETS	-	9	V$TCF11_01	-	1
V$LMO2COM_02	-	3	MA0105 NFKB1 REL	-	8	V$USF_C	-	1
V$NF1_Q6	-	3	MA0114 HNF4 NUCLEAR	-	9	V$ZID_01	-	1
V$NFKAPPAB50_01	-	3	V$AHRARNT_01	-	9	MA0001 AGL3 MADS	-	2
V$YY1_02	-	3	V$AREB6_01	-	9	MA0042 FOXI1 FORKHEAD	-	2
MA0101 REL REL	-	4	V$CHOP_01	-	8	MA0079 SP1 ZN-FINGER, C2H2	-	2
MA0120 ID1 ZN-FINGER, C2H2	-	4	V$ELK1_01	-	8	MA0086 Snail ZN-FINGER, C2H2	-	2
V$GATA1_03	-	4	V$HSF2_01	-	9	MA0089 TCF11-MafG bZIP	-	2
V$USF_02	-	4	V$NFKAPPAB50_01	-	9	MA0103 deltaEF1 ZN-FINGER, C2H2	-	2
MA0082 SQUA MADS	-	5	V$NFKB_Q6	-	8	MA0118 Macho-1 ZN-FINGER, C2H2	-	2
MA0105 NFKB1 REL	-	5	MA0003 TFAP2A AP2	-	10	V$CART1_01	-	2
V$CEBPA_01	-	5	MA0006 Arnt-Ahr bHLH	-	10	V$CDP_02	-	2
V$CREL_01	-	5	MA0120 ID1 ZN-FINGER, C2H2	-	10	V$CEBP_01	-	2
V$NFKB_Q6	-	5	V$E47_02	-	10	V$ER_Q6	-	2
V$OCT1_06	-	5	V$HSF1_01	-	10	V$FOXD3_01	-	2
MA0006 Arnt-Ahr bHLH	-	6	V$RREB1_01	-	10	V$FOXJ2_01	-	2

MA0010 Broad-complex_1 ZN-FINGER, C2H2	-	6	MA0037 GATA3 ZN-FINGER, GATA	-	11	V$GATA2_01	-	2
MA0022 Dorsal_1 REL	-	6	MA0073 RREB1 ZN-FINGER, C2H2	-	10	V$HFH3_01	-	2
MA0023 Dorsal_2 REL	-	6	MA0102 cEBP bZIP	-	11	V$MEF2_03	-	2
MA0070 Pbx HOMEO	-	6	MA0121 ARR10 TRP-CLUSTER	-	11	V$OCT1_04	-	2
MA0090 TEAD TEA	-	6	MA0101 REL REL	-	11	V$P53_01	-	2
V$CEBPB_01	-	6	MA0109 RUSH1-alfa ZN-FINGER, GATA	-	12	V$P53_02	-	2
V$HSF1_01	-	6	V$CEBP_01	-	12	V$PAX2_01	-	2
V$P53_01	-	6	V$CREL_01	-	11	V$SREBP1_01	-	2
V$P53_02	-	6	V$GATA1_02	-	11	V$YY1_01	-	2
MA0049 Hunchback ZN-FINGER, C2H2	-	7	V$PAX5_01	-	12	MA0008 Athb-1 HOMEO-ZIP	-	3
V$CEBP_C	-	7	V$TAXCREB_01	-	11	MA0027 En1 HOMEO	-	3
V$PAX5_01	-	7	MA0016 CFI-USP NUCLEAR RECEPTOR	-	13	MA0036 GATA2 ZN-FINGER, GATA	-	3
V$TST1_01	-	7	MA0057 ZNF42_5-13 ZN-FINGER, C2H2	-	13	V$CDP_01	-	3
MA0003 TFAP2A AP2	-	8	MA0068 Pax4 PAIRED-HOMEO	-	13	V$MEF2_04	-	3
V$AHRARNT_01	-	8	MA0098 c-ETS ETS	-	12	V$OCT1_06	-	3
V$CEBP_Q2	-	8	V$HNF4_01	-	13	V$AREB6_02	-	4
V$TAXCREB_01	-	8	V$LMO2COM_01	-	13	V$AREB6_03	-	4
MA0016 CFI-USP NUCLEAR RECEPTOR	-	9	V$YY1_02	-	13	V$FOXJ2_02	-	4
MA0073 RREB1 ZN-FINGER, C2H2	-	9	V$ZID_01	-	12	V$PBX1_01	-	4
MA0112 ESR1 NUCLEAR	-	9	MA0112 ESR1 NUCLEAR	-	14	V$RSRFC4_01	-	4
MA0114 HNF4 NUCLEAR	-	9	MA0122 Bapx1 HOMEO	-	14	V$SREBP1_02	-	4
V$AP2_Q6	-	9	V$AP2_Q6	-	14	MA0094 Ubx HOMEO	-	5
V$CDPCR3HD_01	-	9	V$CEBPB_02	-	14	MA0015 CF2-II ZN-	-	5

V$HNF4_01	-	9	V$PAX5_02	-	14	FINGER, C2H2 MA0033 FOXL1 FORKHEAD	-	5
V$OCT1_04	-	9	V$SREBP1_01	-	14	MA0052 MEF2A MADS	-	5
V$USF_C	-	9	MA0036 GATA2 ZN-FINGER, GATA	-	15	MA0063 Nkx2-5 HOMEO	-	5
MA0102 cEBP bZIP	-	10	MA0095 YY1 ZN-FINGER, C2H2	-	15	MA0075 Prrx2 HOMEO	-	6
V$CDPCR1_01	-	10	V$AREB6_03	-	15			
V$CHOP_01	-	10	V$GR_Q6	-	15			
V$LMO2COM_01	-	10	MA0035 Gata1 ZN-FINGER, GATA	-	16			
V$PAX5_02	-	10	MA0088 Staf ZN-FINGER, C2H2	-	16			
V$ZID_01	-	10	V$AREB6_02	-	16			
V$GR_Q6	-	11	V$PAX2_01	-	16			
MA0014 Pax5 PAIRED	-	12	V$SREBP1_02	-	16			
V$E47_02	-	12	MA0086 Snail ZN-FINGER, C2H2	-	17			
V$GATA1_02	-	12	MA0103 deltaEF1 ZN-FINGER, C2H2	-	17			
MA0121 ARR10 TRP-CLUSTER	-	13	V$AML1_01	-	17			
V$RREB1_01	-	13	V$GATA2_01	-	17			
MA0037 GATA3 ZN-FINGER, GATA	-	14	V$MZF1_02	-	17			
MA0068 Pax4 PAIRED-HOMEO	-	14	MA0111 Spz1 bHLH-ZIP	-	18			
V$AREB6_01	-	14	V$GATA3_01	-	18			
V$CEBP_01	-	14	V$MZF1_01	-	18			
V$CEBPB_02	-	14	V$SP1_Q6	-	18			
MA0122 Bapx1 HOMEO	-	15	V$YY1_01	-	18			
V$SP1_Q6	-	15	MA0027 En1 HOMEO	-	19			

MA0015 CF2-II ZN-FINGER, C2H2	-	16
MA0027 En1 HOMEO	-	16
MA0057 ZNF42_5-13 ZN-FINGER, C2H2	-	16
MA0109 RUSH1-alfa ZN-FINGER, GATA	-	16
V$PAX2_01	-	17
V$MZF1_01	-	18
V$SREBP1_02	-	18
MA0036 GATA2 ZN-FINGER, GATA	-	19
MA0056 ZNF42_1-4 ZN-FINGER, C2H2	-	19
MA0088 Staf ZN-FINGER, C2H2	-	19
MA0095 YY1 ZN-FINGER, C2H2	-	19
MA0111 Spz1 bHLH-ZIP	-	19
V$AML1_01	-	19
V$SREBP1_01	-	19
MA0035 Gata1 ZN-FINGER, GATA	-	20
MA0039 Klf4 ZN-FINGER, C2H2	-	20
MA0045 HMG-IY HMG	-	20
MA0118 Macho-1 ZN-FINGER, C2H2	-	20
V$AREB6_02	-	20
V$GATA2_01	-	20
V$MZF1_02	-	20
V$YY1_01	-	20
MA0086 Snail ZN-FINGER, C2H2	-	21
MA0103 deltaEF1 ZN-FINGER, C2H2	-	21
MA0056 ZNF42_1-4 ZN-FINGER, C2H2	-	19
MA0118 Macho-1 ZN-FINGER, C2H2	-	19
V$ER_Q6	-	19
MA0039 Klf4 ZN-FINGER, C2H2	-	21
MA0045 HMG-IY HMG	-	21
MA0079 SP1 ZN-FINGER, C2H2	-	21
V$ARP1_01	-	21
V$P300_01	-	21
V$SP1_01	-	22
MA0019 Chop-cEBP bZIP	-	23

V$AREB6_03	-	21
V$ER_Q6	-	21
V$GATA3_01	-	21
MA0079 SP1 ZN-FINGER, C2H2	-	22
V$P300_01	-	22
V$SP1_01	-	22
MA0019 Chop-cEBP bZIP	-	23
V$ARP1_01	-	23

A. All islet peaks not within 2 kb of a known transcipt (intergenic)

B. All intergenic islet peaks that did not overlap any peak from five additional cell lines

C. All intergenic islet peaks that overlapped peaks in all five additional cell lines

D. The number of chromosomes on which the motif was significantly over-represented ($p<.01$)

E. The number of chromosomes on which the motif was significantly under-represented ($p>.99$)

Table 5.6. Functional annotations of genes overlapping islet COREs*

PANTHER Biological Process	PValue	GO Biological Process	PValue
BP00044:mRNA transcription regulation	2.25E-24	GO:0051056~regulation of small GTPase mediated signal transduction	5.72E-08
BP00071:Proteolysis	1.45E-22	GO:0016192~vesicle-mediated transport	5.85E-07
BP00063:Protein modification	4.95E-21	GO:0006512~ubiquitin cycle	9.13E-07
BP00040:mRNA transcription	4.32E-20	GO:0009966~regulation of signal transduction	3.44E-06
BP00104:G-protein mediated signaling	7.57E-16	GO:0007264~small GTPase mediated signal transduction	7.57E-06
BP00143:Cation transport	8.83E-14	GO:0006366~transcription from RNA polymerase II promoter	9.15E-06
BP00064:Protein phosphorylation	1.49E-12	GO:0046903~secretion	1.05E-05
BP00142:Ion transport	1.96E-11	GO:0008104~protein localization	2.43E-05
BP00286:Cell structure	9.20E-11	GO:0032940~secretion by cell	5.91E-05
BP00289:Other metabolism	3.87E-10	GO:0006468~protein amino acid phosphorylation	5.95E-05
BP00102:Signal transduction	8.08E-10	GO:0046578~regulation of Ras protein signal transduction	5.99E-05
BP00060:Protein metabolism and modification	1.17E-08	GO:0007265~Ras protein signal transduction	7.37E-05

* Twelve most enriched PANTHER and GO biological processes

Table 5.7. Islet-selective CORES that extend > 2 kb from the transcription start or termination site of overlapping genes

Table available at http://www.nature.com/ng/journal/v42/n3/extref/ng.530-S5.

Table 5.8. Islet FAIRE enrichment at T2D susceptibility loci

		# FAIRE enriched regions [b]			# SNPs in dbSNP [c]				# T2D-associated SNPs [d]			
						# overlapping FAIRE				# overlapping FAIRE		
Locus	Reference SNP	liberal	moderate	stringent	total # in region	liberal	moderate	stringent	total # in region	liberal	moderate	stringent
TCF7L2	rs7903146[e]	10	7	2	106	3	2	1	4	1	1	-
CDKAL1	rs4712523[e]	38	14	5	242	44	8	2	18	5	-	-
CDKN2A/ CDKN2B	rs2383208[e]	25	10	2	324	23	8	-	3	2	1	-
IGF2BP2	rs4402960[e]	3	3	2	82	4	2	2	29	1	1	1
JAZF1	rs864745[e]	22	10	6	157	11	4	2	6	-	-	-
CDC123/ CAMK1D	rs12779790[e]	8	3	2	135	9	5	2	4	1	1	-
TSPAN8/ LGR5	rs7961581[e]	34	11	5	623	22	8	5	7	-	-	-
THADA	rs7578597[e]	100	57	26	638	85	50	24	109	23	12	6
ADAMTS 9	rs4607103[e]	1	-	-	76	-	-	-	8	-	-	-
NOTCH2/ ADAM30	rs10923931[e]	5	1	1	223	2	-	-	41	-	-	-
FTO	rs8050136[e]	15	7	3	114	6	4	1	37	2	-	-
SLC30A8	rs13266634[e]	80	55	38	377	122	63	21	4	1	1	-
HNF1B	rs7501939[f]	18	9	2	284	11	7	2	3	1	1	1
WFS1	rs10010131[g]	3	2	-	115	-	-	-	45	-	-	-
MTNR1B	rs10830963[h]	3	2	1	96	2	1	1	1	-	-	-
HHEX/ID E	rs1111875[e]	22	6	4	266	13	5	4	4	-	-	-
KCNQ1	rs2237892[i]	4	3	2	285	2	-	-	3	-	-	-
PPARG/S YN2	rs1801282[e]	12	1	0	557	1	-	-	13	-	-	-
KCNJ11	rs5215[e]	7	5	3	88	7	5	3	7	-	-	-
G6PC2	rs560887[h]	28	21	11	177	52	30	20	3	1	1	1

A. Coordinates defined by identifying recombination hotspots flanking the reference SNP

B. Number of FAIRE-enriched sites (sample 3 only) at liberal, moderate and stringent threshold located within locus coordinates

C. Number of SNPs in dbSNP v129 with a reported average heterozygosity > 1% located within locus coordinates

D. Number of SNPs in HapMap CEU r2 > .8 with reference SNP

E. Mohlke, *et al*. Hum Mol Genet. 2008 Oct 15;17(R2):R102-8.

F. Gudmundsson, *et al*. Nat Genet. 2007 Aug;39(8):977-83.

G. Sandhu, *et al.* Nat Genet. 2007 Aug;39(8):951-3.

H. Prokopenko, et al. Nat Genet. 2007 Jan;41(1):77-81.

I. Yasuda, et al. Nat Genet. 2008 Sep;40(9):1092-7.

Table 6.1. Most over- and under-represented TFBS motifs in islet training set

Enriched in positive set					Enriched in negative set				
TFBS id	Database	Positive %	Negative %	Motif score	TFBS id	Database	Positive %	Negative %	Motif score
V$BRACH_01	TRANSFAC	0.1	0.0	3.7	V$NRSF_01	TRANSFAC	0.0	0.5	-4.7
F$ABF1_01	TRANSFAC	5.5	0.8	0.8	V$HEN1_02	TRANSFAC	0.6	5.7	-1.0
V$MEF2_01	TRANSFAC	7.2	1.1	0.8	V$R_01	TRANSFAC	0.6	5.1	-1.0
V$CART1_01	TRANSFAC	36.1	5.7	0.8	MA0138_REST	JASPAR	0.3	3.1	-1.0
V$CLOX_01	TRANSFAC	34.9	5.7	0.8	V$AHRARNT_02	TRANSFAC	0.9	7.3	-0.9
OCT1	JASPAR	22.4	3.8	0.8	MA0105_NFKB1	JASPAR	6.4	44.4	-0.8
V$HNF1_C	TRANSFAC	36.4	6.7	0.7	MA0112_ESR1	JASPAR	1.5	10.4	-0.8
P$ATHB1_01	TRANSFAC	30.3	6.2	0.7	I$HAIRY_01	TRANSFAC	7.3	47.7	-0.8
V$PAX6_01	TRANSFAC	6.2	1.3	0.7	V$SP1_Q6	TRANSFAC	13.2	79.1	-0.8
V$E4BP4_01	TRANSFAC	43.3	9.6	0.7	I$ADF1_Q6	TRANSFAC	4.1	22.9	-0.8
V$NKX61_01	TRANSFAC	58.8	13.6	0.6	MA0016_USP	JASPAR	12.7	66.7	-0.7
MA0135_LHX3	JASPAR	67.9	16.0	0.6	V$ER_Q6	TRANSFAC	3.9	20.1	-0.7
MA0025_NFIL3	JASPAR	59.6	14.2	0.6	F$GAL4_C	TRANSFAC	0.1	0.5	-0.7
P$O2_01	TRANSFAC	0.4	0.1	0.6	P$BZIP910_02	TRANSFAC	3.5	16.4	-0.7
V$PBX1_02	TRANSFAC	17.9	4.4	0.6	V$SREBP1_02	TRANSFAC	7.6	35.5	-0.7
MA0046_HNF1A	JASPAR	46.0	11.8	0.6	F$LAC9_C	TRANSFAC	0.4	2.0	-0.7
V$OCT1_07	TRANSFAC	64.5	16.6	0.6	F$PACC_01	TRANSFAC	12.5	54.1	-0.6
V$FOXJ2_02	TRANSFAC	63.8	16.7	0.6	V$ARP1_01	TRANSFAC	7.6	32.5	-0.6
P$MYBPH3_02	TRANSFAC	27.8	7.6	0.6	V$EGR3_01	TRANSFAC	3.9	16.0	-0.6
V$FREAC2_01	TRANSFAC	39.7	11.4	0.5	V$NGFIC_01	TRANSFAC	6.4	26.4	-0.6

Table 6.2. T2D-associated SNPs with significant allelic differences in TFBS classification ($p < .0001$)

		Allele 1		Allele 2				islet FAIRE sites [a]		
SNP	locus	a1	a1 score	a2	a2 score	score difference	*P*	1	2	3
rs4686697	IGF2BP2	C	20.8	T	30.3	-9.5	3.50E-08			
rs10965246	CDKN2A	C	15.2	T	23.8	-8.6	5.90E-07			
rs4747971	CDC123	C	-2.8	T	5.5	-8.2	1.90E-06			
rs16884074	CDKAL1	C	38.6	T	46.2	-7.6	1.00E-05			l
rs7903146	TCF7L2	C	14.6	T	22	-7.5	1.30E-05		m	m
rs3847554	MTNR1B	C	5.7	T	13	-7.4	1.70E-05			
rs6445424	ADAMTS9	A	27.2	C	19.9	7.3	2.20E-05			
rs6456370	CDKAL1	A	24.6	G	17.4	7.2	2.90E-05			
rs13405776	THADA	C	7	T	13.9	-6.9	6.10E-05			
rs6747229	THADA	A	26.6	T	33.3	-6.8	7.70E-05		l	

a. SNP overlapping islet FAIRE sites from sample 1 (IF1), 2 (IF2), or 3 (IF3) at liberal (l), moderate (m) or stringent (s) threshold

1. Lander, E.S. & Schork, N.J. Genetic dissection of complex traits. *Science* **265**, 2037-48 (1994).

2. Chico, T.J., Milo, M. & Crossman, D.C. The genetics of cardiovascular disease: new insights from emerging approaches. *J Pathol* **220**, 186-97 (2010).

3. O'Rahilly, S. Human genetics illuminates the paths to metabolic disease. *Nature* **462**, 307-14 (2009).

4. Baranzini, S.E. The genetics of autoimmune diseases: a networked perspective. *Curr Opin Immunol* **21**, 596-605 (2009).

5. Bertram, L. & Tanzi, R.E. Genome-wide association studies in Alzheimer's disease. *Hum Mol Genet* **18**, R137-45 (2009).

6. O'Donovan, M.C., Craddock, N.J. & Owen, M.J. Genetics of psychosis; insights from views across the genome. *Hum Genet* **126**, 3-12 (2009).

7. Gubitz, A.K. & Gwinn, K. Mining the genome for susceptibility to complex neurological disorders. *Curr Mol Med* **9**, 801-13 (2009).

8. Shianna, K.V. & Willard, H.F. Human genomics: in search of normality. *Nature* **444**, 428-9 (2006).

9. http://www.ncbi.nlm.nih.gov/projects/SNP/

10. Shastry, B.S. SNPs: impact on gene function and phenotype. *Methods Mol Biol* **578**, 3-22 (2009).

11. Hirschhorn, J.N. & Daly, M.J. Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* **6**, 95-108 (2005).

12. Hirschhorn, J.N. Genetic approaches to studying common diseases and complex traits. *Pediatr Res* **57**, 74R-77R (2005).

13. Perez-Iratxeta, C., Bork, P. & Andrade, M.A. Association of genes to genetically inherited diseases using data mining. *Nat Genet* **31**, 316-9 (2002).

14. Aerts, S. et al. Gene prioritization through genomic data fusion. *Nat Biotechnol* **24**, 537-44 (2006).

15. Turner, F.S., Clutterbuck, D.R. & Semple, C.A. POCUS: mining genomic sequence annotation to predict disease genes. *Genome Biol* **4**, R75 (2003).

16. Adie, E.A., Adams, R.R., Evans, K.L., Porteous, D.J. & Pickard, B.S. SUSPECTS: enabling fast and effective prioritization of positional candidates. *Bioinformatics* **22**, 773-4 (2006).

17. van Driel, M.A., Cuelenaere, K., Kemmeren, P.P., Leunissen, J.A. & Brunner, H.G. A new web-based data mining tool for the identification of candidate genes for human genetic disorders. *Eur J Hum Genet* **11**, 57-63 (2003).

18. Freudenberg, J. & Propping, P. A similarity-based method for genome-wide prediction of disease-relevant human genes. *Bioinformatics* **18 Suppl 2**, S110-5 (2002).

19. Altshuler, D., Daly, M.J. & Lander, E.S. Genetic mapping in human disease. *Science* **322**, 881-8 (2008).

20. Peltonen, L. & McKusick, V.A. Genomics and medicine. Dissecting human disease in the postgenomic era. *Science* **291**, 1224-9 (2001).

21. International HapMap, C. A haplotype map of the human genome. *Nature* **437**, 1299-1320 (2005).

22. Sachidanandam, R. et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928-33 (2001).

23. Shen, R. et al. High-throughput SNP genotyping on universal bead arrays. *Mutat Res* **573**, 70-82 (2005).

24. Carlton, V.E., Ireland, J.S., Useche, F. & Faham, M. Functional single nucleotide polymorphism-based association studies. *Hum Genomics* **2**, 391-402 (2006).

25. Ding, C. & Jin, S. High-throughput methods for SNP genotyping. *Methods Mol Biol* **578**, 245-54 (2009).

26. Craig, D.W. & Stephan, D.A. Applications of whole-genome high-density SNP genotyping. *Expert Rev Mol Diagn* **5**, 159-70 (2005).

27. Li, Y., Willer, C., Sanna, S. & Abecasis, G. Genotype imputation. *Annu Rev Genomics Hum Genet* **10**, 387-406 (2009).

28. Manolio, T.A. et al. Finding the missing heritability of complex diseases. *Nature* **461**, 747-53 (2009).

29. Romeo, S. et al. Rare loss-of-function mutations in ANGPTL family members contribute to plasma triglyceride levels in humans. *J Clin Invest* **119**, 70-9 (2009).

30. Fahmi, S., Yang, C., Esmail, S., Hobbs, H.H. & Cohen, J.C. Functional characterization of genetic variants in NPC1L1 supports the sequencing extremes strategy to identify complex trait genes. *Hum Mol Genet* **17**, 2101-7 (2008).

31. Romeo, S. et al. Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL. *Nat Genet* **39**, 513-6 (2007).

32. Kotowski, I.K. et al. A spectrum of PCSK9 alleles contributes to plasma levels of low-density lipoprotein cholesterol. *Am J Hum Genet* **78**, 410-22 (2006).

33. Cohen, J.C. et al. Multiple rare variants in NPC1L1 associated with reduced sterol absorption and plasma low-density lipoprotein levels. *Proc Natl Acad Sci U S A* **103**, 1810-5 (2006).

34. Ng, S.B. et al. Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet* **42**, 30-5.

35. Romao, I. & Roth, J. Genetic and environmental interactions in obesity and type 2 diabetes. *J Am Diet Assoc* **108**, S24-8 (2008).

36. Stumvoll, M. & Gerich, J. Clinical features of insulin resistance and beta cell dysfunction and the relationship to type 2 diabetes. *Clin Lab Med* **21**, 31-51 (2001).

37. O'Rahilly, S., Barroso, I. & Wareham, N.J. Genetic factors in type 2 diabetes: the end of the beginning? *Science (New York, N.Y.)* **307**, 370-373 (2005).

38. Kaprio, J. et al. Concordance for type 1 (insulin-dependent) and type 2 (non-insulin-dependent) diabetes mellitus in a population-based cohort of twins in Finland. *Diabetologia* **35**, 1060-7 (1992).

39. Hanson, R.L. et al. An autosomal genomic scan for loci linked to type II diabetes mellitus and body-mass index in Pima Indians. *Am J Hum Genet* **63**, 1130-8 (1998).

40. Reynisdottir, I. et al. Localization of a susceptibility gene for type 2 diabetes to chromosome 5q34-q35.2. *Am J Hum Genet* **73**, 323-35 (2003).

41. Silander, K. et al. A large set of Finnish affected sibling pair families with type 2 diabetes suggests susceptibility loci on chromosomes 6, 11, and 14. *Diabetes* **53**, 821-9 (2004).

42. Vionnet, N. et al. Genomewide search for type 2 diabetes-susceptibility genes in French whites: evidence for a novel susceptibility locus for early-onset diabetes on chromosome 3q27-qter and independent replication of a type 2-diabetes locus on chromosome 1q21-q24. *Am J Hum Genet* **67**, 1470-80 (2000).

43. Wiltshire, S. et al. A genomewide scan for loci predisposing to type 2 diabetes in a U.K. population (the Diabetes UK Warren 2 Repository): analysis of 573 pedigrees provides independent replication of a susceptibility locus on chromosome 1q. *Am J Hum Genet* **69**, 553-69 (2001).

44. Ehm, M.G. et al. Genomewide search for type 2 diabetes susceptibility genes in four American populations. *Am J Hum Genet* **66**, 1871-81 (2000).

45. Guan, W., Pluzhnikov, A., Cox, N.J. & Boehnke, M. Meta-analysis of 23 type 2 diabetes linkage studies from the International Type 2 Diabetes Linkage Analysis Consortium. *Hum Hered* **66**, 35-49 (2008).

46. Prokopenko, I. et al. Linkage disequilibrium mapping of the replicated type 2 diabetes linkage signal on chromosome 1q. *Diabetes* **58**, 1704-9 (2009).

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

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

49. Gloyn, A.L. et al. Large-scale association studies of variants in genes encoding the pancreatic beta-cell KATP channel subunits Kir6.2 (KCNJ11) and SUR1 (ABCC8) confirm that the KCNJ11 E23K variant is associated with type 2 diabetes. *Diabetes* **52**, 568-572 (2003).

50. Sandhu, M.S. et al. Common variants in WFS1 confer risk of type 2 diabetes. *Nature genetics* **39**, 951-953 (2007).

51. Gudmundsson, J. et al. Two variants on chromosome 17 confer prostate cancer risk, and the one in TCF2 protects against type 2 diabetes. *Nature genetics* **39**, 977-983 (2007).

52. Winckler, W. et al. Evaluation of common variants in the six known maturity-onset diabetes of the young (MODY) genes for association with type 2 diabetes. *Diabetes* **56**, 685-93 (2007).

53. Sladek, R. et al. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445**, 881-885 (2007).

54. Diabetes Genetics Initiative of Broad Institute of, H. et al. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science (New York, N.Y.)* **316**, 1331-1336 (2007).

55. Scott, L.J. et al. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science (New York, N.Y.)* **316**, 1341-1345 (2007).

56. Zeggini, E. et al. Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science (New York, N.Y.)* **316**, 1336-1341 (2007).

57. Florez, J.C. et al. A 100k Genome-Wide Association Scan for Diabetes and Related Traits in the Framingham Heart Study: Replication and Integration with Other Genome-Wide Datasets. *Diabetes* (2007).

58. Hanson, R.L. et al. A Search for Variants Associated with Young-Onset Type 2 Diabetes in American Indians in a 100k Genotyping Array. *Diabetes* (2007).

59. Hayes, M.G. et al. Identification of Type 2 Diabetes Genes in Mexican Americans Through Genome-wide Association Studies. *Diabetes* (2007).

60. Rampersaud, E. et al. Identification of novel candidate genes for type 2 diabetes from a genome-wide association scan in the Old Order Amish: Evidence for replication from diabetes-related quantitative traits and from independent populations. *Diabetes* (2007).

61. Salonen, J.T. et al. Type 2 Diabetes Whole-Genome Association Study in Four Populations: The DiaGen Consortium. *American Journal of Human Genetics* **81**, 338-345 (2007).

62. Steinthorsdottir, V. et al. A variant in CDKAL1 influences insulin response and risk of type 2 diabetes. *Nature genetics* (2007).

63. Bouatia-Naji, N. et al. A variant near MTNR1B is associated with increased fasting plasma glucose levels and type 2 diabetes risk. *Nat Genet* **41**, 89-94 (2009).

64. Lyssenko, V. et al. Common variant in MTNR1B associated with increased risk of type 2 diabetes and impaired early insulin secretion. *Nat Genet* **41**, 82-8 (2009).

65. Rung, J. et al. Genetic variant near IRS1 is associated with type 2 diabetes, insulin resistance and hyperinsulinemia. *Nat Genet* **41**, 1110-5 (2009).

66. Kong, A. et al. Parental origin of sequence variants associated with complex diseases. *Nature* **462**, 868-74 (2009).

67. Yasuda, K. et al. Variants in KCNQ1 are associated with susceptibility to type 2 diabetes mellitus. *Nat Genet* **40**, 1092-7 (2008).

68. Unoki, H. et al. SNPs in KCNQ1 are associated with susceptibility to type 2 diabetes in East Asian and European populations. *Nat Genet* **40**, 1098-102 (2008).

69. Zeggini, E. et al. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat Genet* **40**, 638-45 (2008).

70. McCarthy, M.I. & Zeggini, E. Genome-wide association studies in type 2 diabetes. *Curr Diab Rep* **9**, 164-71 (2009).

71. Mohlke, K.L., Boehnke, M. & Abecasis, G.R. Metabolic and cardiovascular traits: an abundance of recently identified common genetic variants. *Hum Mol Genet* **17**, R102-8 (2008).

72. Stolerman, E.S. & Florez, J.C. Genomics of type 2 diabetes mellitus: implications for the clinician. *Nat Rev Endocrinol* **5**, 429-36 (2009).

73. Dupuis, J. et al. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nat Genet* **42**, 105-16 (2010).

74. Frayling, T.M. et al. A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* **316**, 889-94 (2007).

75. Valle, T. et al. Mapping genes for NIDDM. Design of the Finland-United States Investigation of NIDDM Genetics (FUSION) Study. *Diabetes care* **21**, 949-958 (1998).

76. Ghosh, S. et al. The Finland-United States investigation of non-insulin-dependent diabetes mellitus genetics (FUSION) study. I. An autosomal genome scan for genes that predispose to type 2 diabetes. *Am J Hum Genet* **67**, 1174-85 (2000).

77. Watanabe, R.M. et al. The Finland-United States investigation of non-insulin-dependent diabetes mellitus genetics (FUSION) study. II. An autosomal genome scan for diabetes-related quantitative-trait loci. *Am J Hum Genet* **67**, 1186-200 (2000).

78. Silander, K. et al. Genetic variation near the hepatocyte nuclear factor-4 alpha gene predicts susceptibility to type 2 diabetes. *Diabetes* **53**, 1141-1149 (2004).

79. Willer, C.J. et al. Screening of 134 single nucleotide polymorphisms (SNPs) previously associated with type 2 diabetes replicates association with 12 SNPs in nine genes. *Diabetes* **56**, 256-264 (2007).

80. Bonnycastle, L.L. et al. Common variants in maturity-onset diabetes of the young genes contribute to risk of type 2 diabetes in Finns. *Diabetes* **55**, 2534-2540 (2006).

81. Scott, L.J. et al. Association of transcription factor 7-like 2 (TCF7L2) variants with type 2 diabetes in a Finnish sample. *Diabetes* **55**, 2649-2653 (2006).

82. Willer, C.J. et al. Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nat Genet* **41**, 25-34 (2009).

83. Newton-Cheh, C. et al. Genome-wide association study identifies eight loci associated with blood pressure. *Nat Genet* (2009).

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

85. Kathiresan, S. et al. Common variants at 30 loci contribute to polygenic dyslipidemia. *Nat Genet* **41**, 56-65 (2009).

86. Willer, C.J. et al. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nat Genet* **40**, 161-9 (2008).

87. Lettre, G. et al. Identification of ten loci associated with height highlights new biological pathways in human growth. *Nat Genet* **40**, 584-91 (2008).

88. Montgomery, S.B. & Dermitzakis, E.T. The resolution of the genetics of gene expression. *Hum Mol Genet* **18**, R211-5 (2009).

89. Birney, E. et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799-816 (2007).

90. Wang, Z., Schones, D.E. & Zhao, K. Characterization of human epigenomes. *Curr Opin Genet Dev* **19**, 127-34 (2009).

91. Wallrath, L.L., Lu, Q., Granok, H. & Elgin, S.C. Architectural variations of inducible eukaryotic promoters: preset and remodeling chromatin structures. *Bioessays* **16**, 165-70 (1994).

92. Park, P.J. ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet* **10**, 669-80 (2009).

93. Barski, A. & Zhao, K. Genomic location analysis by ChIP-Seq. *J Cell Biochem* **107**, 11-8 (2009).

94. Shibata, Y. & Crawford, G.E. Mapping regulatory elements by DNaseI hypersensitivity chip (DNase-Chip). *Methods Mol Biol* **556**, 177-90 (2009).

95. Giresi, P.G. & Lieb, J.D. Isolation of active regulatory elements from eukaryotic chromatin using FAIRE (Formaldehyde Assisted Isolation of Regulatory Elements). *Methods* **48**, 233-9 (2009).

96. Boyle, A.P. et al. High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**, 311-22 (2008).

97. Wasserman, W.W. & Sandelin, A. Applied bioinformatics for the identification of regulatory elements. *Nat Rev Genet* **5**, 276-87 (2004).

98. Portales-Casamar, E. et al. JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res* **38**, D105-10.

99. Matys, V. et al. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res* **34**, D108-10 (2006).

100. Newburger, D.E. & Bulyk, M.L. UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. *Nucleic Acids Res* **37**, D77-82 (2009).

101. Hannenhalli, S. Eukaryotic transcription factor binding sites--modeling and integrative search methods. *Bioinformatics* **24**, 1325-31 (2008).

102. Xie, X., Rigor, P. & Baldi, P. MotifMap: a human genome-wide map of candidate regulatory motif sites. *Bioinformatics* **25**, 167-74 (2009).

103. Prabhakar, S. et al. Close sequence comparisons are sufficient to identify human cis-regulatory elements. *Genome Res* **16**, 855-63 (2006).

104. Wang, Q.F. et al. Detection of weakly conserved ancestral mammalian regulatory sequences by primate comparisons. *Genome Biol* **8**, R1 (2007).

105. Venkatesh, B. & Yap, W.H. Comparative genomics using fugu: a tool for the identification of conserved vertebrate cis-regulatory elements. *Bioessays* **27**, 100-7 (2005).

106. Loots, G.G. Genomic identification of regulatory elements by evolutionary sequence comparison and functional analysis. *Adv Genet* **61**, 269-93 (2008).

107. Van Loo, P. & Marynen, P. Computational methods for the detection of cis-regulatory modules. *Brief Bioinform* **10**, 509-24 (2009).

108. Hallikas, O. et al. Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. *Cell* **124**, 47-59 (2006).

109. Pennacchio, L.A., Loots, G.G., Nobrega, M.A. & Ovcharenko, I. Predicting tissue-specific enhancers in the human genome. *Genome Res* **17**, 201-11 (2007).

110. Blanchette, M. et al. Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. *Genome Res* **16**, 656-68 (2006).

111. Narlikar, L. et al. Genome-wide discovery of human heart enhancers. *Genome Res* **20**, 381-92 (2010).

112. Newton-Cheh, C. & Hirschhorn, J.N. Genetic association studies of complex traits: design and analysis issues. *Mutat Res* **573**, 54-69 (2005).

113. Dean, M. Approaches to identify genes for complex human diseases: lessons from Mendelian disorders. *Hum Mutat* **22**, 261-74 (2003).

114. Thomas, D.C. Are we ready for genome-wide association studies? *Cancer Epidemiol Biomarkers Prev* **15**, 595-8 (2006).

115. Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A. & McKusick, V.A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* **33**, D514-7 (2005).

116. Maglott, D., Ostell, J., Pruitt, K.D. & Tatusova, T. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res* **33**, D54-8 (2005).

117. Birney, E. et al. Ensembl 2006. *Nucleic Acids Res* **34**, D556-61 (2006).

118. Harris, M.A. et al. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* **32**, D258-61 (2004).

119. Smith, C.L., Goldsmith, C.A. & Eppig, J.T. The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biol* **6**, R7 (2005).

120. Kelso, J. et al. eVOC: a controlled vocabulary for unifying gene expression data. *Genome Res* **13**, 1222-30 (2003).

121. Salton, D., Wong, A. & Yang, C.S. A vector space model for automatic indexing. *Communications of the ACM* **18**, 613-620 (1975).

122. Blake, J.A., Richardson, J.E., Bult, C.J., Kadin, J.A. & Eppig, J.T. MGD: the Mouse Genome Database. *Nucleic Acids Res* **31**, 193-5 (2003).

123. Becker, K.G., Barnes, K.C., Bright, T.J. & Wang, S.A. The genetic association database. *Nat Genet* **36**, 431-2 (2004).

124. Bairoch, A. et al. The Universal Protein Resource (UniProt). *Nucleic Acids Res* **33**, D154-9 (2005).

125. Camon, E. et al. The Gene Ontology Annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro. *Genome Res* **13**, 662-72 (2003).

126. Alfarano, C. et al. The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Res* **33**, D418-24 (2005).

127. Peri, S. et al. Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Res* **32**, D497-501 (2004).

128. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. & Hattori, M. The KEGG resource for deciphering the genome. *Nucleic Acids Res* **32**, D277-80 (2004).

129. Apweiler, R. et al. InterPro--an integrated documentation resource for protein families, domains and functional sites. *Bioinformatics* **16**, 1145-50 (2000).

130. Ueda, H. et al. Association of the T-cell regulatory gene CTLA4 with susceptibility to autoimmune disease. *Nature* **423**, 506-11 (2003).

131. Bottini, N. et al. A functional variant of lymphoid tyrosine phosphatase is associated with type I diabetes. *Nat Genet* **36**, 337-8 (2004).

132. Begovich, A.B. et al. A missense single-nucleotide polymorphism in a gene encoding a protein tyrosine phosphatase (PTPN22) is associated with rheumatoid arthritis. *Am J Hum Genet* **75**, 330-7 (2004).

133. Guo, D. et al. A functional variant of SUMO4, a new I kappa B alpha modifier, is associated with type 1 diabetes. *Nat Genet* **36**, 837-41 (2004).

134. Kochi, Y. et al. A functional variant in FCRL3, encoding Fc receptor-like 3, is associated with rheumatoid arthritis and several autoimmunities. *Nat Genet* **37**, 478-85 (2005).

135. Pimm, J. et al. The Epsin 4 gene on chromosome 5q, which encodes the clathrin-associated protein enthoprotin, is involved in the genetic susceptibility to schizophrenia. *Am J Hum Genet* **76**, 902-7 (2005).

136. Gharani, N., Benayed, R., Mancuso, V., Brzustowicz, L.M. & Millonig, J.H. Association of the homeobox transcription factor, ENGRAILED 2, 3, with autism spectrum disorder. *Mol Psychiatry* **9**, 474-84 (2004).

137. Klein, R.J. et al. Complement factor H polymorphism in age-related macular degeneration. *Science* **308**, 385-9 (2005).

138. Rivera, A. et al. Hypothetical LOC387715 is a second major susceptibility gene for age-related macular degeneration, contributing independently of complement factor H to disease risk. *Hum Mol Genet* **14**, 3227-36 (2005).

139. Helgadottir, A. et al. A variant of the gene encoding leukotriene A4 hydrolase confers ethnicity-specific risk of myocardial infarction. *Nat Genet* **38**, 68-74 (2006).

140. Gold, B. et al. Variation in factor B (BF) and complement component 2 (C2) genes is associated with age-related macular degeneration. *Nat Genet* **38**, 458-62 (2006).

141. Laitinen, T. et al. Characterization of a common susceptibility locus for asthma-related traits. *Science* **304**, 300-4 (2004).

142. Monsuur, A.J. et al. Myosin IXB variant increases the risk of celiac disease and points toward a primary intestinal barrier defect. *Nat Genet* **37**, 1341-4 (2005).

143. Vella, A. et al. Localization of a type 1 diabetes locus in the IL2RA/CD25 region by use of tag single-nucleotide polymorphisms. *Am J Hum Genet* **76**, 773-9 (2005).

144. Maraganore, D.M. et al. High-resolution whole-genome association study of Parkinson disease. *Am J Hum Genet* **77**, 685-93 (2005).

145. Grupe, A. et al. A scan of chromosome 10 identifies a novel locus showing strong association with late-onset Alzheimer disease. *Am J Hum Genet* **78**, 78-88 (2006).

146. Wheeler, D.L. et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **34**, D173-80 (2006).

147. Hirschman, L., Colosimo, M., Morgan, A. & Yeh, A. Overview of BioCreAtIvE task 1B: normalized gene lists. *BMC Bioinformatics* **6 Suppl 1**, S11 (2005).

148. http://www.diabetes.niddk.nih.gov.libproxy.lib.unc.edu/dm/pubs/statistics.

149. Frayling, T.M. Genome-wide association studies provide new insights into type 2 diabetes aetiology. *Nature reviews.Genetics* **8**, 657-662 (2007).

150. Barroso, I. et al. Candidate gene association study in type 2 diabetes indicates a role for genes involved in beta-cell function as well as insulin action. *PLoS biology* **1**, E20 (2003).

151. Freeman, H. & Cox, R.D. Type-2 diabetes: a cocktail of genetic discovery. *Human molecular genetics* **15 Spec No 2**, R202-9 (2006).

152. Altshuler, D. et al. The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nature genetics* **26**, 76-80 (2000).

153. Grant, S.F. et al. Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. *Nature genetics* **38**, 320-323 (2006).

154. Gaulton, K.J., Mohlke, K.L. & Vision, T.J. A computational system to select candidate genes for complex human traits. *Bioinformatics (Oxford, England)* **23**, 1132-1140 (2007).

155. Silander, K. et al. A large set of Finnish affected sibling pair families with type 2 diabetes suggests susceptibility loci on chromosomes 6, 11, and 14. *Diabetes* **53**, 821-829 (2004).

156. Saaristo, T. et al. Cross-sectional evaluation of the Finnish Diabetes Risk Score: a tool to identify undetected type 2 diabetes, abnormal glucose tolerance and metabolic syndrome. *Diabetes & vascular disease research : official journal of the International Society of Diabetes and Vascular Disease* **2**, 67-72 (2005).

157. Lazar, M.A. How obesity causes diabetes: not a tall tale. *Science (New York, N.Y.)* **307**, 373-375 (2005).

158. Lowell, B.B. & Shulman, G.I. Mitochondrial dysfunction and type 2 diabetes. *Science (New York, N.Y.)* **307**, 384-387 (2005).

159. Rhodes, C.J. Type 2 diabetes-a matter of beta-cell life and death? *Science (New York, N.Y.)* **307**, 380-384 (2005).

160. Schwartz, M.W. & Porte, D., Jr. Diabetes, obesity, and the brain. *Science (New York, N.Y.)* **307**, 375-379 (2005).

161. Harris, M.A. et al. The Gene Ontology (GO) database and informatics resource. *Nucleic acids research* **32**, D258-61 (2004).

162. Kelso, J. et al. eVOC: a controlled vocabulary for unifying gene expression data. *Genome research* **13**, 1222-1230 (2003).

163. Smith, C.L., Goldsmith, C.A. & Eppig, J.T. The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome biology* **6**, R7 (2005).

164. Camon, E. et al. The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic acids research* **32**, D262-6 (2004).

165. Mootha, V.K. et al. Erralpha and Gabpa/b specify PGC-1alpha-dependent oxidative phosphorylation gene expression that is altered in diabetic muscle. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 6570-6575 (2004).

166. Mootha, V.K. et al. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature genetics* **34**, 267-273 (2003).

167. Patti, M.E. et al. Coordinated reduction of genes of oxidative metabolism in humans with insulin resistance and diabetes: Potential role of PGC1 and NRF1. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 8466-8471 (2003).

168. Guan, W, et al. Meta-analysis of 23 type 2 diabetes linkage studies from the International Type 2 Diabetes Linkage Analysis Consortium. *Human Hered* **66**, 35-49 (2007).

169. Morley, M. et al. Genetic analysis of genome-wide variation in human gene expression. *Nature* **430**, 743-747 (2004).

170. Fingerlin, T.E. et al. Variation in three single nucleotide polymorphisms in the calpain-10 gene not associated with type 2 diabetes in a large Finnish cohort. *Diabetes* **51**, 1644-1648 (2002).

171. Horikawa, Y. et al. Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus. *Nature genetics* **26**, 163-175 (2000).

172. Weedon, M.N. et al. Meta-analysis and a large association study confirm a role for calpain-10 variation in type 2 diabetes susceptibility. *American Journal of Human Genetics* **73**, 1208-1212 (2003).

173. Willer, C.J. et al. Tag SNP selection for Finnish individuals based on the CEPH Utah HapMap database. *Genetic epidemiology* **30**, 180-190 (2006).

174. Blanchette, M. et al. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome research* **14**, 708-715 (2004).

176. Conneely, K.N. & Boehnke, M. So many correlated tests, so little time! Rapid adjustment of p-values for multiple correlated tests. *Am.J.Hum.Genet.* (2007).

177. Skol, A.D., Scott, L.J., Abecasis, G.R. & Boehnke, M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat Genet* **38**, 209-13 (2006).

178. Knoblauch, H. et al. Common haplotypes in five genes influence genetic variance of LDL and HDL cholesterol in the general population. *Human molecular genetics* **11**, 1477-1485 (2002).

179. Tai, E.S. et al. Association between the PPARA L162V polymorphism and plasma lipid levels: the Framingham Offspring Study. *Arteriosclerosis, Thrombosis, and Vascular Biology* **22**, 805-810 (2002).

180. Chiang, S.H., Chang, L. & Saltiel, A.R. TC10 and insulin-stimulated glucose transport. *Methods in enzymology* **406**, 701-714 (2006).

181. Guerrero, C., Martin-Encabo, S., Fernandez-Medarde, A. & Santos, E. C3G-mediated suppression of oncogene-induced focus formation in fibroblasts involves inhibition of ERK activation, cyclin A expression and alterations of anchorage-independent growth. *Oncogene* **23**, 4885-93 (2004).

182. Dumont, P., Leu, J.I., Della Pietra, A.C., 3rd, George, D.L. & Murphy, M. The codon 72 polymorphic variants of p53 have markedly different apoptotic potential. *Nature genetics* **33**, 357-365 (2003).

183. Meyre, D. et al. Variants of ENPP1 are associated with childhood and adult obesity and increase the risk of glucose intolerance and type 2 diabetes. *Nature genetics* **37**, 863-867 (2005).

184. Bondarenko, V.A., Liu, Y.V., Jiang, Y.I. & Studitsky, V.M. Communication over a large distance: enhancers and insulators. *Biochemistry and cell biology = Biochimie et biologie cellulaire* **81**, 241-251 (2003).

185. Cassar, A., Holmes, D.R., Jr., Rihal, C.S. & Gersh, B.J. Chronic coronary artery disease: diagnosis and management. *Mayo Clin Proc* **84**, 1130-46 (2009).

186. Iizuka, K. & Horikawa, Y. ChREBP: a glucose-activated transcription factor involved in the development of metabolic syndrome. *Endocr J* **55**, 617-24 (2008).

187. Lee, E.C. et al. Identification of a new functional domain in angiopoietin-like 3 (ANGPTL3) and angiopoietin-like 4 (ANGPTL4) involved in binding and inhibition of lipoprotein lipase (LPL). *J Biol Chem* **284**, 13735-45 (2009).

188. Sung, H.Y. et al. Human tribbles-1 controls proliferation and chemotaxis of smooth muscle cells via MAPK signaling pathways. *J Biol Chem* **282**, 18379-87 (2007).

189. Murphy, C., Murray, A.M., Meaney, S. & Gafvels, M. Regulation by SREBP-2 defines a potential link between isoprenoid and adenosylcobalamin metabolism. *Biochem Biophys Res Commun* **355**, 359-64 (2007).

190. Saridakis, V. et al. The structural basis for methylmalonic aciduria. The crystal structure of archaeal ATP:cobalamin adenosyltransferase. *J Biol Chem* **279**, 23646-53 (2004).

191. Fogarty, M.P., Xiao, R., Prokunina-Olsson, L., Scott, L.J. & Mohlke, K.L. Allelic expression imbalance at high-density lipoprotein cholesterol locus MMAB-MVK. *Hum Mol Genet* (2010).

192. Schadt, E.E. et al. Mapping the genetic architecture of gene expression in human liver. *PLoS Biol* **6**, e107 (2008).

193. Perrine, C.L. et al. Glycopeptide-preferring polypeptide GalNAc transferase 10 (ppGalNAc T10), involved in mucin-type O-glycosylation, has a unique GalNAc-O-Ser/Thr-binding site in its catalytic domain not found in ppGalNAc T1 or T2. *J Biol Chem* **284**, 20387-97 (2009).

194. Lowe, C.E. et al. Large-scale genetic fine mapping and genotype-phenotype associations implicate polymorphism in the IL2RA region in type 1 diabetes. *Nat Genet* **39**, 1074-82 (2007).

195. Pomerantz, M.M. et al. The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. *Nat Genet* **41**, 882-4 (2009).

196. Tuupanen, S. et al. The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. *Nat Genet* **41**, 885-90 (2009).

197. Ahituv, N. et al. Medical sequencing at the extremes of human body mass. *Am J Hum Genet* **80**, 779-91 (2007).

198. Nejentsev, S., Walker, N., Riches, D., Egholm, M. & Todd, J.A. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science* **324**, 387-9 (2009).

199. Valle, T. et al. Mapping genes for NIDDM. Design of the Finland-United States Investigation of NIDDM Genetics (FUSION) Study. *Diabetes Care* **21**, 949-58 (1998).

200. Saaristo, T. et al. Cross-sectional evaluation of the Finnish Diabetes Risk Score: a tool to identify undetected type 2 diabetes, abnormal glucose tolerance and metabolic syndrome. *Diab Vasc Dis Res* **2**, 67-72 (2005).

201. Scott, L.J. et al. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* **316**, 1341-5 (2007).

202. Barski, A. et al. High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823-37 (2007).

203. Kooner, J.S. et al. Genome-wide scan identifies variation in MLXIPL associated with plasma triglycerides. *Nat Genet* **40**, 149-51 (2008).

204. Rada-Iglesias, A. et al. Whole-genome maps of USF1 and USF2 binding and histone H3 acetylation reveal new aspects of promoter structure and candidate genes for common human disorders. *Genome Res* **18**, 380-92 (2008).

205. Guenther, M.G., Levine, S.S., Boyer, L.A., Jaenisch, R. & Young, R.A. A chromatin landmark and transcription initiation at most promoters in human cells. *Cell* **130**, 77-88 (2007).

206. Ferretti, V. et al. PReMod: a database of genome-wide mammalian cis-regulatory module predictions. *Nucleic Acids Res* **35**, D122-6 (2007).

207. Miller, W. et al. 28-way vertebrate alignment and conservation track in the UCSC Genome Browser. *Genome Res* **17**, 1797-808 (2007).

208. Kumar, P., Henikoff, S. & Ng, P.C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* **4**, 1073-81 (2009).

209. Houten, S.M., van Woerden, C.S., Wijburg, F.A., Wanders, R.J. & Waterham, H.R. Carrier frequency of the V377I (1129G>A) MVK mutation, associated with Hyper-IgD and periodic fever syndrome, in the Netherlands. *Eur J Hum Genet* **11**, 196-200 (2003).

210. Cuisset, L. et al. Molecular analysis of MVK mutations and enzymatic activity in hyper-IgD and periodic fever syndrome. *Eur J Hum Genet* **9**, 260-6 (2001).

211. Gaulton, K.J., et al. A map of open chromatin in human pancreatic islets. *Nat Genet* **42**, 255-9 (2010).

212. Leibiger, I.B., Leibiger, B. & Berggren, P.O. Insulin signaling in the pancreatic beta-cell. *Annu Rev Nutr* **28**, 233-51 (2008).

213. King, H., Aubert, R.E. & Herman, W.H. Global burden of diabetes, 1995-2025: prevalence, numerical estimates, and projections. *Diabetes Care* **21**, 1414-31 (1998).

214. Akirav, E., Kushner, J.A. & Herold, K.C. Beta-cell mass and type 1 diabetes: going, going, gone? *Diabetes* **57**, 2883-8 (2008).

215. Rhodes, C.J. Type 2 diabetes-a matter of beta-cell life and death? *Science* **307**, 380-4 (2005).

216. Kahn, S.E. The importance of the beta-cell in the pathogenesis of type 2 diabetes mellitus. *Am J Med* **108 Suppl 6a**, 2S-8S (2000).

217. Polonsky, K.S., Sturis, J. & Bell, G.I. Seminars in Medicine of the Beth Israel Hospital, Boston. Non-insulin-dependent diabetes mellitus - a genetically programmed failure of the beta cell to compensate for insulin resistance. *N Engl J Med* **334**, 777-83 (1996).

218. Frayling, T.M. Genome-wide association studies provide new insights into type 2 diabetes aetiology. *Nat Rev Genet* **8**, 657-62 (2007).

219. Gudmundsson, J. et al. Two variants on chromosome 17 confer prostate cancer risk, and the one in TCF2 protects against type 2 diabetes. *Nat Genet* **39**, 977-83 (2007).

220. Saxena, R. et al. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* **316**, 1331-6 (2007).

221. Sladek, R. et al. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445**, 881-5 (2007).

222. Steinthorsdottir, V. et al. A variant in CDKAL1 influences insulin response and risk of type 2 diabetes. *Nat Genet* **39**, 770-5 (2007).

223. Unoki, H. et al. SNPs in KCNQ1 are associated with susceptibility to type 2 diabetes in East Asian and European populations. *Nat Genet* (2008).

224. Yasuda, K. et al. Variants in KCNQ1 are associated with susceptibility to type 2 diabetes mellitus. *Nat Genet* (2008).

225. Zeggini, E. et al. Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* **316**, 1336-41 (2007).

226. Grarup, N. et al. Association testing of novel type 2 diabetes risk alleles in the JAZF1, CDC123/CAMK1D, TSPAN8, THADA, ADAMTS9, and NOTCH2 loci with insulin release, insulin sensitivity, and obesity in a population-based sample of 4,516 glucose-tolerant middle-aged Danes. *Diabetes* **57**, 2534-40 (2008).

227. Grarup, N. et al. Studies of association of variants near the HHEX, CDKN2A/B, and IGF2BP2 genes with type 2 diabetes and impaired insulin release in 10,705

Danish subjects: validation and extension of genome-wide association studies. *Diabetes* **56**, 3105-11 (2007).

228. Pascoe, L. et al. Common variants of the novel type 2 diabetes genes CDKAL1 and HHEX/IDE are associated with decreased pancreatic beta-cell function. *Diabetes* **56**, 3101-4 (2007).

229. Perry, J.R. & Frayling, T.M. New gene variants alter type 2 diabetes risk predominantly through reduced beta-cell function. *Curr Opin Clin Nutr Metab Care* **11**, 371-7 (2008).

230. Henikoff, S. Nucleosome destabilization in the epigenetic regulation of gene expression. *Nature Reviews Genetics* **9**, 15-26 (2008).

231. Keene, M.A. & Elgin, S.C. Micrococcal nuclease as a probe of DNA sequence organization and chromatin structure. *Cell* **27**, 57-64 (1981).

232. Levy, A. & Noll, M. Chromatin fine structure of active and repressed genes. *Nature* **289**, 198-203 (1981).

233. Wu, C. The 5' ends of Drosophila heat shock genes in chromatin are hypersensitive to DNase I. *Nature* **286**, 854-60 (1980).

234. Wu, C., Wong, Y.C. & Elgin, S.C. The chromatin structure of specific genes: II. Disruption of chromatin structure during gene activity. *Cell* **16**, 807-14 (1979).

235. Giresi, P.G., Kim, J., McDaniell, R.M., Iyer, V.R. & Lieb, J.D. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res* **17**, 877-85 (2007).

236. Hogan, G.J., Lee, C.K. & Lieb, J.D. Cell cycle-specified fluctuation of nucleosome occupancy at gene promoters. *PLoS Genet* **2**, e158 (2006).

237. Buck, M.J. & Lieb, J.D. A chromatin-mediated mechanism for specification of conditional transcription factor targets. *Nat Genet* **38**, 1446-51 (2006).

238. Bucher, P. et al. Assessment of a novel two-component enzyme preparation for human islet isolation and transplantation. *Transplantation* **79**, 91-7 (2005).

239. Latif, Z.A., Noel, J. & Alejandro, R. A simple method of staining fresh and cultured islets. *Transplantation* **45**, 827-30 (1988).

240. Boj, S.F., Parrizas, M., Maestro, M.A. & Ferrer, J. A transcription factor regulatory circuit in differentiated pancreatic cells. *Proc Natl Acad Sci U S A* **98**, 14481-6 (2001).

241. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* **18**, 1851-8 (2008).

242. Boyle, A.P., Guinney, J., Crawford, G.E. & Furey, T.S. F-Seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics* **24**, 2537-8 (2008).

243. Rozowsky, J. et al. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat Biotechnol* **27**, 66-75 (2009).

244. Buck, M.J., Nobel, A.B. & Lieb, J.D. ChIPOTle: a user-friendly tool for the analysis of ChIP-chip data. *Genome Biol* **6**, R97 (2005).

245. Wingender, E. et al. TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res* **28**, 316-9 (2000).

246. Frith, M.C. et al. Detection of functional DNA motifs via statistical over-representation. *Nucleic Acids Res* **32**, 1372-81 (2004).

247. Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W.W. & Lenhard, B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res* **32**, D91-4 (2004).

248. Kim, T.H. et al. Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell* **128**, 1231-45 (2007).

249. Gunton, J.E. et al. Loss of ARNT/HIF1beta mediates altered gene expression and pancreatic-islet dysfunction in human type 2 diabetes. *Cell* **122**, 337-49 (2005).

250. Giardine, B. et al. Galaxy: a platform for interactive large-scale genome analysis. *Genome Res* **15**, 1451-5 (2005).

251. Bao, L., Zhou, M. & Cui, Y. CTCFBSDB: a CTCF-binding site database for characterization of vertebrate genomic insulators. *Nucleic Acids Res* **36**, D83-7 (2008).

252. Dennis, G., Jr. et al. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol* **4**, P3 (2003).

253. Thomas, P.D. et al. PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res* **13**, 2129-41 (2003).

254. Su, A.I. et al. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A* **101**, 6062-7 (2004).

255. Prokopenko, I. et al. Variants in MTNR1B influence fasting glucose levels. *Nat Genet* **41**, 77-81 (2009).

256. Sandhu, M.S. et al. Common variants in WFS1 confer risk of type 2 diabetes. *Nat Genet* **39**, 951-3 (2007).

257. Luco, R.F. et al. A conditional model reveals that induction of hepatocyte nuclear factor-1alpha in Hnf1alpha-null mutant beta-cells can activate silenced genes postnatally, whereas overexpression is deleterious. *Diabetes* **55**, 2202-11 (2006).

258. Luco, R.F., Maestro, M.A., Sadoni, N., Zink, D. & Ferrer, J. Targeted deficiency of the transcriptional activator Hnf1alpha alters subnuclear positioning of its genomic targets. *PLoS Genet* **4**, e1000079 (2008).

259. Ishihara, H. et al. Pancreatic beta cell line MIN6 exhibits characteristics of glucose metabolism and glucose-stimulated insulin secretion similar to those of normal islets. *Diabetologia* **36**, 1139-45 (1993).

260. Nagy, P.L., Cleary, M.L., Brown, P.O. & Lieb, J.D. Genomewide demarcation of RNA polymerase II transcription units revealed by physical fractionation of chromatin. *Proc Natl Acad Sci U S A* **100**, 6364-9 (2003).

261. Odom, D.T. et al. Control of pancreas and liver gene expression by HNF transcription factors. *Science* **303**, 1378-81 (2004).

262. Bell, G.I. & Polonsky, K.S. Diabetes mellitus and genetically programmed defects in beta-cell function. *Nature* **414**, 788-91 (2001).

263. Oliver-Krasinski, J.M. & Stoffers, D.A. On the origin of the beta cell. *Genes Dev* **22**, 1998-2021 (2008).

264. Di Lorenzo, T.P., Peakman, M. & Roep, B.O. Translational mini-review series on type 1 diabetes: Systematic analysis of T cell epitopes in autoimmune diabetes. *Clin Exp Immunol* **148**, 1-16 (2007).

265. McCarthy, M.I. & Zeggini, E. Genome-wide association scans for Type 2 diabetes: new insights into biology and therapy. *Trends Pharmacol Sci* **28**, 598-601 (2007).

266. Xi, H. et al. Identification and characterization of cell type-specific and ubiquitous chromatin regulatory structures in the human genome. *PLoS Genet* **3**, e136 (2007).

267. Heintzman, N.D. et al. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108-12 (2009).

268. Cuddapah, S. et al. Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome Res* **19**, 24-32 (2009).

269. Atouf, F., Czernichow, P. & Scharfmann, R. Expression of neuronal traits in pancreatic beta cells. Implication of neuron-restrictive silencing factor/repressor element silencing transcription factor, a neuron-restrictive silencer. *J Biol Chem* **272**, 1929-34 (1997).

270. Fujitani, Y. et al. Targeted deletion of a cis-regulatory region reveals differential gene dosage requirements for Pdx1 in foregut organ differentiation and pancreas formation. *Genes Dev* **20**, 253-66 (2006).

271. Gerrish, K., Van Velkinburgh, J.C. & Stein, R. Conserved transcriptional regulatory domains of the pdx-1 gene. *Mol Endocrinol* **18**, 533-48 (2004).

272. Sander, M. et al. Homeobox gene Nkx6.1 lies downstream of Nkx2.2 in the major pathway of beta-cell formation in the pancreas. *Development* **127**, 5533-40 (2000).

273. Edwards, C.A. et al. The evolution of the DLK1-DIO3 imprinted domain in mammals. *PLoS Biol* **6**, e135 (2008).

274. Lyssenko, V. et al. Clinical risk factors, DNA variants, and the development of type 2 diabetes. *N Engl J Med* **359**, 2220-32 (2008).

275. Bouatia-Naji, N. et al. A polymorphism within the G6PC2 gene is associated with fasting plasma glucose levels. *Science* **320**, 1085-8 (2008).

276. Helgason, A. et al. Refining the impact of TCF7L2 gene variants on type 2 diabetes and adaptive evolution. *Nat Genet* **39**, 218-25 (2007).

277. Guelen, L. et al. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* **453**, 948-51 (2008).

278. Dillon, N. Gene regulation and large-scale chromatin organization in the nucleus. *Chromosome Res* **14**, 117-26 (2006).

279. Gilbert, N. et al. Chromatin architecture of the human genome: gene-rich domains are enriched in open chromatin fibers. *Cell* **118**, 555-66 (2004).

280. Hindorff, L.A. et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* **106**, 9362-7 (2009).

281. Song, L. & Crawford, G.E. DNase-seq: A High-Resolution Technique for Mapping Active Gene Regulatory Elements across the Genome from Mammalian Cells. *CSH Protoc* **2010**, **(**2010).

282. Gaulton, K.J. et al. A map of open chromatin in human pancreatic islets. *Nat Genet* **42**, 255-9 (2010).

283. Wang, Z. et al. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat Genet* **40**, 897-903 (2008).

284. Robertson, G. et al. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods* **4**, 651-7 (2007).

285. Chorley, B.N. et al. Discovery and verification of functional single nucleotide polymorphisms in regulatory genomic regions: current and developing technologies. *Mutat Res* **659**, 147-57 (2008).

286. Bulyk, M.L. Computational prediction of transcription-factor binding site locations. *Genome Biol* **5**, 201 (2003).

287. Lenhard, B. & Wasserman, W.W. TFBS: Computational framework for transcription factor binding site analysis. *Bioinformatics* **18**, 1135-6 (2002).

288. Lantz, K.A. et al. Foxa2 regulates multiple pathways of insulin secretion. *J Clin Invest* **114**, 512-20 (2004).

289. Perera, H.K. et al. Expression and shifting subcellular localization of the transcription factor, Foxd3, in embryonic and adult pancreas. *Gene Expr Patterns* **6**, 971-7 (2006).

290. Ohlsson, H., Karlsson, K. & Edlund, T. IPF1, a homeodomain-containing transactivator of the insulin gene. *EMBO J* **12**, 4251-9 (1993).

291. Hu, M., Yu, J., Taylor, J.M., Chinnaiyan, A.M. & Qin, Z.S. On the detection and refinement of transcription factor binding sites using ChIP-Seq data. *Nucleic Acids Res*. Jan 6, 2010 [Epub ahead of print]

292. Conde, L. et al. PupaSuite: finding functional single nucleotide polymorphisms for large-scale genotyping purposes. *Nucleic Acids Res* **34**, W621-5 (2006).

293. Ponomarenko, J.V. et al. rSNP_Guide, a database system for analysis of transcription factor binding to DNA with variations: application to genome annotation. *Nucleic Acids Res* **31**, 118-21 (2003).

294. Reumers, J. et al. Joint annotation of coding and non-coding single nucleotide polymorphisms and mutations in the SNPeffect and PupaSuite databases. *Nucleic Acids Res* **36**, D825-9 (2008).

295. Schmitt, A.O., Assmus, J., Bortfeldt, R.H. & Brockmann, G.A. CandiSNPer: a web-tool for the identification of candidate SNPs for causal variants. *Bioinformatics* **26**, 969-70 (2010).

296. Fong, C. et al. GWAS analyzer: integrating genotype, phenotype and public annotation data for genome-wide association study analysis. *Bioinformatics* **26**, 560-4 (2010).

297. Shen, T.H., Carlson, C.S. & Tarczy-Hornoch, P. Evaluating the accuracy of a functional SNP annotation system. *BMC Bioinformatics* **10 Suppl 9**, S11 (2009).

298. Shen, T.H., Carlson, C.S. & Tarczy-Hornoch, P. SNPit: a federated data integration system for the purpose of functional SNP annotation. *Comput Methods Programs Biomed* **95**, 181-9 (2009).

299. Borowiak, M. & Melton, D.A. How to make beta cells? *Curr Opin Cell Biol* **21**, 727-32 (2009).

300. Simonis, M., Kooren, J. & de Laat, W. An evaluation of 3C-based methods to capture DNA interactions. *Nat Methods* **4**, 895-901 (2007).

301. van Berkum, N.L. & Dekker, J. Determining spatial chromatin organization of large genomic regions using 5C technology. *Methods Mol Biol* **567**, 189-213 (2009).

302. Tiwari, V.K. & Baylin, S.B. Mapping networks of protein-mediated physical interactions between chromatin elements. *Curr Protoc Mol Biol* **Chapter 21**, Unit 21 161-13 (2010).

303. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289-93 (2009).

304. Wallace, C. et al. The imprinted DLK1-MEG3 gene region on chromosome 14q32.2 alters susceptibility to type 1 diabetes. *Nat Genet* **42**, 68-71 (2010).

305. Raum, J.C. et al. FoxA2, Nkx2.2, and PDX-1 regulate islet beta-cell-specific mafA expression through conserved sequences located between base pairs -8118 and -7750 upstream from the transcription start site. *Mol Cell Biol* **26**, 5735-43 (2006).

306. Schuit, F., Moens, K., Heimberg, H. & Pipeleers, D. Cellular origin of hexokinase in pancreatic islets. *J Biol Chem* **274**, 32803-9 (1999).