The post-GWAS era: paving the way from association to functional insight

Stephanie Rosse

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2013

Reading Committee:

Chris Carlson, Chair

Stephanie Fullerton

Paul Auer

Alex Reiner

Program Authorized to Offer Degree:

Public Health Genetics

University of Washington

ABSTRACT

The post-GWAS era: paving the way from association to functional insight

Stephanie Rosse

Chair of the Supervisory Committee:
Christopher Carlson, Ph.D.
Department of Epidemiology

Considerable effort has been devoted to uncovering genetic variation that can be used to tailor prevention and treatment strategies for common diseases. Genome-wide association studies (GWAS) have proven to be an extremely powerful approach to identify susceptibility loci for common diseases and other complex traits. However, the profusion of GWAS results over the last eight years has not been accompanied by rapid translation into clinical benefit. In part, this unmet challenge stems from difficulty in interpreting the underlying biology of GWAS loci, and the unwieldiness of functional follow-up studies. Most GWAS signals are localized outside of protein coding regions, implicating transcriptional regulation as an important mechanism for susceptibility of complex disease. Unfortunately, the tools developed to predict and test perturbations to amino acid changes are inadequate for studying regulatory mechanisms. Therefore, interpreting the functional mechanisms underlying GWAS signals will likely be one of the most pressing challenges in the translation of genetic information into clinical application. Each chapter of this dissertation will explore various methods to translate GWAS associations into clinically actionable information.

The first chapter develops a statistical and bioinformatics framework to fine-map known loci associated with fasting glucose levels that could be used to study a wide range of GWAS loci. This study leverages the differences in genomic architecture between populations to identify likely functional variants within known signals identified in a single population. Chapter two introduces a

novel high-throughput *in vitro* assay to test variants previously prioritized for laboratory analysis. The functional interpretation of particular genetic variants will, in many circumstances, require recruitment of participants possessing particular genetic variants of interest. For that reason, chapter 3 examines the feasibility of applying current ethical recommendations for recruiting genetic research participants based on their individual research results. Specifically, a content analysis of current consent documents retrieved from the database of Genotypes and Phenotypes (dbGaP) was conducted to evaluate the consistency of consent documentation with ethical recommendations for re-contacting participants to invite their enrollment into a new study based on their genetic research results.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Genome-wide association studies (GWAS) have been a major advancement towards precision medicine, leading to ~2,000 robust associations with complex diseases since 2005 [1]. However, the clinical utility of these results is limited by modest effect sizes exerted by each variant, and the unclear functionality of the associated variants. Despite the benefit anticipated from genetic research findings, improved preventive and therapeutic strategies from GWAS results will likely take longer than the expected timeframe to integrate scientific discoveries into clinical application [1]. The reason for this is due in part to the nature of GWAS loci and the need for higher throughput functional assays. Although robust GWAS signals identify specific genomic regions that contain a functional variant (or variants) the signal often implicates many correlated variants, spanning multiple genes. Unlike Mendelian disease, variants associated with complex phenotypes exert modest effects, and tend to be positioned outside of coding regions. Thus, an important step toward precision medicine will be the identification of causal variants and the interpretation of their effects on gene expression. Despite these challenges, there are several approaches that will help translate the vast catalog of trait-associated variants into a deeper understanding of molecular mechanisms underlying common disease.

Refining European GWAS signals through trans-ethnic meta-analysis

The power of GWAS to detect an association is influenced by the frequency of the causal variant, and its correlation with other variants, a phenomenon known as linkage disequilibrium (LD). This genetic architecture varies between populations, with the extent of LD increasing in ancestral populations that were more geographically distant from Africa. Population substructure can lead to false positive associations between non-causal variants and a trait or disease (phenotype) [2]. As a

result, genome-wide association studies have historically restricted analyses to more homogeneous populations and attempted to control for global ancestry. To date the majority of GWAS have been conducted in populations of European descent [3]. Given that populations of European ancestry have, on average, more extensive LD than populations with African ancestry, association signals are typically very broad, with each disease-associated tagSNP representing many potential functional variants. Laboratory follow-up of variants is both expensive and time intensive, making broad GWAS signals particularly difficult to interpret definitively. Furthermore, because there are differences in allele frequencies between populations, there exists uncertainty in the generalizability of GWAS results across populations.

Differences in genetic architecture between populations can be leveraged to differentiate likely causal variants from non-functional variants in LD with the causal variant. Under the premise that functional variants will have a more significant and consistent association than non-functional variants, different patterns of LD and allele frequencies between populations can help to refine GWAS signals in trans-ethnic studies. Furthermore, trans-ethnic studies of GWAS loci will help to broaden benefits to public health by exploring the generalizability of current results. In addition to fine-mapping known loci, this approach may help to identify additional functional variants specific to different populations in the same regions as known variants (allelic heterogeneity).

Chapter 1 of this dissertation conducts a trans-ethnic study to evaluate the generalizability of GWAS loci associated with fasting glucose levels. This study aims to leverage the differences in genomic architecture between populations to refine GWAS signals and prioritize likely functional variants for laboratory follow-up. In addition, this study will also search for novel secondary signals within known loci independently associated with fasting glucose. This analysis will be restricted to 14 previously identified fasting glucose loci, and other loci associated with metabolic traits. Therefore,

identification of other population specific loci across the genome is beyond the scope of this study, but remains an important aim for future research.

Prioritizing variants for laboratory follow-up through bioinformatics framework

In addition to the complications caused by LD, interpreting the biology of GWAS signals is difficult because the vast majority of associations are located outside of coding sequence. This suggests that the underlying variants associated with common diseases exert their effects through transcriptional regulation. Unfortunately, regulatory mechanisms are quite complicated and require more onerous strategies to disentangle than highly penetrant coding variants uncovered by genetic linkage studies. A major challenge in the translation of genetic associations into clinical benefit is prioritizing variants within a GWAS signal for more cumbersome functional assays.

Fortunately, we now have a wealth of information about potential regulatory regions that is publicly available to use for annotation of variants across the genome. Data produced by the Encyclopedia of DNA Elements (ENCODE) Project Consortium[4] is accessible through the UCSC Genome Browser and attempts to provide a comprehensive catalog of all functional elements in the human genome. Though such a collection will likely never be complete, given the capricious nature of regulatory mechanisms (in that they vary with cell type, environment, and time) the combinatorial groupings of various biological datasets across various assays are useful for detecting regulatory elements. By aligning these datasets with genomic variants, GWAS associated variants can be interrogated with a functional lens to select a few candidate SNPs falling in suspected regulatory regions. In some instances this may dramatically reduce the number of variants to take forward and assay within a cell line or other model system.

Development of a high-throughput functional assay

Once an association between a genetic variant and a trait has been detected and replicated, the next challenge is to elucidate the molecular mechanism behind the association. For regulatory regions this can be particularly difficult since there are often several functional candidate variants, which often fall in non-coding regions with one or more candidate genes. Development of a model system to rapidly investigate how allelic variants alter transcription levels of a target gene would provide critical insight for interpreting GWAS loci. Unfortunately, current models are likely to introduce too much background noise to reproducibly detect modest changes in gene expression and require testing one genetic variant at a time. Chapter 2 will describe the next step in the translational pipeline by introducing a novel allelic reporter assay to test functional hypotheses. This reporter assay combines flow cytometry and next generation sequencing to simultaneously evaluate transcriptional effects from many allelic plasmids that were transfected in a single event. Although this novel assay was developed with a poly-A tail variant known to alter mRNA degradation, it could easily be used to test other types of putative regulatory variants for function.

Comparison of consent documentation with ethical guidelines for genotype driven recruitment of participants using the NIH data repository

The majority of the 'post GWAS' strategies described thus far have relied on the use of large centralized data repositories containing information that cannot readily be linked to a person's identity. Although a genomic commons is intended for the reuse of information by secondary investigators, the utility of this resource is limited because ascertainment of additional genetic or phenotypic information may be necessary to pursue some research questions. In many cases the biology behind a genetic association with disease is unclear. Recruiting research participants known

to carry a particular variant may be a powerful tool to gain substantial new insight. In circumstances where a variant is rare in the general population, genotype driven recruitment (GDR) may be the only feasible option, in light of the declining federal support for genetic research and inability to find an adequate number of participants from the general population. Unfortunately, recruiting participants from a genetic repository presents several challenges for informed consent beyond the logistical issues arising from de-identification of data.

A primary concern of GDR is avoiding harm from the return of unwanted information while maintaining the transparency of new research aims and/or reason for the participant's eligibility. Despite the many challenges, this type of study design may be essential to maximizing the benefits from the expense and effort the public has invested in genetic research. The last Chapter of this dissertation describes an analysis of consent forms made available in the National Institutes of Health's database of Genotypes and Phenotypes (dbGAP), aimed at identifying potential conflicts with recent ethical recommendations for GDR which would hinder this recruitment strategy.

Scope of dissertation

The rapid evolvement of scientific advancements has introduced many new strategies for moving beyond the assumptions made in single-variant association testing to methods capable of detecting more biologically plausible genetic associations. Beyond the strategies that will be discussed in the details of this dissertation there are many new statistical and functional approaches emerging. For instance, recent advances in parallel processing allow for the use of more computationally intensive nonparametric statistical methods. These statistical approaches may help to identify relationships among networks of intergenic SNPs around biologically relevant genetic pathways (ie epistasis) that were previously undetectable by GWAS [5-7]. Furthermore, it has been

hypothesized that a significant portion of heritability in common diseases missed by GWAS will be uncovered by analysis of gene-environment interactions and a better understanding of the interactions with the microbiome [8-10].

Beyond the aforementioned hypothesis-generating approaches, there are new methods for high-throughput genome-wide *in vivo* discovery of enhancers that may prove to be a promising approach [11]. Single nucleotide variants only represent part of human genetic variation underlying deleterious mutations. Therefore, it will also be necessary to look at structural variation beyond that of a single nucleotide, such as small insertion or deletion events, large CNVs, retrotransposition events and inversions that have been mostly ignored until recently [12-14]. A predictive framework will need to be established for distinguishing functional versus neutral mutations for these types of variants as well [11]. One single approach will not be sufficient to translate genetic information into clinical use. In fact, it is likely that each of these approaches will need to be simultaneously pursued in order to extract the full benefit of genomic research.

In addition to the significant scientific challenges that impede the translation of genetic information into new treatments and preventions, there are many ethical, legal and social challenges that must be addressed as we move toward functional genomic studies using data within central repositories. For instance, given that most large-scale genomic studies have been conducted almost exclusively in populations of European descent, there is substantial concern that the health benefits arising from application of this genetic information will exacerbate existing health disparities in marginalized populations [3]. There are also significant legal implications that must be considered in the development of diagnostic methods, therapeutic treatments and prevention strategies. For instance, there has been much debate over the patentability of genetic information [15]. This dissertation will only briefly scratch the surface of issues related to biobanking such as privacy, identifiability, obtaining informed consent and the return of incidental findings. Other significant

ethical, legal and social challenges facing the translation of genetic information into clinical benefit, such as equitable access, patenting of genetic information, need for increased community input and many risks related to group harms are beyond the scope of this analysis.

CHAPTER 1

Population Architecture using Genomics and Epidemiology (PAGE): Trans-ethnic insight on the genetics of fasting glucose

1.1 ABSTRACT

Genome-wide association studies (GWAS) have discovered a large number of common variants associated with type 2 diabetes mellitus (T2DM) in populations of European descent. Despite this success, the functional variants within these loci and the underlying biology driving these associations remain largely unknown. Recent extension of GWAS across ethnicities has demonstrated that the majority of signals are directionally consistent and similar in magnitude across major population groups. In contrast, differences in linkage disequilibrium and allele (LD) frequencies between populations can help localize causal variants and potentially discover secondary independent signals within known loci. Understanding trans-ethnic genetic influences on the underlying traits of T2DM, such as concentrations of fasting plasma glucose (FPG), will provide substantial insight into disease pathway(s) of T2DM.

This study fine-mapped 14 previously identified loci for fasting glucose, by investigating approximately 200,000 variants in multiple populations (13,582 Hispanic American, 17,414 African American, 966 Asian American, and 437 American Indian populations) from 5 studies in the population architecture using genetic epidemiology (PAGE) consortium. Eleven loci (2p23.3 *GCKR,*2q31.1 *G6PC2,* 3q21.1 *ADCY5,*3q26.2 *SLC2A2/EIF5A2,* 7p21.2 *DGKB,* 7p13 *GCK,* 9p24.2 *GLIS3, TCF7L2,*11p11.2 *CRY2,*11q12.2 *FADS2,* and 11q14.3 *MTNR1B*) were replicated in the trans-ethnic meta-analysis, and evidence was provided for fine-mapping in five of those validated FPG loci. These findings suggest that loci associated with glycemic traits may have generalizable

effects across studied populations and that trans-ethnic meta-analysis may assist in prioritizing likely functional candidates.

1.2 INTRODUCTION

Type 2 diabetes mellitus (T2DM) and its associated complications are a major and growing health burden worldwide [16]. Exploring the genetics of traits associated with T2DM, such as fasting plasma glucose (FPG), across ethnicities may improve our understanding of the underlying biology. Examining GWAS signals in more than one population may help to detect the underlying causal variant(s) responsible for the statistical association, and may provide insight into whether biological function differs across ethnic groups.

Divergent evolutionary and migratory histories have influenced both nucleotide diversity and patterns of linkage disequilibrium (LD) across populations [17, 18]. These differences in genetic architecture can reduce the power of initial discovery of disease-associated variants in certain populations when studied independently. However, trans-ethnic meta-analysis may address this issue by increasing the observed frequency of the functional SNP or proxies correlated with the functional SNP. Greater haplotype diversity, as seen in populations with African ancestry, may help to identify functional variants, due in part to reduced LD with neighboring SNPs [19]. For example, a recent study of FPG loci in African Americans (AA) from the PAGE consortium identified an independent signal tagged by rs77719485 in the known *G6PC2* region (unpublished data). Thus, trans-ethnic fine-mapping studies have the potential to narrow the set of candidate functional variants at GWAS loci. It is posited that the evidence of association will be stronger for variants in high LD with the underlying causal variants than the majority of GWAS index SNPs that were selected to be maximally informative of genetic structure rather than function. Therefore, dense genotyping of

variants at known GWAS loci may help to localize the underlying causal variants by identifying those with stronger associations than the index variants [20-23]. The MetaboChip is a high-density custom genotyping array designed to replicate and fine-map known GWAS signals for metabolic and atherosclerotic/cardiovascular traits [24, 25]. The fine-mapping SNPs spanned a wide range of allele frequencies including rare (minor allele frequency (MAF)<0.005) and less common (0.005≤MAF<0.05) SNPs selected from the International HapMap Project and the August 2009 release of the 1000 Genomes Project. SNPs annotated as nonsynonymous, essential splice site or stop codon were included regardless of MAF, design score, or the presence of nearby SNPs [25] . The MetaboChip contains densely spaced SNPs at 14 known FPG loci. Thus, this platform is particularly useful for fine-mapping known loci without adding the uncertainty associated with imputation methods in other fine-mapping approaches [26].

Allelic heterogeneity, in which different variants at the same region impact a common phenotype, has been identified at some GWAS loci associated with T2DM in populations from Southeast Asia and populations of European descent [25, 27]. Most recently, a second independent association, rs77719485 in the *G6PC2* region, was identified in African Americans from PAGE (unpublished data). As such, current evidence suggests that exploring known genetic variants across populations can help identify novel associations. Despite this, allelic heterogeneity has not been comprehensively evaluated for FPG across multiple populations.

A critical step in the extension of GWAS results is establishing a functional role for specific SNPs within a trait associated locus. To aid in functional discovery of FPG loci, 14 previously identified loci were fine-mapped by investigating approximately 200,000 variants in multiple populations (13,582 Hispanic American, 17,414 African American, 966 Asian, and 437 American Indian populations) from five studies in the population architecture using genetic epidemiology

(PAGE) consortium. As a secondary aim, this study searched for pleiotropic signals in other cardiovascular and metabolic trait loci included on the MetaboChip platform.

1.3 SUBJECTS AND METHODS

1.3.1 Study population

The Population Architecture using Genomics and Epidemiology (PAGE I) consortium was funded by the National Human Genome Research Institute to investigate the epidemiologic architecture of well-replicated genetic variants associated with human diseases or traits [28]. The PAGE I study consisted of a coordinating center and four consortia: Epidemiologic Architecture for Genes Linked to Environment (EAGLE), which is based on data from Vanderbilt University Medical Center's DNA biobank (BioVU); the Multiethnic Cohort Study (MEC); the Women's Health Initiative (WHI); and Causal Variants Across the Life Course (CALiCo), a consortium of five cohort studies: Atherosclerosis Risk in Communities (ARIC), Coronary Artery Risk Development in Young Adults (CARDIA), the Cardiovascular Health Study (CHS), the Hispanic Community Health Study/Study of Latinos (SOL), and the Strong Heart Study [28]. The PAGE II consortium consists of MEC, WHI, CALiCo, ARIC, CARDIA, CHS/SOL, SHS, as well as Mount Sinai School of Medicine's (MSSM) DNA biobank, (Bio*Me*).

This analysis included participants from MEC, WHI, ARIC, CARDIA, CHS, SOL, and MSSM. Participants with BMI<16.5 kg/m^2 and BMI>70 kg/m^2 were excluded from further analysis with the assumption that these extremes could be attributable to data coding errors, an underlying illness or possibly to a familial syndrome. Analysis was also limited to adults defined as age 18 years or older. All studies were approved by Institutional Review Boards at their respective sites, and all study participants provided written informed consent.

1.3.2 Anthropometric and fasting glucose measurements

FPG concentrations were measured using standard assays, at laboratories specific to each PAGE site. For WHI, ARIC, CARDIA, CHS, SOL, and MSSM, BMI was calculated from height and weight measured at time of study enrollment. In MEC, self-reported height and weight were used to calculate baseline BMI. Individuals self-reporting that they have ever been diagnosed with diabetes, individuals who reported taking diabetes medications, and individuals with FPG concentrations consistent with diabetes (i.e., ≥126 mg/dl or ≥7 mmol/L) were excluded from analysis.

1.3.3 Genotyping and quality control

Genotyping was performed using the MetaboChip, the design of which has been described elsewhere [25]. In brief, the MetaboChip, a custom Illumina iSelect genotyping array of nearly 200,000 SNP markers, is designed to cost effectively analyze putative association signals identified through GWAS meta-analyses of many glucose-related metabolic and cardiovascular traits and to fine-map established loci [25]. MetaboChip SNPs were selected from the catalogs developed by the International HapMap and 1000 Genomes Projects [25]. More than 122,000 SNPs were included to fine-map 257 GWAS loci of 23 traits [25]. The boundaries around each GWAS index SNP were determined by identifying all SNPs with $r^2 \geq 0.5$ with the index SNP, and then expanding the initial boundaries by 0.02cM in either direction using the HapMap-based genetic map. All 1000 Genomes Pilot 1 SNPs were considered as potential fine mapping SNPs, unless the SNP allele frequency was <0.01 in all three HapMap populations (CEU, YRI and HBC/JPT) [25]. SNPs were excluded if (a) the Illumina design score was <0.5 or (b) there were SNPs within 15bp in both directions of the SNP of interest with allele frequency of >0.02 among Europeans (CEU). SNPs annotated as nonsynonymous, essential splice site, or stop codon were included regardless of allele frequency, design score, or

nearby SNPs in the primer [25]. A total of 14 glucose GWAS loci identified were represented for signal fine mapping (see table 1.1).

Table 1.1 Characterization of 14 genomic regions fine-mapped for fasting glucose concentrations

Chr	Locus	Base pair range (Build 36)	N SNPs
1	*PROX1*	212191441 , 212234131	100
2	*GCKR*	27243138 , 27805162	782
2	*G6PC2*	169460886 , 169522901	184
3	*ADCY5*	124459609 , 124689609	676
3	*SLC2A2*	172014805 , 172251865	568
7	*DGKB*	14151613 , 15112045	3000
7	*GCK, YKT6*	44188528 , 44232602	98
9	*GLIS3*	4233162 , 4300558	341
10	*TCF7L2*	112957728 , 113043029	404
11	*CRY2*	45662738 , 46119405	715
11	*MADD, MTCH2*	46878217 , 48047879	1653
11	*FADS3*	61262159 , 61508200	548
11	*MTNR1B*	92306695 , 92364969	142
15	*C2CD4A, C2CD4B*	59886474 , 60307401	721

Samples were genotyped at the University of Southern California Epigenome Center (MEC), Translational Genomics Research Institute (WHI), and Human Genetics Center of the University of Texas-Houston (ARIC, CARDIA, CHS). Each center genotyped 90 HapMap YRI (Yoruba in Ibadan, Nigeria) samples to facilitate cross-study quality control (QC), as well as 2-3% study-specific blinded replicates to assess genotyping quality. Genotypes were called separately for each study using GenomeStudio with the GenCall 2.0 algorithm. Samples were called using study-specific cluster definitions (based on samples with call rate >95%) and kept in the analysis if call rate was >95%. SNPs with GenTrain score <0.6, cluster separation score <0.4, call rate <0.95, and Hardy-Weinberg Equilibrium $p < 1 \times 10^{-6}$ were excluded. This analysis also excluded SNPs based on Mendelian errors in 30 YRI trios >1, replication errors >2 with discordant calls (when comparing across studies) in 90 YRI samples >3, and discordant calls for 90 YRI genotyped in PAGE versus

HapMap database >3. For African American meta-analysis SNPs were excluded if present in less than three studies. Given the small number of studies involved for the other ethnicities, SNPs were not required to be present in more than one study.

The present analysis for Bio*Me* derives from 12,726 samples genotyped using the Illumina HumanOmniExpress+ v1.1 BeadChip, which queries >240,000 exonic markers and an additional >718,000 genome-wide markers. For the current analyses, related and cryptically related individuals were removed. Related persons were identified using PLINK by estimating identical-by-descent (IBD) statistics for all pairs. Individuals were defined as "related" when IBD > 0.185. When apparent first-degree relative pairs were identified, the member with the lower call rate was excluded. EIGENSOFT [29, 30] was used to determined principal components of ancestry in each study separately and participants who mapped outside the clusters of African Americans, or Hispanic Americans were excluded from further analyses as described elsewhere [24]. Specifically, samples with an inbreeding coefficient (F) above 0.15 were removed from further analysis [31]. Further exclusions include SNPs with MAF <1%, those whose distribution was inconsistent with the Hardy-Weinberg Equilibrium expectation ($P < 0.001$), SNPs with low call rate and those with evidence of batch/plate effects. There were 711,270 genotyped SNPs available for imputation that was performed using IMPUTE2 using the 1000 Genome data (March 2012 version 3).

1.3.4 Statistical analysis

Models were analyzed independently in each study population and ethnicity to evaluate the association between glucose and SNP using multivariate linear regression. An additive genetic model adjusted for age, sex, study site (as applicable), smoking status (current versus former/never), and ancestry principal components was used in each study. For evaluation of signal fine mapping, all models assessing the association between glucose and SNPs also adjusted for BMI (continuous).

Fixed-effect models with inverse variance weighting were used to pool the study-specific association results as implemented in METAL[32]. Q-statistics and I^2 were used to measure heterogeneity across studies and across ethnicities. METAL was used to first meta-analyze results across each study for each ethnicity independently and then meta-analyzed METAL results for each ethnicity. For generalization of index SNPs from GWAS, a nominal significance level (p=0.05) was used. For signal fine mapping, the locus-wide P value of 5.6×10^{-5}, which is 0.05 divided by 886 (average number of SNPs on the MetaboChip within each of the 14 glucose loci) was used to establish significance.

LD pattern comparisons between AA, Hispanic Americans (HA), Asian American, and European populations (EA) in the trans-ethnic results was facilitated through the use of data from the 1000 Genomes Phase 1 study populations (AFR, AMR, ASN, EUR, respectively) [33]. LocusZoom plots [34] were used to graphically display the fine-mapping results. SNP positions from NCBI build 37 were used and recombination rates were estimated from 1000 Genomes Project data.

1.4 FINE-MAPPING RESULTS

This study consisted of 13,582 Hispanic, 966 Asian, 437 American Indian, and 124 Hawaiian Americans from five studies within PAGE and included a total of 17,414 African Americans, of which 15,852 were studied in previous PAGE analysis (unpublished data). The total size of the study for the trans-ethnic meta-analysis was 32,523. The average age of this cohort was 53 years (see tables 1.2 and 1.3). About 75% of study participants are women. Within and across studies, men tended to have lower BMI than women and slightly higher mean fasting glucose. A total of 9,932 genetic variants were tested across 14 FPG loci. In search of novel loci in MetaboChip regions, a total of 169,718 variants were studied.

1.4.1 Replication of known fasting glucose loci

Among the 14 FPG GWAS loci identified in studies of European ancestry, 11 loci (2p23.3 *GCKR,* 2q31.1 *G6PC2,* 3q21.1 *ADCY5,* 3q26.2 *SLC2A2/EIF5A2,* 7p21.2 *DGKB,* 7p13 *GCK,* 9p24.2 *GLIS3,*10q25.2 *TCF7L2,* 11p11.2 *CRY2,*11q12.2 *FADS2,* and 11q14.3 *MTNR1B*) contained index SNPs with significant evidence of replicated association at $p<0.05$ (see table 1.4). Of the 11 replicated regions, associations at seven loci (2p23.3 *GCKR,* 2q31.1 *G6PC2,* 3q21.1 *ADCY5,* 3q26.2 *SLC2A2,* 7p21.2 *DGKB,* 7p13 *GCK,* and 11q14.3 *MTNR1B)* also reached the regional discovery threshold of 5.6×10^{-5}, which is 0.05 with Bonferroni correction for the average number of SNPs (n=885) across FPG loci.

The lead SNP (the most significant SNP in the trans-ethnic analysis) in most loci had a minor allele frequency >0.05, with little evidence for heterogeneity. Among the 11 replicated loci, the lead SNPs in *GCKR* (rs1260326, $P=3.49 \times 10^{-10}$), *GCPC2* (rs560887, $P=7.25 \times 10^{-23}$), and *MTNR1B* (rs10830963, $P=1.89 \times 10^{-17}$) were the same index SNPs identified through GWAS (see table 1.4).

Table 1.2 Summary characteristics for all 32,523 study participants

	Male				Female				Total			
	N	Mean (SD)	Min	Max	N	Mean (SD)	Min	Max	N	Mean (SD)	Min	Max
Age (years)	6926	48 (12)	18	92	21,286	55 (9)	18	93	28,212	53 (10)	18	93
BMI (kg/m^2)	6926	28.3 (0.1)	16.6	58	21,286	29.8 (0.1)	16.5	59.9	28,212	29.4 (0.1)	16.5	59.9
Glucose (mmol/L)	6926	5.3 (0.002)	0.3	7	21,286	5.2 (0.001)	1.1	6.9	28,212	5.2 (0.001)	0.3	7

Table 1.3 Characteristics of all participants, stratified by cohort and ethnicity

Multiethnic Cohort Study-MEC

	Male				Female			
	N	Mean (SD)	Min	Max	N	Mean (SD)	Min	Max
Japanese Americans								
Age (years)	307	61 (8)	45	76	325	59 (8)	45	75
BMI (kg/m^2)	307	25.2 (3.3)	16.6	40.0	325	24.8 (16.8)	16.8	48.9
Glucose (mmol/L)	307	5.0 (0.7)	0.3	7.0	325	4.9 (0.7)	2.9	6.8
Hawaiians								
Age (years)	37	54 (7)	46	74	85	55 (7)	45	72
BMI (kg/m^2)	37	29.4 (4.2)	21.8	40.6	85	28.4 (5.6)	20.0	50.4
Glucose (mmol/L)	37	5.1 (1.0)	1.4	6.9	85	4.8 (0.7)	1.4	6.3

Table 1.3 Characteristics of all participants, stratified by cohort and ethnicity, continued

Women's Health Initiative (WHI)

	Male				Female			
	N	Mean (SD)	Min	Max	N	Mean (SD)	Min	Max
Hispanic Americans								
Age (years)					3228	60 (7)	49	79
BMI (kg/m^2)					3228	28.4 (5.3)	16.5	55.8
Glucose (mmol/L)					3228	5.1 (0.6)	3.2	6.9
Asian Americans								
Age (years)					418	65 (7)	50	79
BMI (kg/m^2)					418	24.3 (4.2)	16.6	56.4
Glucose (mmol/L)					418	5.4 (0.6)	4.0	6.9
Native Americans								
Age (years)					437	60 (7)	50	79
BMI (kg/m^2)					437	29.4 (6.1)	17.3	51.5
Glucose (mmol/L)					437	5.28 (0.6)	2.9	6.9

Hispanic Community Health Study / Study of Latinos (SOL)

	Male				Female			
	N	Mean (SD)	Min	Max	N	Mean (SD)	Min	Max
Age (years)	4067	43.4 (14.1)	18	75	5805	43.9 (13.5)	18	75
BMI (kg/m^2)	4067	28.8 (5.1)	17.1	57.5	5805	29.7 (6.0)	16.5	59.5
Glucose (mmol/L)	4067	5.4 (0.5)	3.5	6.9	5805	5.2 (0.5)	3.1	6.9

Mount Sinai School of Medicine (MSSM)

	Male				Female			
	N	Mean (SD)	Min	Max	N	Mean (SD)	Min	Max
Age (years)	569	47.9 (13.7)	20	89	993	48.8 (14.7)	19	89
BMI (kg/m^2)	569	27.8 (5.8)	18.6	58.0	993	31.0 (7.8)	18.6	59.5
Glucose (mmol/L)	569	4.8 (0.7)	2.5	6.9	993	4.7 (0.7)	2.3	6.9

Table 1.3 Characteristics of all participants, stratified by cohort and ethnicity, continued

Pankow et. al study

	Male				Female			
	N	Mean (SD)	Min	Max	N	Mean (SD)	Min	Max
Age (years)	1946	54.7 (8.0)	44	92	9995	59.2 (7.6)	33	93
BMI (kg/m^2)	1946	27.9 (4.8)	18.5	54.9	9995	30.6 (6.4)	18.5	59.9
Glucose (mmol/L)	1946	5.3 (0.7)	1.4	6.9	9995	5.2 (0.6)	1.2	6.9

Table 1.4 *Generalizability of known fasting glucose loci for 11 replicated regions*

	SNP							Effect		Significance		
	rsID	AFR CAF/r^2	AMR CAF/r^2	ASN CAF/r^2	EUR CAF/r^2	Position (hg19)	Alleles†	mmol/L (±SE)	Pval	P_{het}	Nom.	Reg.
11q14.3-*MTNR1B*												
LEAD	rs10830963	0.03/1	0.23/1	0.44/1	0.29/1	92708710	C/G	-0.06 (0.007)	1.89E-17	0.04	Yes	Yes
INDEX												
2p23.3-*GCKR*												
LEAD	rs1260326	0.88/1	0.59/1	0.44/1	0.59/1	27730940	A/G	-0.03 (0.006)	3.49E-10	0.85	Yes	Yes
INDEX												
2q31.1-*G6PC2*												
LEAD	rs560887	0.99/1	0.79/1	0.97/1	0.69/1	169763148	A/G	-0.07 (0.007)	7.25E-23	0.91	Yes	Yes
INDEX												
3q21.1-*ADCY5*												
LEAD	rs7614016	0.09/0.6	0.27/0.95	0.00/1	0.18/0.92	123070426	A/G	-0.03 (0.007)	1.32E-6	0.60	Yes	Yes
INDEX	rs11708067	0.12/0.6	0.28/0.95	0.00/1	0.18/0.92	123065778	G/A	-0.03 (0.006)	9.10E-6	0.07	Yes	Yes
3q26.2-*SLC2A2*												
LEAD	rs10513689	0.41/0.66	0.19/0.95	0.02/0.9	0.14/1	170727658	A/G	-0.05 (0.008)	8.04E-8	0.29	Yes	Yes
INDEX	rs11920090	0.34/0.66	0.19/0.95	0.02/0.9	0.14/1	170717521	A/T	-0.02 (0.006)	0.001	0.0009	Yes	No
7p21.2-*DGKB*												
LEAD	rs2214618	0.91/<0.2	0.67/0.41	0.52/<0.2	0.63/ 0.27	15029541	A/G	-0.02 (0.005)	8.41E-6	0.87	Yes	Yes
INDEX	rs2191349	0.51/<0.2	0.50/0.41	0.66/<0.2	0.51/ 0.27	15064309	A/C	0.02 (0.005)	0.0009	0.77	Yes	No

Table 1.4 Generalizability of known fasting glucose loci for 11 replicated regions, continued

	SNP							Effect		Significance		
	rsID	AFR CAF/r^2	AMR CAF/r^2	ASN CAF/r^2	EUR CAF/r^2	Position (hg19)	Alleles †	mmol/ L (±SE)	Pval	P_{het}	Nom.	Reg.
7p13-*GCK*												
LEAD	rs2908286	0.20/0.99	0.19/0.96	0.20/0.88	0.18/1	44234737	A/G	0.06 (0.006)	4.4E-21	0.34	Yes	Yes
INDEX	rs1799884	0.20/0.99	0.19/0.96	0.18/0.88	0.18/1	44229068	A/G	0.06 (0.006)	7.3E-21	0.32	Yes	Yes
9p24.2-*GLIS3*												
LEAD	rs75385796	0.00/ <0.2	0.00/ <0.2	0.02/<0.2	0.00/<0.2	4247921	A/T	0.31 (0.1)	0.004	0.05	Yes	No
INDEX	rs7034200	0.64/ <0.2	0.53/ <0.2	0.44/<0.2	0.52/<0.2	4289050	A/C	0.01 (0.01)	0.004	0.70	Yes	No
10q25.2 *TCF7L2*												
LEAD	rs11195489	0.03/<0.2	0.00/<0.2	0.00/<0.2	0.00/<0.2	113007238	A/G	-0.05 (0.02)	0.001	0.38	Yes	No
INDEX	rs10885122	0.20/<0.2	0.85/<0.2	0.93/<0.2	0.89/<0.2	113042093	A/G	-0.01 (0.006)	0.02	0.84	Yes	No
11p11.2-*CRY2*												
LEAD	rs12794698	0.07/0.68	0.50/1	0.24/1	0.49/1	45871861	A/G	-0.02 (0.005)	0.0005	0.14	Yes	No
INDEX	rs11605924	0.05/0.68	0.50/1	0.24/1	0.49/1	45873091	C/A	-0.02 (0.005)	0.001	0.16	Yes	No
11q12.2-*FADS2*												
LEAD	rs2238003	0.03/<0.2	0.36/<0.2	0.31/<0.2	0.27/<0.2	61523735	A/G	-0.02 (0.006)	0.0003	0.32	Yes	No
INDEX	rs174547	0.04/<0.2	0.54/<0.2	0.41/<0.2	0.36/<0.2	61570783	A/G	0.02 (0.006)	0.002	0.18	Yes	No

In the remaining three loci, (*GLIS3*, *FADS2*, and *TCF7L2*), where the lead SNP was not in high LD with the index SNP in any of the studied populations, there was not sufficient evidence to determine whether the lead variant was a secondary independent signal. The lead SNP in the *GLIS3* region, rs75385796, was rare and observed more frequently in Asian Americans from PAGE (coding allele frequency (CAF)=0.03 in Asian Americans from PAGE). As a result of the small proportion of individuals in this study with Asian ancestry, this SNP only reached nominal significance (P=0.03) with significant evidence of heterogeneity (P_{het}=0.05). Similarly, the lead SNP in *TCF7L2*, rs11195489, was only observed in AA (CAF=0.04 in AA from PAGE) and did not reach the threshold for discovery. Lastly, the lead SNP in *FADS2*, rs2238003, was observed in all study populations and was not in LD with the index. However, like *GLIS3* and *TCF7L2* lead SNPs, rs2238003 did not reach the discovery threshold and was not taken explored further in conditional analysis.

1.4.2 Fine-mapping fasting glucose loci

To investigate whether LD patterns across populations can be leveraged to narrow association signals, both in terms of number of significant variants and physical region, locus zoom plots[34] for a large GWAS[35] conducted in Europeans were compared to PAGE results from both HA (the largest subgroup specific to this study) and the trans-ethnic meta-analysis across all four populations. In five out of the seven loci reaching regional significance (*MTNR1B*, *GCKR*, *G6PC2*, *SLC2A2*, and *DGKB*), trans-ethnic differences in LD assisted with narrowing association signals (see figures 1.1-1.2).

Figure 1.1 *Fine-Mapping of regions: index SNP is the lead significant SNP across ethnicities (MTNR1B, GCKR, G6PC2)*

EA
Manning et al.
N=95,500

HA-PAGE
N=13,582

Multi-ethnic
PAGE
N=32,3999

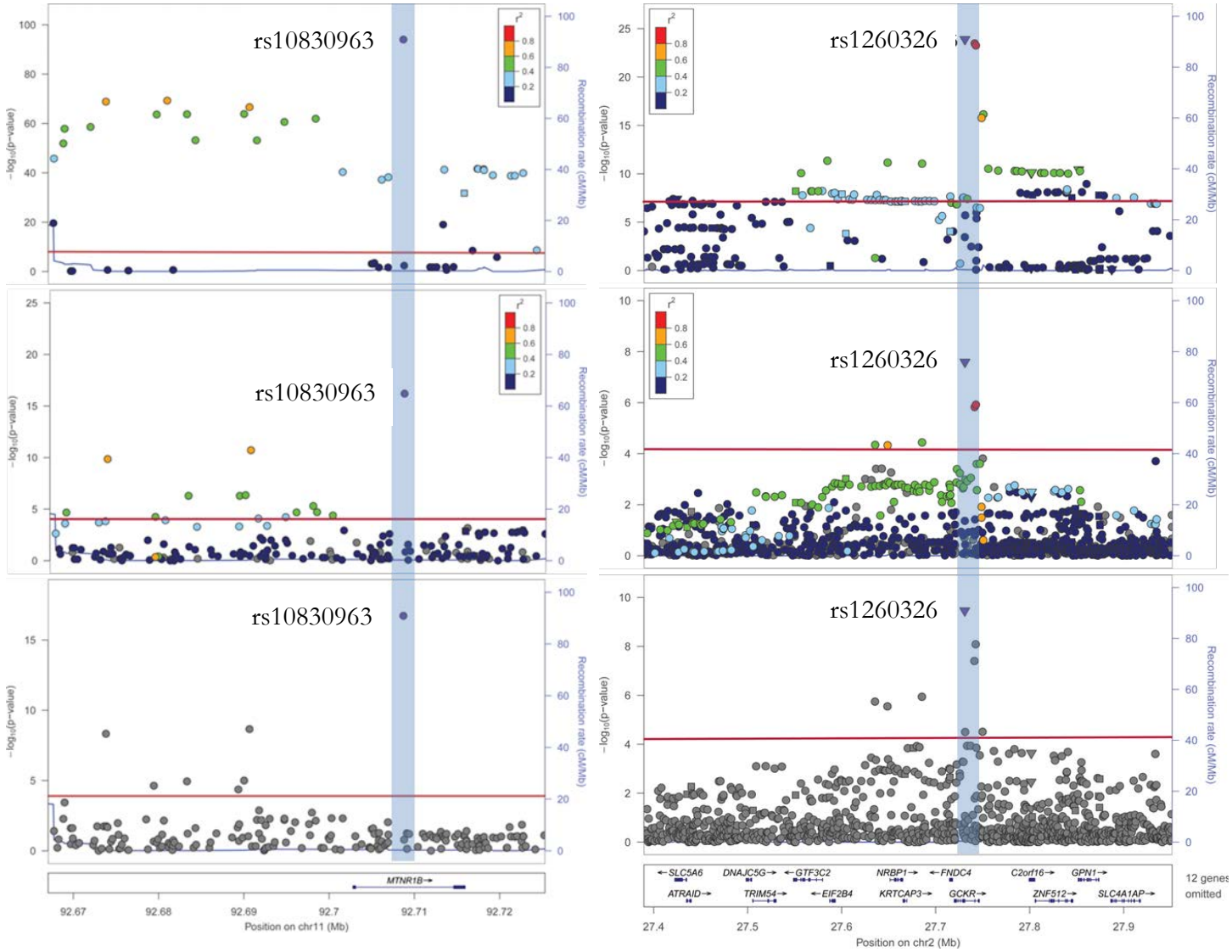


Figure 1.1 *Fine-Mapping of regions: index SNP is the lead significant SNP across ethnicities (MTNR1B, GCKR, G6PC2) continued*

EA Manning et al. N=95,500

HA-PAGE N=13,582

Multi-ethnic PAGE N=32,399

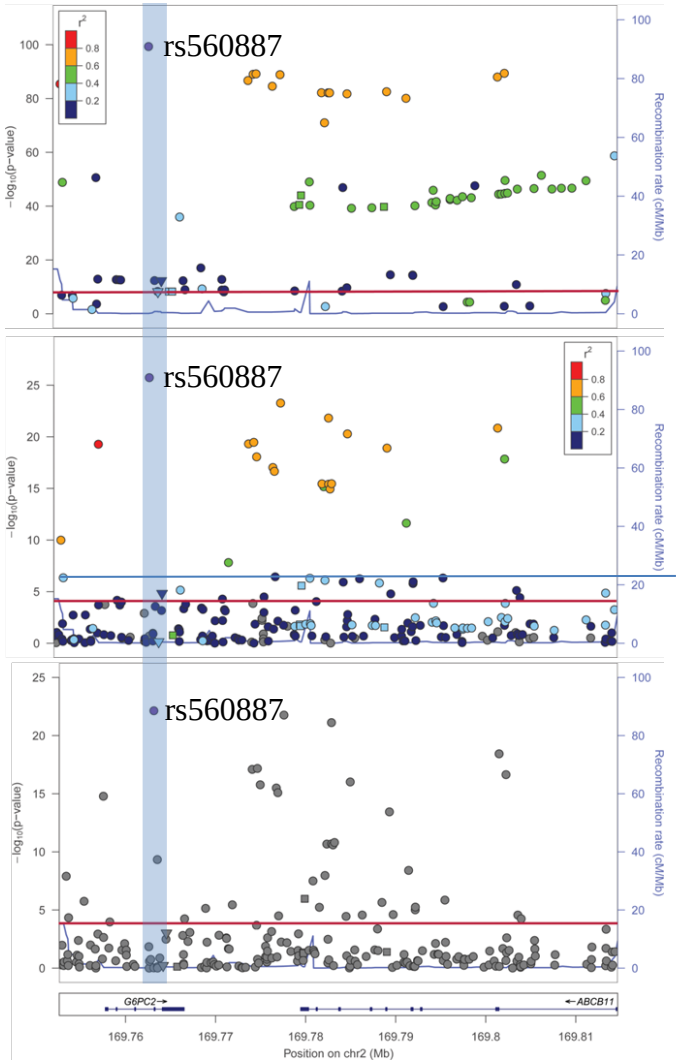


Figure 1.1 shows regional plots of the results from three loci (*MTNR1B, GCKR* respectively on previous page and *G6PC2* to the left) where the index SNP was the most significant SNP in Hispanics and the trans-ethnic results from PAGE. For each locus, results from the European GWAS [35], HA from PAGE, and the trans-ethnic analysis from PAGE are shown from the top to bottom panels, respectively. The horizontal red line represents the significance thresholds for each analysis. Among these loci, fine-mapping was most apparent at the *GCKR* locus, where two variants highly correlated with the lead variant, rs1260326, were as significant as the index in the GWAS, but were slightly less significant in the PAGE results. The GWAS signals at the *MTNR1B* and *G6PC2* loci were more refined in Europeans than the *GCKR* locus, lessening the impact of fine-mapping for these regions.

Of the remaining eight loci, where the lead SNP in the trans-ethnic analysis differed from the index SNP, the lead SNP in five loci (rs7614016 in *ADCY5*, rs10513689 in *SLC2A2*, rs2214618 in *DGKB*, rs2908286 in *GCK*, and rs12794698 in *CRY2*) were in moderate-to-strong LD with the index SNP in one or more 1000G Phase 1 populations, r^2 ranging from 0.41 to 1.0 in AFR, AMR or ASN (see figure 1.2).

Figure 1.2 *Fine-Mapping of regions: index and lead significant SNP are different across ethnicities (SLC2A2, ADCY5, GCK, DGKB)*

EA Manning et al. N=95,500

HA-PAGE N=13,582

Multi-ethnic PAGE N=32,399

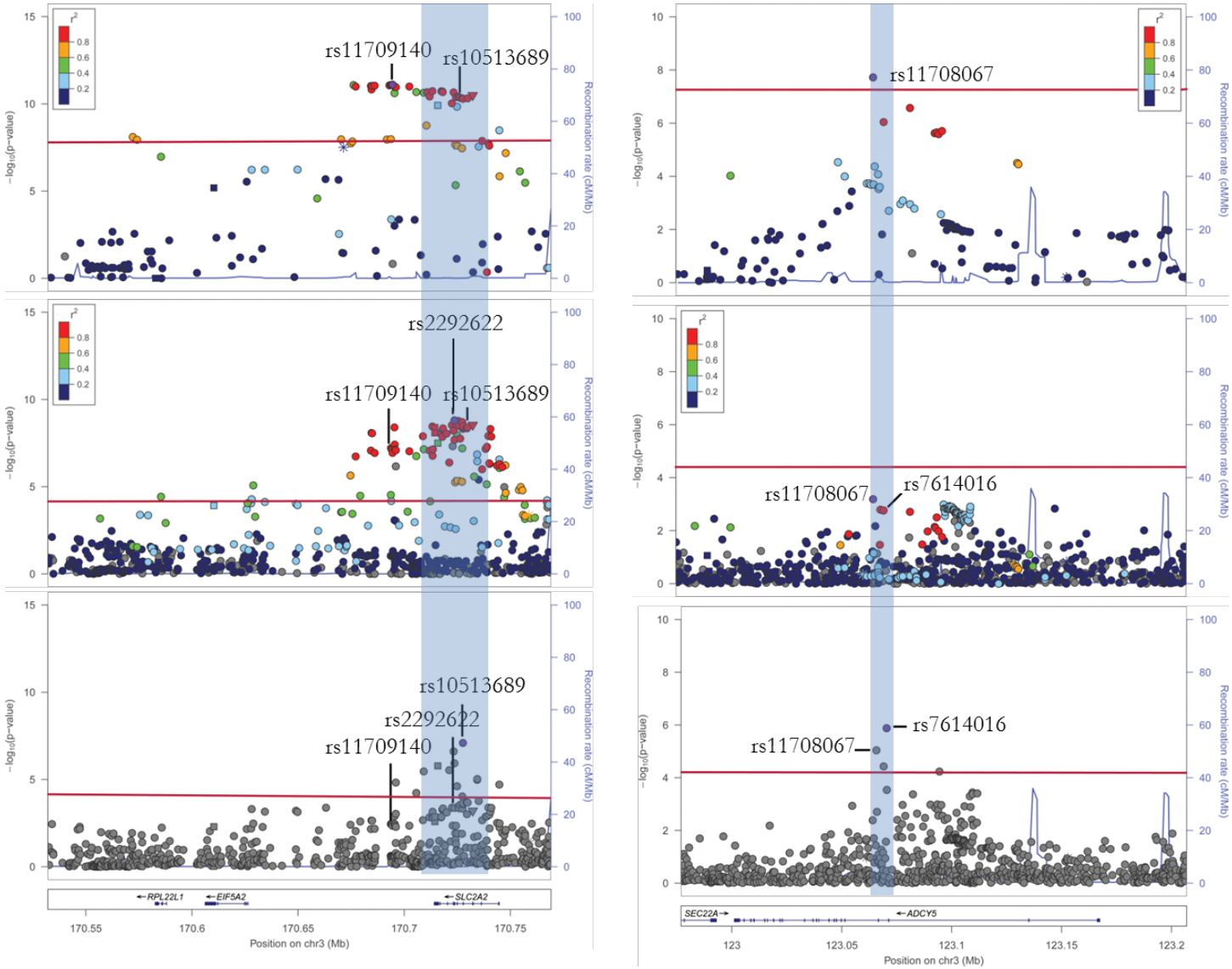
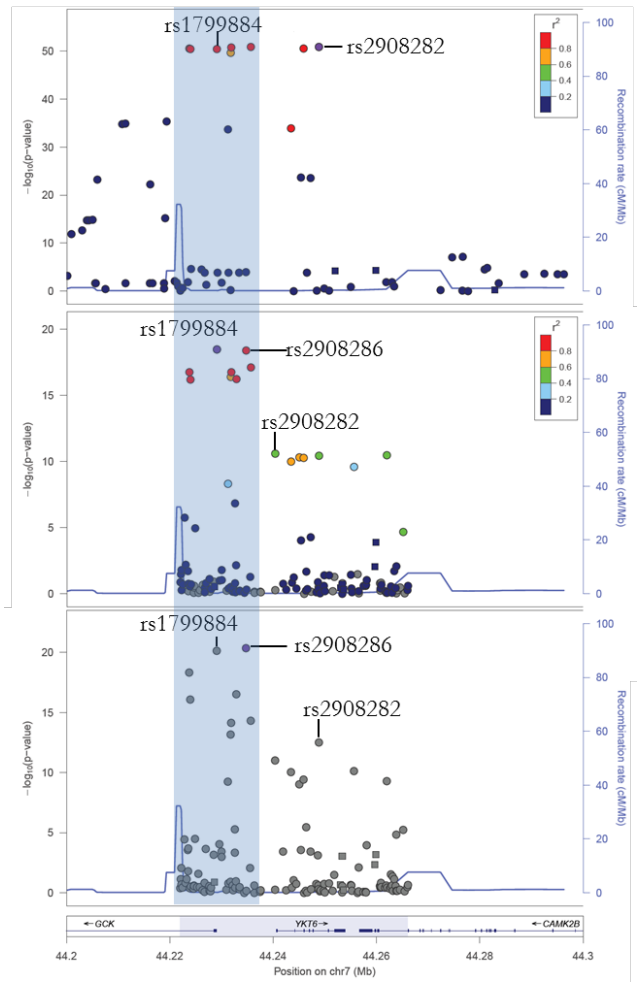


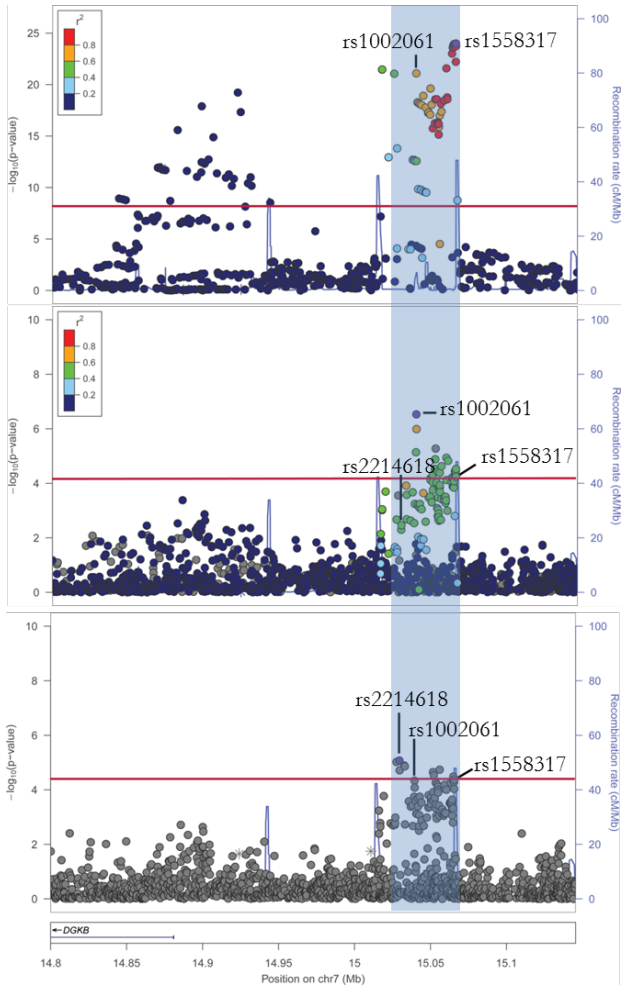
Figure 1.2 *Fine-Mapping of regions: index and lead significant SNP are different across ethnicities (SLC2A2, ADCY5, GCK, DGKB) continued*

EA
Manning et al.
N=95,500

HA-PAGE
N=13,582

Multi-ethnic
PAGE
N=32,399





For fine-mapping purposes, it was hypothesized that functional variants, or strong proxies for functional variants, were likely to reach more significant p-values than bystander variants unrelated to concentrations of FPG. Therefore, in assessing whether fine-mapping was accomplished, at each locus the number of highly significant variants (5.6×10^{-5}) in LD with the lead variant ($r^2>0.4$ in at least one subpopulation) from both HA and the trans-ethnic meta-analysis results were compared to the number of correlated ($r^2>0.4$ in EUR) variants reaching genome-wide significance ($5x10^{-8}$) in Europeans from the MAGIC study population [2].

Using this approach, fine-mapping was particularly evident in three loci, *GCKR*, *G6PC2* and *SLC2A2*. In the *GCKR* locus, the number of likely candidate variants was reduced from 26 in the GWAS signal to just six variants in HA from the PAGE study. Similarly, in *G6PC2*, 41 variants present in the GWAS signal were reduced to 21 variants in the HA study population. However, trans-ethnic meta-analysis was unable to refine the signal further in either of these loci. Alternatively, fine-mapping for the *SLC2A2* locus was most apparent in the trans-ethnic meta-analysis where differences in genetic architecture between populations allowed a broad signal in Europeans and HA to be refined. In HA, the number of variants that were highly significant and correlated was much greater than that of the GWAS signal (approximately 80 versus 32), and all but one of the variants in the GWAS signal were also present in the HA signal. In contrast to population specific results, there were only 11 variants that were highly significant in the trans-ethnic results.

The fine-mapping at *DGKB* and *MTNR1B* was less evident. Results in the *DGKB* region were interesting in that the lead significant variant was different in the results for each population specific analysis and in the trans-ethnic analysis. Of note, the lead significant variant was never found to pass the significance threshold across analyses, though the lead in the trans-ethnic (rs1558317) was correlated with lead HA ($r^2=0.5$) and close to the significance threshold (see figure 1.2). Although the number of significant variants was substantially reduced in HA and trans-ethnic

results the discrepancies across results make interpretation of fine-mapping more difficult. At the *MTNR1B* fine-mapping was less evident because the extent of LD at this locus was less extensive when compared to other GWAS loci for fasting glucose. As such, strong candidates had already been identified at the *MTNR1B* locus, though trans-ethnic analysis provided further support for these previously identified SNPs.

1.5 DEVELOPMENT OF BIOINFORMATICS FRAMEWORK

1.5.1 The Framework

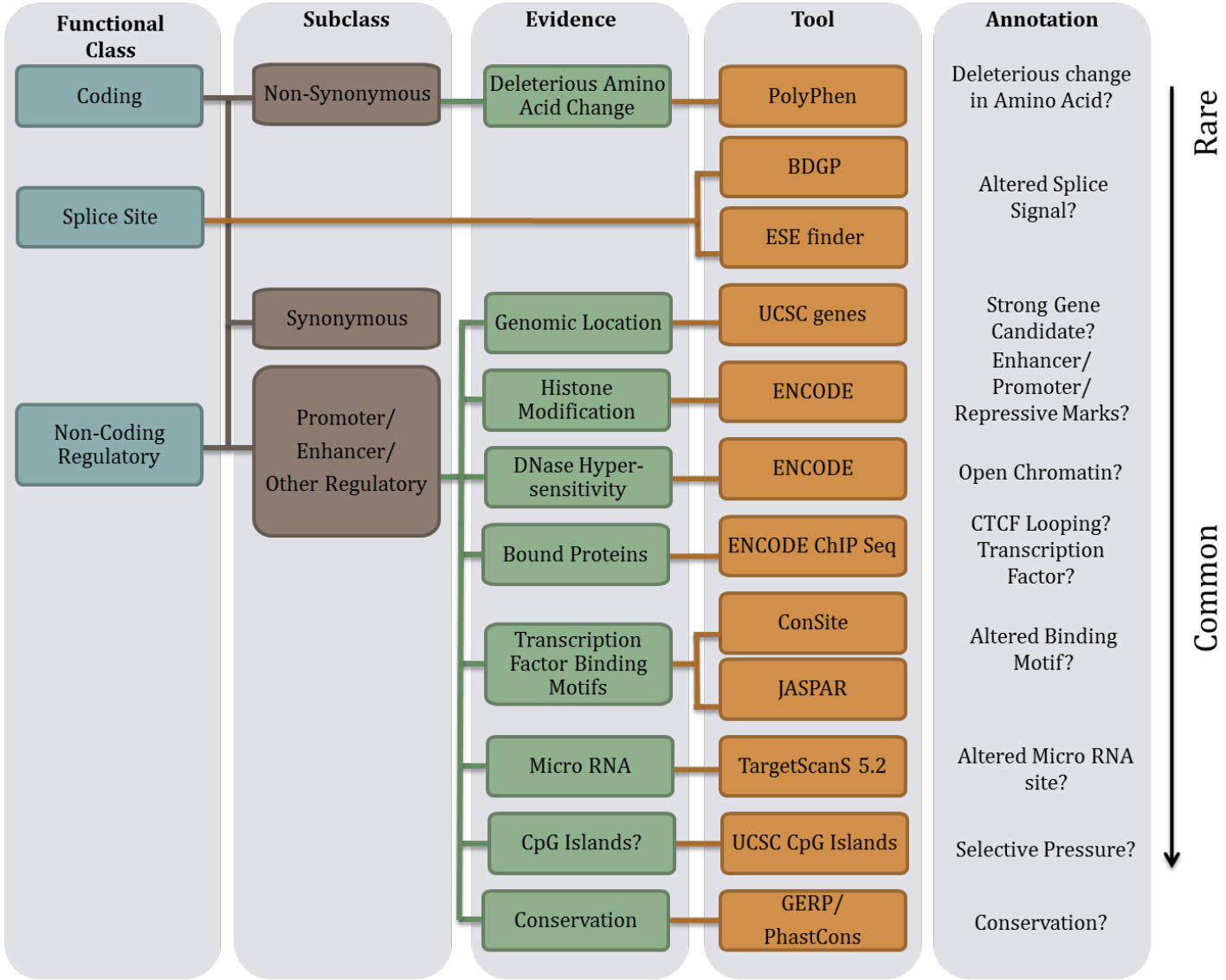
Genome-wide association (GWA) studies of fasting glucose have successfully replicated several susceptibility loci in populations of European ancestry [35]. Although trans-ethnic meta-analysis can be used to fine-map broad GWAS signals, results from this strategy may still implicate several statistically indistinguishable variants. Furthermore, most of the loci identified thus far are positioned in non-coding regions suggesting that the underlying functional variants responsible for the associations likely impact regulatory regions. To further prioritize likely causal variants and generate testable functional hypotheses about the underlying mechanisms, a framework was developed to utilize publicly available biological datasets (see figure 1.3).

Several bioinformatics databases are available for the functional characterization of putative disease causing loci such as HaploReg [36] (maintained by the Broad Institute), and the University of California, Santa Cruz (UCSC) genome browser [37]. Annotation of non-protein-coding regions operates under the hypothesis that trait-associated alleles exert their effects by influencing transcriptional levels through multiple regulatory mechanisms. HaploReg is useful for an initial survey of a large number of correlated variants for regulatory DNA elements such as DNaseI

Hypersensitivity (DHS), transcription factor binding sites, histone modifications, eQTLs, protein-binding motif analysis, and evolutionarily conserved regions. This database extracts regulatory information from ENCODE and NIH Roadmap, and creates a large transcription factor (TF) motif library of position weight matrices (PWMs) from sources such as TRANSFAC and JASPAR. These datasets are then integrated with known variants from the 1000 Genomes Phase I populations. LD with each of variant is also calculated allowing retrieval of functional information for tagSNPs and the sets of variants correlated with the tagSNPs.

Although HaploReg is a useful tool for an initial screen, compiling various functional datasets at a single nucleotide can be misleading, such as reporting the dbSNP function for unrelated genes in the same region. In addition, because the DNase I nuclease is unable to cleave DNA in the region where a transcription factor or other regulatory protein is bound [38], at nucleotide resolution functional SNPs disrupting protein binding may lack DHS signal. Although DNase I footprinting [39] can be used to detect dips in DHS signal consistent with protein binding, HaploReg does not contain this data. Furthermore, reproducibility of these high-throughput biological datasets is an important step in establishing true signal from background noise and in comparing data generated from different laboratories [40]. Unfortunately, this quality-control pipeline has only been used in a limited number of datasets from ENCODE/Roadmap and is not available through HaploReg. Therefore, after a permissive initial screen for variants with functional evidence from any of the biological datasets within HaploReg, the UCSC genome browser is useful for a deeper interrogation of potential causal variants. Additional tools such as ESE finder [41] are also useful for predicting alterations in transcripts when variants are found in splice site regions.

Figure 1.3 *Bioinformatics Framework*



Functional hypotheses were made for (1) all significant lead SNPs and (2) all variants tagged by the lead SNPs (r^2>0.4 in one or more 1000 genomes populations). This list was generated using HaploReg. Each list was then annotated for potential regulatory evidence consistent with enhancers, promoters, insulators, silencers, and other effects related to gene expression. Although SNPs in high LD with the lead SNPs were often included in tests of association, frequentist testing is not reflective of actual function [42]. As such, the underlying functional SNP may not be the SNP reaching the top significant association. Therefore, all correlated SNPs were aligned with a combined browser view of currently available ENCODE tracks in the UCSC Genome Browser and compared each allelic region for altered transcription factor binding site (TFBS) motifs using JASPAR, ConSite, and HaploReg. For variants falling near potential splice sites the ESE finder was to search for differential splicing signals.

Since distal enhancers often facilitate cell-type specific expression, it is helpful to look for evidence in a variety of cell lines in addition to those related to the trait. Chromatin remodeling and histone modification can provide additional evidence for cell-specific regulatory elements. The "Histone Modification" tracks within the ENCODE Analysis Hub (http://ftp.ebi.ac.uk/pub/databases/ensembl/encode/integration_data_jan2011/hub.txt) and "Uniformly Signal" within the Roadmap Epigenomics Data Complete Collection at Wash U VizHub (http://vizhub.wustl.edu/VizHub/RoadmapReleaseAll.txt), can be used to query data that has been processed through the uniform quality pipeline. The methylation and acetylation of histone proteins changes chromatin accessibility for transcription and such marks can serve as a powerful tool for identifying both enhancer and promoter regions. There are three histone modification marks that are particularly informative for most regulatory regions, namely H3K4me1, H3k27ac, and H3k4me3. The H3K4me1 histone mark is associated with enhancers downstream of transcription start sites and the H3k27ac histone mark is similarly thought to enhance transcription. Alternatively the

H3K4Me3 mark is associated with active promoters. Although a large number of cell lines and tissues were queried to assess cell-type specific regulation, HepG2 (liver) and PANC1 (pancreas) carcinoma cell lines from ENCODE assays and normal adult liver, pancreas, and pancreatic islets tissue from Roadmap assays were of particular interest for regulation of fasting glucose.

Regulatory regions are susceptible to DNase I cutting and ENCODE has assayed hypersensitivity in a large collection of cell types. The precision of the DNase I cluster track is somewhat better than that of chromatin methylation and acetylation patterns detected by ChIP-seq. As such, the DNase I hypersensitivity track was used to provide a more specific demarcation of open chromatin within broader histone modification signals.

The ChIP-Seq TFBS track provides evidence for the binding of specific proteins. Using the ChIP-seq method, proteins that bind to specific regions were identified and could therefore be related to differential binding affinities between alleles. As an example, CTCF is a cohesion protein that can act as an activator, a repressor/silencer or an insulator. When binding chromatic insulators it can prevent interactions between promoters and nearby enhancers or silencers. However it also mediates long-range chromatin looping, which can bring enhancers into proximity with a gene's promoter [43]. Therefore, after establishing evidence of regulatory histone modification or open chromatin structure, ChIP-seq TFBS was used to identify potential disruption of protein binding.

Though less specific than ChIP-Seq methodology, JASPAR, ConSite and HaploReg PWM databases were used to query a larger number of conserved TFBS to predict alterations in conserved binding motifs between reference and alternate alleles. After identification of proteins that bind to a particular region through ChIP-seq, these PWM libraries can be particularly useful for generating hypotheses about perturbation of protein binding. Use of 46-way PhastCons track in the Genome Browser was used as secondary evidence for a regulatory region, but lack of conservation did not rule out a functional candidate because regulatory elements can be species specific.

Transcription factor binding motifs are scattered throughout the genome, and may or may not be related to functional regulation. Although open chromatin structure is now known to be actively regulated, there remain some stochastic properties that would allow transcription factors to temporarily rest in a non-functional region of the genome. As such, the co-occurrence of both epigenetic remodeling and protein binding from independent assays were considered to provide stronger evidence than the presence of either signal alone. Furthermore, because ChIP-seq signals are broader than the underlying functional region, DHS can be used to more precisely demark the region of interest. Finally, variants that have evidence from each of these assays and also disrupt the binding affinity as predicted through PWM of conserved binding motifs are considered to have the strongest bioinformatics evidence of *cis* regulatory effects.

Replicated signals from the trans-ethnic meta-analysis were also explored for previous evidence of expression quantitative trait loci (eQTLs) using several data sources. Liver[44], brain cerebellum[45], brain frontal cortex[45], brain temporal cortex[45], and brain pons[45] eQTLS were obtained using the GTEx (Genotype-Tissue Expression) eQTL Browser. For each defined fine-mapping region, this analysis searched for lead SNPs or SNPs with an $r^2>0.5$ in any 1000 Genomes Project population with a lead SNP falling in a *cis* eQTLS. The traits "fasting glucose-related traits," "fasting plasma glucose," and "type 2 diabetes" were used to define the phenotype.

1.5.2 Bioinformatics Results for Fine-mapped fasting glucose loci

To further explore fine-mapping results, a bioinformatics analysis was performed using the framework outlined in figure 1.3, which utilizes data from the UCSC Genome Browser and Haploreg (see table 1.5)[36, 37]. This analysis found that in the three loci where the index SNP was also the lead significant variant in HA and in the trans-ethnic analyses (*GCKR*, *MTNR1B,* and *G6PC2*), that the index SNPs were among the strongest functional annotations for the SNPs within these

signals. In addition, strong functional hypotheses for laboratory follow-up were generated in the *GCK*, *SLC2A2*, and *ADCY5* loci (see table 1.6).

Table 1.5 Description of bioinformatics tools used for functional follow-up of non-coding regions

Dataset	Genomic class	Description	Data source/program
1	Non-synonymous coding	Exonic positions wherein the variant would cause an amino acid replacement	dbSNP version 131
2	Promoter	1kb regions upstream of annotated transcription start sites	RefSeq
3	TFBS	Transcription factor binding sites (TFBS) predicted in promoter & non-promoter regulatory elements	UCSC Table Browser[46]: ChIPseq Transcription factor[4] PWM-scan[a] JASPAR, CONSITE, HaploReg
4	Non-coding RNA	All types of experimentally supported non-coding RNA, including microRNAs	RNAdb 2.0[47] & miRBase 17.0[48]
5	MicroRNA target site	Computationally predicted microRNA target sites within 3' UTRs	TargetScanS 5.2[49]
6	Enhancer element	Experimentally supported enhancer elements in any tissue	VISTA Enhancer Browser UCSC Table Browser[46]: ENCODE ChIP-seq Histone Modification[4]
7	Candidate non-specific regulatory element	Open chromatin loci in at least one human cell type, as assessed by DNase I hypersensitivity (DHS) mapping	UCSC Table Browser[46]: Duke and UW DNase I HS data from > 50 cell types[4]
8	Insulator elements	CTCF binding sites assessed by ChIP-seq technology	UCSC Table Browser[46]: ChIP-seq TFBS[4]
9	eQTL	Allele-specific differences in expression levels	GTEx eQTL Browser[50]
10	Conserved Element		UCSC Table Browser[46]: PhastCons 46-way conservation[51]
11	Splice Site		BDGP

[a]PWM-scan was applied using positional weight matrices (PWMs) from the Transfac database

Table 1.6 *Fine-mapped FPG loci with strongest functional candidates from bioinformatics analysis*

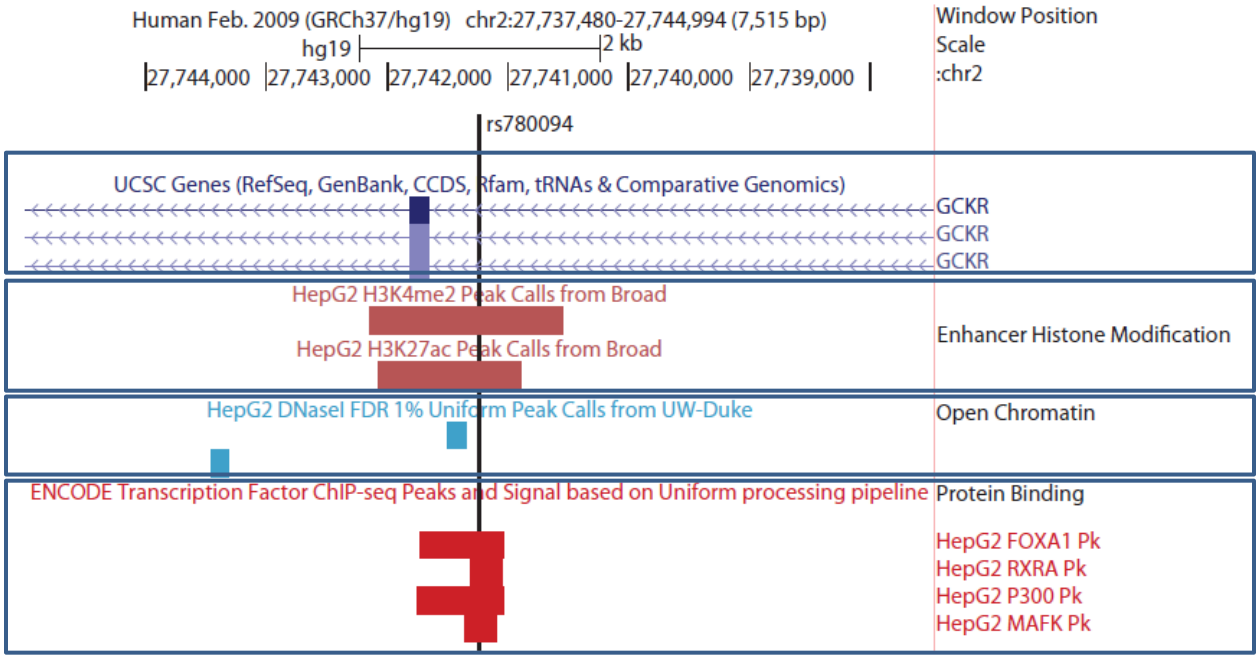
Locus/ Predicted regulated gene	Lead SNP	Strong functional candidate	r^2 with Index SNP EA/ASN/ AMR/AFR	Genomic location of functional candidate	# cell lines showing open chromatin	# cell lines showing histone modifications	Proteins bound	Altered motifs	Conservation (PhastCons)
11q14.3-*MTNR1B*	rs10830963	rs10830963	NA	92708710	4	2	FOXA2	TAL1	0
2p23.3-*GCKR*	rs1260326	rs780094	0.9/0.9/0.9/0.6	27741237	0	2, including liver carcinoma cells (HepG2)	FOXA2,MAFK,RXRA in liver carcinoma cells (HepG2)	Foxi1, MAFK	0
3q21.1-*ADCY5*	rs7614016	rs2877716	0.9/1/0.9/0.3	123094451	5	17, including liver carcinoma cells (HepG2)	CTCF, RAD21	DMRT1	0.15
3q26.2-*SLC2A2*	rs10513689	rs5398	0.4/<0.2/0.5/<0.2	170715830	6	16, including liver carcinoma cells (HepG2)	FOXA1, HDAC2, HNF4A, JUND, MAFK, P300, RXRA, TCF4	FOXA1, HDAC2,TCF4, P300	1
7p13-*GCK*	rs2908286	rs1004558	1/1/0.5/0.5	44234737	pancreatic duct cells	8, including liver carcinoma cells (HepG2)	GABP, HEY1, POL2 in HepG2 cells P300, CMYC, TR4, TBP, MAx, IRF1, NRSF, ELK4, in other cell lines	CTCF, GLI, HDAC2	0

The index variant for *GCKR*, rs1260326, was the lead variant in HA and in the trans-ethnic study. Bioinformatics annotation revealed that rs1260326 is a missense mutation, though PolyPhen2 prediction indicated the mutation is likely benign. Although disruption of protein coding is a likely candidate for having functional consequences, it does not preclude the potential for other non-coding regulatory variants to contribute the underlying cause of the association signal. In fact, functional analysis revealed an additional candidate, rs780094, which is located in the intronic region of *GCKR*. In a liver carcinoma cell line (HepG2), the region containing rs780094 was shown to bind three transcription factors (FOXA2, MAFK, and RXRA) and have histone modifications consistent with enhancer activity (see figure 1.4). Given that *GCKR* is produced in hepatocytes, enhancer histone marks HepG2 cells in this region suggest that rs780094 could disrupt transcription factor binding for regulation of *GCKR* expression and thus alter the expression of a key enzyme of glucose metabolism. Given the high LD between rs780094 and rs1260326 (r^2 ranging from 0.81 to 0.91 across study populations), distinguishing the potential functional effects of each variant will require laboratory follow-up.

In the *G6PC2* locus, the strongest association across populations was consistently rs560887. Bioinformatics annotation suggests that this variant is the most likely functional candidate showing weak signal for promoter histone marks in normal pancreatic cells (from Roadmap). In the *MTNR1B* locus the index SNP (rs10830963) was also the lead significant variant across ethnicities, and bioinformatics analysis suggests that rs10830963 is the most likely functional SNP. The rs10830963 variant is located in the intronic region of *MTNR1B* that binds FOXA2 in HepG2 cells. However, there was only weak evidence for open chromatin structure and histone modification consistent with regulatory elements. Because variants on GWAS platforms were selected based on the number of variants they are highly correlated with and thus informative for, it is serendipitous for the index to be the actual causal variant. However, the index SNPs for these three loci reflect

some of the strongest associations and effects observed thus far for FPG[52] and in part this may be a reflection of their causal relationship with FPG.

Figure 1.4 *UCSC Genome Browser view of predicted functional SNP in the GCKR locus*



UCSC Genome Browser view of the fine-mapped *GCKR* region (chr2: 27,744,000-27,744,994). The first panel shows that rs780094 is located in an intronic region of *GCKR*. The remaining 3 panels show that rs780094 falls in a region with enhancer histone marks, but misses the DHS signal of open chromatin and that 4 transcription factors bind to the region in HepG2 cells, respectively.

Fine-mapping for the *SLC2A2* locus successfully reduced a broad European GWAS signal to just 11 variants that were highly significant in the trans-ethnic study. Bioinformatics analysis identified one strong functional candidate among the highly significant variants in the trans-ethnic results. The rs5398 variant is a synonymous mutation of *SLC2A2* that also appears to be located in a strong enhancer in the HepG2 cell line and normal pancreatic tissue. In HepG2 cells there are a number of proteins that bind the region containing rs5398. In addition to *in vitro* binding of FOXA1, TCF4, HNF4 and p300 transcription factors, motif analysis of this region revealed differences in the

binding affinities between the alternate alleles of rs5398 for each of these transcription factors (see figure 1.5). Thus, it is possible that rs5398 is the functional variant underlying the *SLC2A2* signal and laboratory follow-up is warranted for this polymorphism and others that were significant in the trans-ethnic signal.

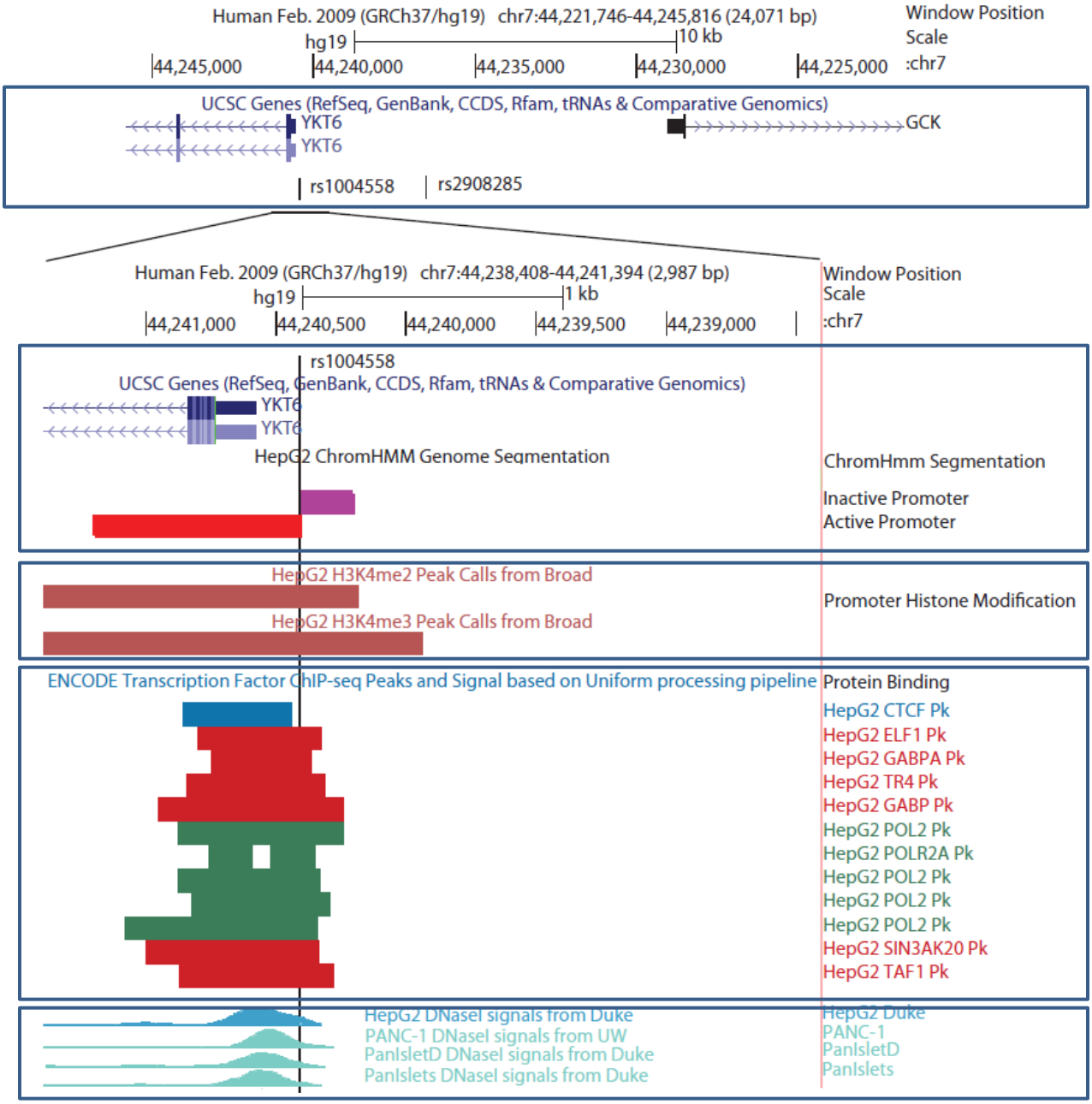
Figure 1.5 *UCSC Genome Browser view of predicted functional SNP in the SLC2A2 locus*

UCSC Genome Browser view of the fine-mapped *SLC2A2* region (chr3: 170,714,984-170,716,675). The first panel shows that rs5398 is located in a coding region of SLC2A2. ChromHmm segmentation defines the chemical signatures in this region as reflecting enhancer activity. The remaining three panels show that rs5398 falls in a region with enhancer histone marks, open chromatin structure and transcription factor binding in HepG2 cells, respectively.

Like *SLC2A2*, in the *GCK* region the number of highly significant and correlated variants was higher in the HA population than in the previous GWAS results. In contrast, the trans-ethnic results did not substantially reduce the number of highly significant variants or the physical genomic region. The most significant variant also varied across populations. Bioinformatics analysis suggested that the strongest functional candidate was rs1004558. This variant is in moderate LD with each of the lead variants across 1KG study populations (average r^2=0.5 with rs2908286). The rs1004558 variant is located in a region that binds GABP, HEY1, POL2, and TR4 in HepG2 cell line and also contains weak promoter histone modification marks across many cell types including HepG2 (see figure 1.6)

The lead variant in the *DGKB* region also differed substantially across populations. In Europeans there were 29 variants that were genome-wide significant and in high LD (r^2>0.4) with the lead variants (rs1558317). In HA, the number of variants that were highly significant and in LD (r^2>0.4) with the lead (rs1002061) was reduced to 18. In the trans-ethnic results this number was further reduced to 15. Additionally, the physical genomic region of the signal was also reduced from ~210Kb to ~2Kb. Despite the success of fine-mapping in this region, bioinformatics annotation was unable to identify any strong functional candidates.

Figure 1.6 *UCSC Genome Browser view of predicted functional SNP in the GCK locus*



UCSC Genome Browser view of the *GCK* fine-mapping region. The relative positions of the lead SNP (rs2908285) and the predicted functional SNP (rs1004558) in relation to GCK are shown in the top panel (ch7:44,221,746-44,245,816). The region containing rs1004558 was magnified in the lower panels to visualize various functional datasets (chr7:44,238,408-44,241,394). The second panel shows that the ChromHMM genome segmentation defines the chemical signatures in this region as reflecting a boundary between active and inactive promoters. The third, fourth, and fifth panels show evidence of promoter histone modification, protein binding regions, and open chromatin in liver carcinoma cells (HepG2), respectively.

In the *ADCY5* region there were no highly significant variants in the results for HA from PAGE, and only 1 variant (rs11708067) reached genome-wide significance in Europeans. Alternatively, in the multi-ethnic results there were three significant results (rs11708067, rs11720108, and rs7614016). Of these three variants, the index variant (rs11708067) had the strongest functional annotation. Histone marks in normal pancreatic cells and open chromatin structure (DNAse I hypersensitivity) suggest that rs11708076 is located in a weak enhancer region. However, ENCODE assays for various transcription factors did not illuminate any potential perturbation in protein binding. In addition, bioinformatics analysis revealed that rs2877716 was a strong functional candidate in high LD with rs11708067 across populations (r^2 of 0.8 EUR from 1KG to 1 in all other 1KG study populations). This variant was also highly significant in the trans-ethnic results ($P=5.8 \times 10^{-5}$), and in a fine-mapping study of Europeans on the MetaboChip by Scott et al. ($P=3.95 \times 10^{-13}$)[53]. The predicted functional variant, rs2877716, was also significant at the replication threshold (P= 0.003) in HA from PAGE. Bioinformatics annotation revealed that rs2877716 is located in a region that binds the cohesion protein CTCF and the transcription factor RAD21. Saxena et al. found that the C allele of rs2877716 was significantly associated with 2-hour glucose levels ($P=4.2 \times 10^{-16}$ and with T2DM (odds ratio=1.12, 95% CI 1.09-1.15, $P=4.8 \times 10^{-18}$)[54]. The adenylate cyclases have been implicated in the GLP-1 and GIP pathways dependent on cAMP-induced insulin release by beta-cells [55, 56] and disrupting regulation of *APCY5* through the alteration of a transcription factor binding site (such as rs2877716) could explain the underlying biology of this association. However, the polymorphisms in *APCY5* have not yet been tested in the laboratory. Thus, trans-ethnic meta-analysis confirmed a likely functional candidate in the *APCY5* that was previously only identified in replication studies.

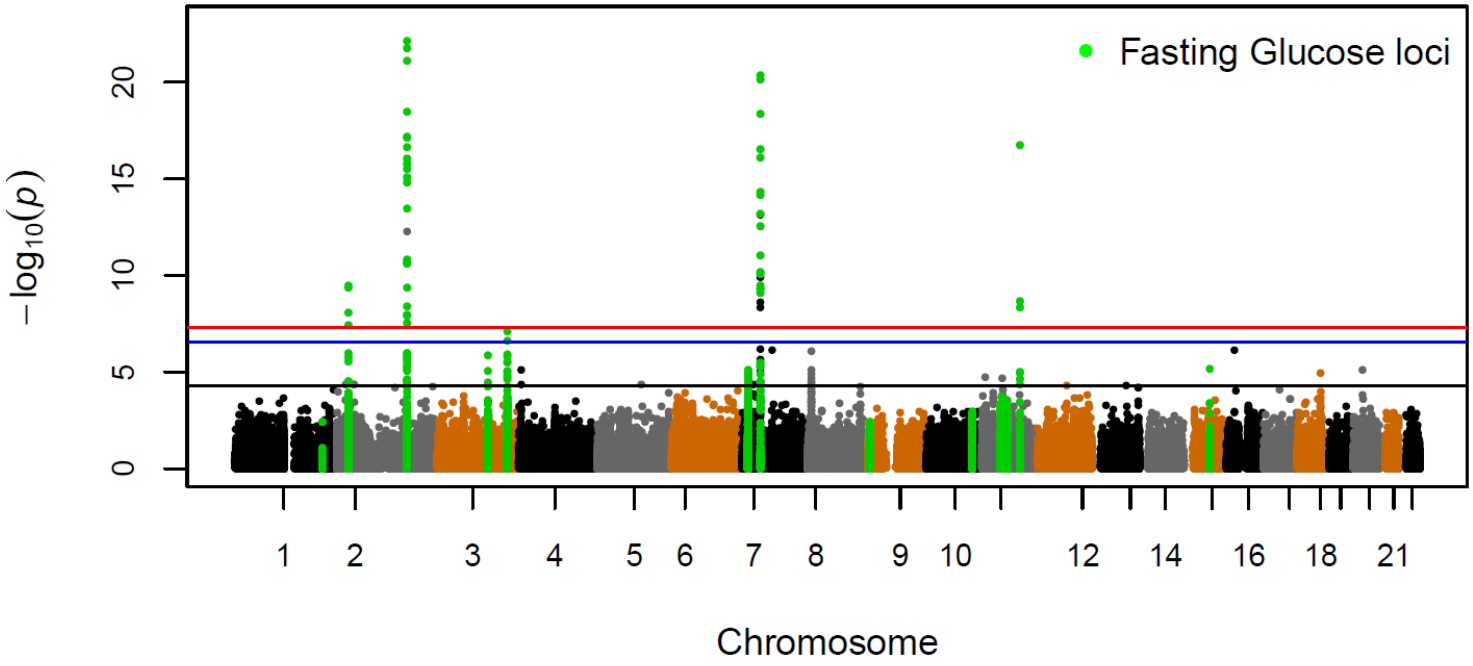
Of the 11 loci that were taken forward for eQTL analysis, only one SNP, rs10739037 ($P_{eQTL}=1.27 \times 10^{-12}$) showed association with expression levels of *GLIS3* in liver tissue. However, the

lead SNP, rs7034200 was not in LD (r^2>0.2) with rs10739037 in any 1000 Genome Populations. For the remaining brain tissues, no *cis* effect was observed for any variants in the fine-mapping regions.

1.6 METABOCHIP-WIDE ANALYSIS OF FASTING GLUCOSE LOCI

The MetaboChip-wide analysis (excluding the SNPs in the 14 FPG loci) identified five SNPs in two loci (2q31.1-*G6PC2* and 7p13-*GCK*) associated with FPG in trans-ethnic analysis (see figure 1.7). Rs477224 reached MetaboChip-wide significance (5.14×10^{-13}) without significant heterogeneity across populations in the trans-ethnic results. This variant is located approximately 7kb upstream of *G6PC2* and is not in LD with the significantly associated variants in the defined fine-mapping region encompassing *G6PC2*. However, a previous association (rs1402837; $P=6.8 \times 10^{-10}$) [57] in the promoter region of *G6PC2* was weakly correlated with rs477224 (r^2=0.21 in 1KG EUR), but this LD was negligible in all populations except those of European descent. As such, this region could be a novel association, although conditional analysis including rs1402837 (two SNPs included in the model) is necessary to confirm this. Bioinformatics follow-up did not reveal a strong functional candidate.

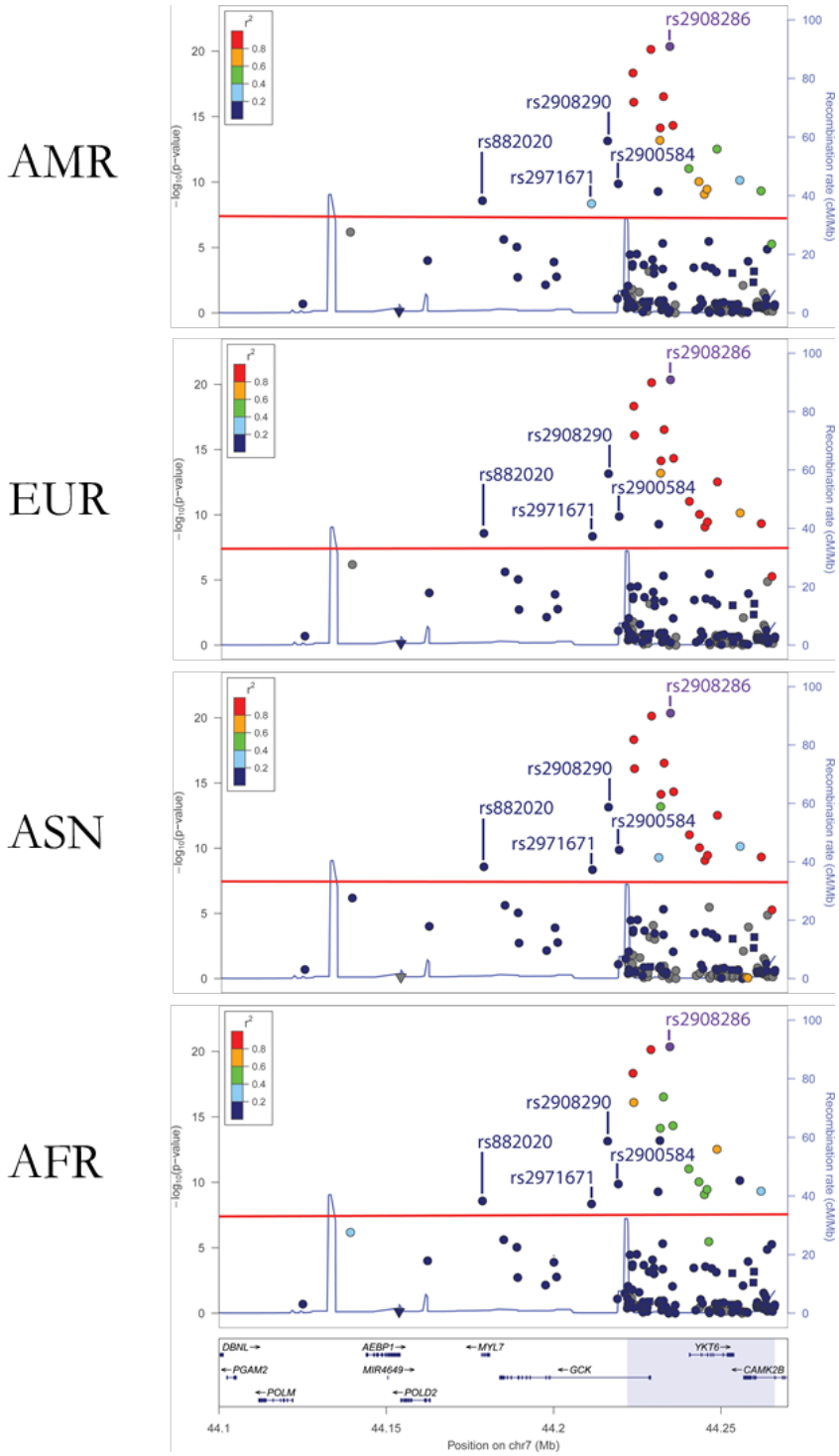
Figure 1.7 *Manhattan Plot for trans-ethnic meta-analysis MetaboChip-wide results*



Variants reaching the MetaboChip-wide threshold that are not located in the defined fine-mapping regions are in close proximity to regions previously associated with FPG. Five SNPs in two loci (1 SNP in 2q31.1-*G6PC2* and four SNPs 7p13-*GCK*) were identified. The rs477224 polymorphism is located 7kb upstream of *G6PC2* and is not in LD with the significantly associated variants in the defined fine-mapping region encompassing *G6PC2*. In the region near *GCK,* four SNPs reached MetaboChip-wide significance (rs2908290, rs2300584, rs2971671, and rs882020). Three of these variants are in the same LD block (rs2908290, rs2300584, and rs2971671, with r^2 ranging from 0.4-1 across populations). In contrast, rs882020 appears to be uncorrelated with these three variants. Like rs477224 near *G6PC2*, these variants were not in LD with previously identified loci and were not located in defined fine-mapping regions.

In the region near *GCK,* four SNPs reached MetaboChip-wide significance (rs2908290, $P=7.08 \times 10^{-4}$; rs2300584, $P=1.34 \times 10^{-10}$; rs2971671, $P=4.44 \times 10^{-9}$; and rs882020, $P=2.63 \times 10^{-9}$). Three of these variants are in the same LD block (rs2908290, rs2300584, and rs2971671, with r^2 ranging from 0.4-1 across populations). In contrast, rs882020 does not appear to be correlated with these three variants. Like rs477224 in the near-*G6PC2* region, these variants were not in LD with previously identified loci and were not located in the defined fine-mapping regions. Although these variants appear to be independent, given their physical proximity to known loci, future conditional analysis is warranted to confirm that the associations are truly independent (see figure 1.8).

Figure 1.8 Regional plots of the trans-ethnic results using LD in four 1000 Genomes Project populations for potential novel loci in the GCK region



Shown to the left are the two potential novel loci in the *GCK* region (rs2908290, rs2900584, and rs2971671 belong to the same haplotype block, and rs882020 is not correlated with these variants). LD between each variant in the trans-ethnic results and the lead variant in the *GCK* fine-mapping region, rs2908286, is shown for the respective 1000 Genomes Project Phase I study populations. The fine mapping region is highlighted in blue in the lowest panel.

Bioinformatics analysis suggested that the strongest functional candidate for the rs2908290 LD block is rs2908274. The rs2908274 variant is in moderate LD with these variants and located in a region of open chromatin structure with evidence of promoter histone modification and binds the transcription factors MAX, NRF1, ZNF263. The strongest functional candidate from rs882020 is an intronic variant in *GCK* (rs76323047) that is in moderate LD with rs882020. This variant is located in a region of open chromatin and binds EGR1, HEY1 and POL2 in a leukemia cell line (K562).

1.7 DISCUSSION

Patterns of LD and allele frequencies, referred to as genetic architecture, differ substantially between ethnic groups, and on average, European-ancestry populations have more highly correlated SNPs and extended haplotypes in comparison to populations of African or Asian ancestry. Hispanic Americans represent an admixed population with highly variable contributions of African, European and New World ancestry. As such, leveraging these differences across and within populations is a powerful approach to localize causal variants and discover novel associations that may have previously been undetectable in studies restricted to one particular ethnicity.

In this large multi-ethnic study population of more than 32,000 participants (comprised primarily of Hispanic and African American populations), high-density genotyping on the MetaboChip platform was used to fine-map 14 known FPG loci by leveraging differences in genomic architecture across populations. This study replicated 11 regions and found that in only three of the 11 loci the index SNP was consistently the most significant variant across populations. In five of the remaining eight loci, where the lead SNP was not the same as the index, the lead variant was in moderate or high LD with the original index SNP. In the remaining three replicated loci, although both the lead and index variants reached a nominal significance threshold, there was

not sufficient evidence to suggest the lead variant was an independent signal. In two of the three regions in which the lead variant was not correlated with the index, the lead SNP was very rare in all but one ethnic group. Thus, it is likely that this study was underpowered to detect a potentially novel locus in this study, particularly when the SNP was largely restricted to populations of Asian ancestry.

In addition to confirming previous fine-mapping findings, several strong novel functional candidates were proposed that can help guide future functional studies. The bioinformatics framework developed in this study suggested strong functional candidates in *ADCY5* (rs2877716), *SLC2A2* (rs5398) and *GCK* (rs1004558) that have not previously been reported. Each of these loci was shown to bind known transcription factors in liver carcinoma cells (HepG2) and have biochemical signatures consistent with regulatory activity. The rs5398 variant in the *SLC2A2* is of particular note. Although it is a synonymous variant the bioinformatics analysis provided strong evidence of a 3' enhancer in the region harboring rs5398. In addition to ChIP-seq evidence that TCF4, P300, and FOXA1 bind to this region, motif analysis predicted that the binding affinity was stronger for the alternate allele. A gain of function polymorphism for *SLC2A2*, which encodes a glucose transporter, is consistent with the estimated effect of -0.03 (± 0.004) mmol/L. Furthermore, variants in TCF4 are among the strongest susceptibility loci for T2DM [58]. This analysis also implicated rs780094 in *GCKR*, and rs10830963 in *MTNR1B* as likely functional candidates and each of these variants have been implicated in risk for T2DM[59, 60].

Particular strengths of this study include the large study population, high-density genotyping and representation of multiple ethnic groups. However, there were several limitations that should be noted. Despite the large size of this study, limited statistical power in Asian Americans may account for the inability to replicate variants at known FPG susceptibility loci, particularly for variants with weaker effects. Furthermore, although this study included populations from four major ethnic groups, the greatest proportion of participants were Hispanic and African American. Given that

there may be population-specific effects as a result of gene-gene, gene-environment, and divergent allele frequencies across populations, this study was limited in its ability to detect associations more prominent in Asian populations[52, 61]. Lastly, fine-mapping approaches only serve as an initial step in determining the underlying causal variant(s) driving association signals by prioritizing likely causal candidates for more onerous laboratory follow-up. To further meet this objective, functional candidates were identified using bioinformatics databases. However, given that the functional evidence detected by these datasets is incomplete, future functional studies in the laboratory are critical in determining the underlying causal variants[42]. That being said, the combination of fine-mapping with bioinformatics data is particularly useful for reducing both the physical genomic regions of interest and prioritizing candidates for molecular characterization. As such, fine-mapping is an essential step in functional interpretation of GWAS signals because laboratory follow-up of all possible variants in GWAS loci is prohibitively expensive and time-intensive.

To further refine the regions associated with fasting glucose, future work should include conditional analyses for each of the seven loci reaching the significance threshold for discovery of secondary signals. For each region, models should incorporate the lead SNP in the respective region. If after adjustment for the lead SNP, the secondary SNPs of interest remain significant it suggests that an independent signal related to a different functional variant is present in the region.

Given the critical roles of *GCK, GCKR, G6PC2,* and *MTNR1B* on glucose metabolism and insulin secretion[60, 62, 63], FPG-associated variants within these loci have been subject to substantial investigation. Numerous studies have extended the association between these variants and FPG to a role in susceptibility to T2DM. These efforts have revealed that the *GCK*-rs1799884 variant is associated with impaired glucose regulation[64], that *GCKR*-rs780094, and *G6PC2*-rs560887 variants are associated with modest increases in risk of T2DM[59, 65], and a large prospective study has shown that the risk allele of *MTNR1B*-rs10830963 is predictive of T2DM [60]. Recently, the rs560887-G allele

was shown to be a functional variant underlying the *G6PC2* signal via enhanced pre-mRNA splicing[66]. Furthermore, rs2232316, which was weakly correlated with the putative independent signal detected in this study, rs477224 (r^2=0.23 in ASN from 1KG) was shown to increase the binding of the transcription factor Foxa2 and thereby increase expression of *G6PC2* [66].

This trans-ethnic study comprehensively fine-mapped known common variants associated with concentrations of FPG. Genomic regions harboring known risk variants were refined, novel functional candidates were proposed and new independent signals in previously FPG-implicated genes were identified. Thus, these results suggest that trans-ethnic meta-analysis can help in translating GWAS results into new biological insight.

CHAPTER 2

FlowSeq: An *in vitro* method to test regulatory effects of genetic variants

2.1 ABSTRACT

Translating genetic association results into novel therapeutics and prevention strategies is challenging for the large fraction of identified variants without an obvious function. Characterization of these variants by interrogating the putative functional alleles through traditional transient expression assays is expensive, time-intensive and highly sensitive to minor discrepancies in laboratory procedure. This study developed a reporter assay that couples fluorescence activated cell sorting (FACS) with next generation sequencing to screen candidate variants for function. This strategy allows simultaneous testing of multiple variants and quantification of their relative contribution to altered levels of reporter protein expression. As proof of principle, this method was applied to analyze a poly-A tail variant (rs6732914) in the glucagon gene that showed evidence of altered mRNA degradation in a traditional cell culture assay. This novel method successfully reproduced the expected allelic difference in gene expression associated with post-transcriptional regulation. This assay provides a rapid and sensitive approach to quantify relative effects of regulatory sequence variation and could provide the necessary insight to translate information from genome-wide association studies (GWAS) into target genes for therapeutic intervention.

2.2 INTRODUCTION

Much effort has been expended on the identification of genetic variants associated with common diseases and traits. The majority of variants identified through genome-wide association studies (GWAS) involve variation outside of coding regions [67], and it is hypothesized that variants associated with complex diseases exert their effects through regulation of gene expression [39]. However, interpreting the biological mechanisms underlying association signals is difficult without conducting additional laboratory experiments. Despite the large number of statistical associations that have flooded the literature over the last ten years [68], there has been little translation from statistical association to the identification of functional variants. Compared to the identification of genetic variants, laboratory characterization of functional variants remains expensive and time consuming. Given that each disease-associated polymorphism is a proxy for a set of correlated variants [69], comprehensive laboratory follow-up through the standard variant-by-variant approach is impractical. As such, there exists great need for the development of a more efficient laboratory assays to assess functional consequences of genetic variations.

Before testing phenotypic consequences of common risk alleles it is necessary to identify candidates that are most likely to influence a regulatory element from a set of correlated variants underlying a GWAS signal. Prioritizing variants can be accomplished by conducting one of the emerging study designs to fine-map GWAS signals [70, 71] or through the bioinformatics annotation of variants with publicly available regulatory datasets [39]. After selecting candidates for laboratory follow-up, identifying an appropriate model to test a particular functional hypothesis is critical. The mechanisms influencing gene expression are often organism-specific and can vary temporally, between tissues, and even between cell types within a tissue [72, 73]. Thus, selecting an inappropriate model system for functional characterization can produce false negative results. The ideal model for testing susceptibility variants is often unavailable because untransformed cells are typically unable to

proliferate *in vitro*. In practice, cell lines derived from a specific tissue or tumor tissue offer the best models because, unlike other model systems, these cells are more likely to express the regulatory factors necessary to reproduce differential expression patterns.

GWAS results introduced two important considerations for functional follow-up: (1) the effects of common variation on complex disease are modest when studied individually, and (2) the lack of protein-coding associations at most loci suggests that the causal variant(s) in the region influence gene expression [1]. Although transient transfection of allelic reporter constructs in cell culture-based assays has long been used to measure large effects resulting from sequence variation, the robust detection of smaller differences is hindered by the background noise inherent to temporary episomal expression, largely introduced by variable efficiencies of transfection. In traditional approaches, the same reporter gene is used for each allelic plasmid, so the test plasmids must be transfected separately from one another. In order to control for differential transfection efficiencies, each test allele plasmid is independently cotransfected with and quantitated against, a control reporter gene (typically renilla or red fluorescent protein). Extending this method to test GWAS risk alleles assumes it is possible to quantitate test plasmids relative to control plasmids with enough precision to detect relatively modest changes in fluorescence or mRNA abundance. In addition, there are several nuisance parameters that can bias results, including the number of plasmids transfected into each cell and loss of free plasmids as cells divide. The permanent introduction of allelic DNA into the cellular genome through stable transfection can address these nuisance parameters, but the procedure is much slower and results can be confounded by regulatory elements present in the native cellular DNA near the insertion site. Thus, application of transfection methods for detecting small effect sizes associated with GWAS loci will need to reduce background noise in order to detect subtle changes in transcription while allowing more rapid interrogation of risk alleles.

In addition to balancing precision and efficiency, testing regulatory variants requires a versatile method capable of evaluating diverse regulatory mechanisms. Gene expression is influenced by a dynamic interplay of transcriptional activity and post-transcriptional regulation of mRNA abundance [74, 75]. Although fluorescent reporter assays can be used to measure both transcript abundance and fluorescent intensity of the reporter, only the latter will fully capture differences in translational efficiency between regulatory variants. Given the diversity of regulatory mechanisms, the expression plasmid used for testing risk alleles should ideally be adaptable to *cis*-regulatory elements (promoters), distal regulatory elements (enhancers), and regions that influence the stability or translational efficiency of mRNA, such as elements in the 3' and 5' untranslated regions (UTR). Thus, the design of a comprehensive method to test regulatory variation will ideally consider both the complexity of the underlying biology and the measurement used to quantify differences in gene expression.

This study aimed to develop a reporter assay that could be adapted to test many types of regulatory mechanisms and detect small differences in expression. To this end, several high throughput technologies were coupled to enable rapid and precise interrogation of regulatory variants. By combining fluorescence activated cell sorting with next generation sequencing, this novel method reproduced the detection of a known functional polymorphism, provided a more efficient approach to screen multiple variants in a region and reliably quantified their relative effects in gene expression.

2.3 METHODS

2.3.1 Allelic Plasmid Construction

Through a traditional luciferase assay, the presence of a functional polymorphism (rs6732914) was detected in 3' end of the glucagon (*GCG*) transcript. This polymorphism alters the cleavage site where the immature mRNA is cleaved to add the poly-A tail. In transient transfection experiments it demonstrated that the A allele exhibited higher expression levels than the G allele, which may be due to differential mRNA turnover[i] . To test this approach, a chimeric gene was constructed containing the green fluorescent protein (GFP) coding sequence from pEGFP (Clontech, Genbank CVU76561) fused to the *GCG* 3' UTR. The chimeric gene was cloned into the pCI-neo plasmid (Promega part #TB215) multiple cloning site, driven by the CMV viral promoter. Standard altered sites II mammalian site-directed mutagenesis protocols were used (Promega) to (a) alter the allele at the rs6732914 SNP, (b) remove the SV40 poly-A signal that lies 3' from the multiple cloning site so that it would not out-compete the *GCG* poly-A signal, and (c) repair the Amp gene, thereby restoring Amp resistance. The chloramphenicol acetyltransferase (*CAT*) gene, driven by the SV40 promoter, was not disabled because this experiment required only a single round of mutagenesis to generate the construct of interest; therefore, *CAT* was used as an internal control. The neomycin phosphotransferase selectable marker was used to select cells with integrated plasmid after incubation in media with Invitrogen Geneticin selectiveantibiotic (G418). These two allelic constructs (pGCG-Gdel and pGCG-Adel) were sequence confirmed using primers spaced approximately every 500 bp around the plasmid, and differ only at the rs6732914 SNP.

2.3.2 FlowSeq Reporter Assay

This study developed a novel reporter assay that combines flow cytometry and next generation sequencing. As described above, constructs were designed so that the cytomegalovirus (*CMV*) early promoter drives expression of a fusion transcript containing the GFP coding sequence spliced to allelic variants of the *GCG* 3'UTR and poly-A signal. After allelic plasmids were constructed and quantitated, an equimolar mixture of the two allelic plasmids was transfected into human hepatoma PLC-PRF-5 (PLC) cells. After selection for stably integrated plasmids, the cells were trypsin digested off plates, and sorted into quartile samples on the basis of GFP expression using FACS. The relative frequency of each allelic construct was then measured for each quartile by high throughput sequencing.

To ensure that expression was measured in cells with integrated plasmid, cells were selected for using the Neomycin Phosphotransferase Selectable Marker [Promega]. Because stable expression of foreign DNA is subject to local regulatory influences from the cells' neighboring chromosomal sequence, random selection of single clones could bias or attenuate our measurement of transgene expression. To address this, analyses were conducted across entire plates with the intention of averaging these local effects across large numbers of insertion sites for each allele.

PLC cells were first transferred from Invitrogen's Dulbecco's Modified Eagle's Medium (DMEM) to 100mm plate and allowed to grow to 85-90% confluence. Cells were then trypsinized according to Invitrogen's protocol and placed into a 1.5µl tube containing 2.5µg of *GCG*-G, 2.5µg of *GCG*-A allelic plasmid DNA (total of 5 µg plasmid DNA), and 100µl of Nucleofactor solution. From this solution, 2.5 million cells were collected and transferred into Amaxa cuvettes with 0.81 µl of Nucleofactor solution. Cuvettes were then placed into an electroporator, allowing introduction of either *GCG*-A or *GCG*-G allelic plasmid into the cells. These cells were then transferred onto two

separate warm 60mm plates and incubated at 37°C. After 48 hours of incubation cells were trypsinized from the plates and washed with 3 ml of phosphate buffered saline (PBS). Cells were then split from each plate and transferred onto four 60mm plates with media and the selective antibiotic for Neomycin, G418. This process was repeated every two days over the course of two weeks. After this incubation period, cells were trypsinized from the plates and washed with 3 ml of PBS. Trypsin was neutralized with 2ml of media before transferring cells into 15ml tubes. Cells then underwent centrifugation with 3,500 rcf spins for 1 minute. Pellets were filtered and resuspended in flow cytometry tubes (Falcon) containing 2% PBS solution and a viable cell marker, Propidium Iodide (PI). In addition to the cotransfection of each allele, cells were independently transfected cells with allelic plasmid A or G using the same methods outlined above. As a positive control transgene expression of GFP reporter gene was driven by SV40 strong promoter, and SV40 polyA tail, again following the same methods as in the sample.

2.3.3 Fluorescence-activated cell sorting (FACS)

FACS was used to divide single, viable PLC/PRF cells into four quartiles of the log transformed GFP distribution, reported as units of fluorescein isothiocyanate (FITC). Samples were acquired and sorted using FACSDiva software (BD Biosciences) on a three-laser (488, 633, and 405 nm) FACSAria flow cytometer (BD Biosciences) with the standard optical configurations. Configuration of the cytometer was done prior to the experiment using the cytometer setup and tracking application in the BD FACSDiva software Version 6.1.3 according to the manufacturer's instructions. Immediately before acquisition, the cells were filtered through a 35-μm filter (Cell Strainer, BD Biosciences) to remove cell clumps. Before the flow sorting, a total of 30,000 events, or cell reads, in each sample were acquired to define the gating regions. Viable cells were defined as those expressing less than 103 PI units. Proxies for the physical attributes of the cells were used to

select single, living cells. The forward scatter channel (FSC) measuring light excitation in the column of the suspended cell corresponds to particle size. Light measured approximately perpendicular to the excitation line, known as side scatter (SSC), provides information about the granular content within a particle. A combination of these two measurements was used to differentiate healthy single cells from cellular debris, apoptotic cells and clumps of cells. For instance, subcellular debris and clumps of cells will have lower FSC and SSC measurements, while clumped cells will show higher FSC and SSC. For each quartile, a total of 75,000 to 100,000 cells were sorted directly into 400-µL of media solution and stored until quantitative polymerase chain reaction (PCR) in -80°C. Total sorting time was 30 minutes (median time per sample, 15 minutes). Post data analysis of the GFP distribution in each sample was conducted in FlowJo version 7.6 software (Ashland, OR) and in R statistical computing software v2.12.2.

2.3.4 MiSeq Library Sequencing for Allelic Quantitation

DNA was extracted from each sample using Qiagen, DNeasy Blood and Tissue kit (catalogue number 69504). Stably integrated plasmid sequence was amplified by PCR from each sample. Primers were modified and designed with NEXTERA adapters to guarantee compatibility with Nextera **XT DNA Sample Preparation Kit (catalogue number FC-131-1024)**. The modified primers were as follows:

GCG_FW2A:GCCTCCCTCGCGCCATCAGTGGCACTTCATCCAGCACAAGCTG

GCG_FW2B:GCCTTGCCAGCCCGCTCAGTGGCACTTCATCCAGCACAAGCTG

GCG_RV2A:GCCTCCCTCGCGCCATCAGCGCCATTCAGGCTGCGCAACTGTT

GCG_RV2B:GCCTTGCCAGCCCGCTCAGCGCCATTCAGGCTGCGCAACTGTT

Using these modified primers DNA was amplified from the samples. PCR products were converted to Illumina libraries using standard NEXTERA protocols (Illumina), to produce

multiplexed paired-end libraries with unique indexes from Illumina to facilitate multiplex sequencing of libraries within a single MiSeq run. Samples were sequenced on the Illumina MiSeq platform generating paired-end read chemistry with 2 x 150 bp sequencing at each end of the template.

2.3.5 Statistical Analysis

Raw Illumina fastq files were processed and analysed through standard assembly pipeline. Fastq files were de-multiplexed, quality filters applied and analyzed using Fastx Toolkit (v0.0.13). The quantity of allelic genomic sequence in each of the eight samples (corresponding to the four quartiles of gene expression in two replicate plates of a single cotransfection event) was ascertained by searching for 13-mer nucleotide sequences corresponding to the forward and reverse allelic inserts ('CAATGGtAAAGAA' , 'CAATGGcAAAGAA,' 'TTCTTaCCATTG', 'TTCTTgCCATTG') using the grep utility at the command line on the demultiplexed files. Alignment of the sequences was visually inspected in Integrative Genomics Viewer (IGV 2.1.x) for quality assurance.

This analysis tested for statistically significant log fold changes between cotransfected allelic plasmids (A and G) using a Poisson Generalized Linear Model in R statistical environment (v2.12.2). Blinded modelling of this method with synthesized data revealed that this approach could be used to simultaneously test at least 24 allelic plasmids.

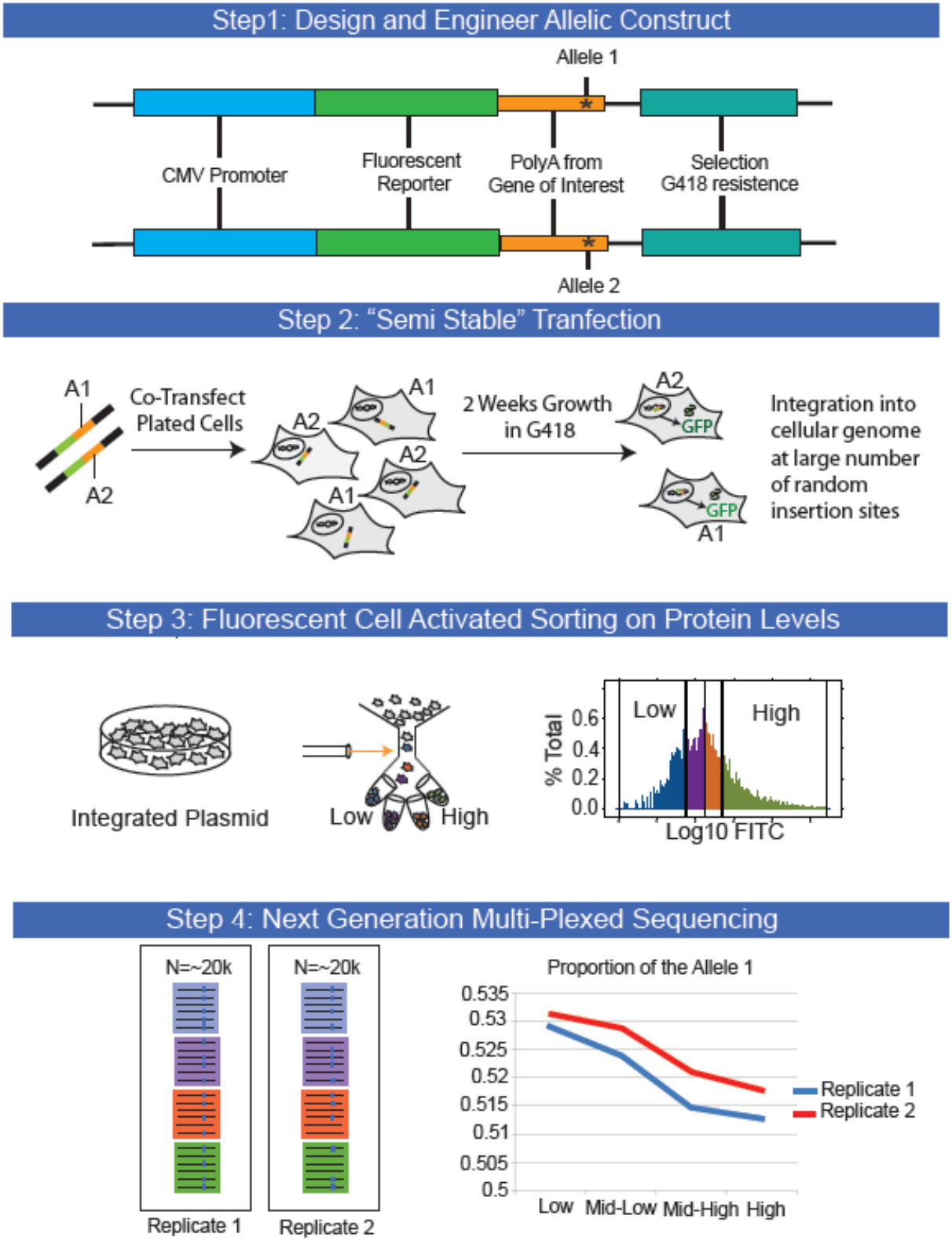
2.4 RESULTS

2.4.1 Description of the Method

The aim of this study was to develop a method that, in a single experiment, could test many potential regulatory variants associated with a complex disease or trait and quantify the relative effects of each allele on gene expression. The work flow for this procedure is shown in figure 2.1.

First, bioinformatics tools are used to prioritize strong functional candidates and identify the appropriate cell line to test the corresponding hypothesis. For this experiment, a 3'UTR variant in the glucagon gene was tested using the strong cytomeglovirus (CMV) promoter to drive expression of our fluorescent gene. This experiment was conducted in PLC/PRF cells (a hepatoma derived cell line) because polyadenylation is not known to be a cell type specific, and because pancreas beta cell derived lines are difficult to culture and transfect. The second step in this method is to 'semi-stably' cotransfect cells and then select such that each surviving cell has integrated an allelic plasmid into the genome. Plasmids (pGCG-A_{del} and pGCG-G_{del} were cotransfected into approximately 200,000 PLC/PRF cells at an equimolar ratio. Cells were selected for integrated plasmid using G418 over the course of two weeks. Since clones with a single integration site were not selected, this experiment was faster than traditional stable transfection approaches and averaged the effects of the cells' regulatory elements across thousands of cells. In the third step, entire plates of cells were sorted into quartiles based on CMV driven green fluorescent protein (GFP) expression, in which mRNA turnover was altered by an allelic polyA tail. In the last step of this experiment DNA was extracted from each quartile sample, barcoded and multiplex sequenced on the MiSeq platform using a single lane. After removing low-quality reads, the proportion of each allele present in a phenotypic pool was calculated.

Figure 2.1 *FlowSeq Experiment Pipeline*



2.4.2 Fluorescent Activated Cell Sorting

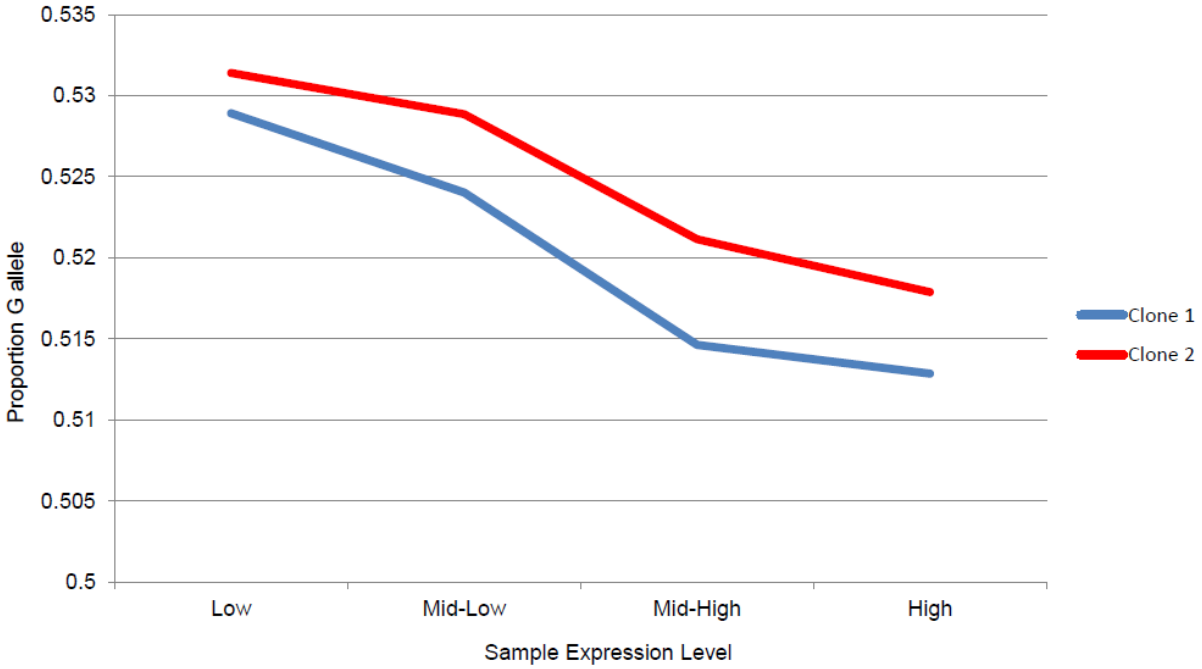
This study aimed to develop a reporter assay that could be adapted to test many types of regulatory mechanisms and detect minor differences in expression. Use of fluorescent protein allowed interpretation of the impact of alleles on protein expression rather than the more intermediate phenotype of mRNA abundance. It was hypothesized that some regulatory variants may reduce expression levels close to that of control cells that were not transfected with fluorescent protein. The design of the probes during the sequencing preparation ensured that analysis was based solely on engineered allelic DNA. As such, a negative control was not necessary and absent from this experiment in order to avoid the inadvertent removal of transfected cells expressing fluorescent levels close to that of an untransfected control. However, because fluorescence is impacted by cellular attributes or number of cells measured, propidium iodide (PI) was used as a viable marker to select living cells and light scatter to select for measurement of fluorescence in a single cell. For gating approximately 30,000 'events' or fluorescent measurements passed to FACSDiva (28,844 and 37,094 for clone 1 and 2 respectively). Of those there were approximately 30% (10,754 and 11,527 for clone 1 and 2 respectively) that passed the light scatter gating indicative of single intact cells. Using PI to isolate living cells from dead another 10% were removed. In total, 343,855 cells from clone 1 and 308,813 cells for clone 2 were sorted into four quartiles of reporter protein expression allowing comparison of the phenotypic effects in each individual cell from 3'UTR alleles across a very large number of events.

2.4.3 Statistical Analysis

Using the Armitage Test for trend in two replicate plates (each plate representing one transfection event) a statistically significant and reproducible negative association between the

proportion of G allele ($p<0.005$) and eGFP gene expression (Figure 2.2) was found. Using a standard normal distribution to approximate fold change relative to the standard deviation (=1) this analysis estimated an approximate 11% increase in mean expression from the A allele (mutant) over mean expression from the G allele (wildtype). These results were consistent with previous findings (Unpublished data). Reduced expression levels from the G allele relative to the reference A allele were confirmed using a reporter assay combining fluorescence activated cell sorting (FACs) with next-generation sequencing. An equimolar mixture of the A and G allelic plasmids were cotransfected into PLC/PRF cells. After two weeks of G418 selection for cells with at least one stably integrated plasmid into the cellular genome, FACs was used to separate viable, individual cells into quartiles on the basis of GFP expression. DNA was extracted and PCR amplified from each quartile sample. PCR products were converted to Illumina libraries using standard NEXTERA protocols (Illumina), resulting in multiplexed paired-end libraries each containing a unique index from Illumina. A single multiplex MiSeq run was performed to sequence all of the libraries. The proportion of reads matching the G allele was highest for cells in the lowest quartile of GFP expression, and the trend toward increasing A allele representation was monotonic and significant across all four samples. These single-cell results from stable transfection are concordant with the differential expression observed in the transient transfection results.

Figure 2.2 Results for glucagon 3'UTR regulatory variant (rs6732914)



Reduced expression levels from the G allele relative to the reference A allele were confirmed using a novel reporter assay combining fluorescence activated cell sorting (FACs) with next-generation sequencing. As opposed to traditional reporter assays this approach allows simultaneous transfection and subsequent assessment of allelic effects in multiple alleles. An equimolar mixture of the A and G allelic plasmids were co-transfected into PLC/PRF cells. After two weeks of G418 selection for cells with at least one stable integration of plasmid into the cellular genome, we used FACS to separate individual cells into quartiles on the basis of GFP expression. DNA was extracted and PCR amplified from each quartile sample. PCR products were converted to Illumina libraries using standard NEXTERA protocols (Illumina), resulting in multiplexed paired-end libraries each containing a unique index from Illumina. A single multiplex MiSeq run was performed to sequence all of the libraries. The proportion of reads matching the G allele was highest for cells in the lowest quartile of GFP expression, and the trend toward increasing A allele representation was monotonic and significant across all four quartiles. These single-cell results from stable transfection confirm the differential expression observed in the transient transfection results.

2.5 DISCUSSION

This study improved upon traditional reporter assays by combining several high throughput technologies to enable rapid and precise interrogation of regulatory variants. Similar to traditional reporter assays, this system involved the design and introduction of allelic expression plasmids into

an appropriate cell line, followed by fluorescence measurement as a reflection of gene expression. However, unlike traditional assays, all allelic plasmids were simultaneously transfected into cells of interest, thus removing bias from independent transfections and subsequent adjustment with reporter controls. Like other recent studies [76, 77] Fluorescent Activated Cell Sorting (FACS) was used to measure expression of the reporter protein in each individual cell and measured the relative allelic proportions in DNA prepared from cells sorted into quartiles of reporter expression using multiplexed high-throughput sequencing. Unlike previous methods, a hybrid of transient and stable transfection was used to select cells with stable integration of a plasmid, but at a very large number of integration sites (a method referred to here as "semi-stable" transfection, Figure 1). In doing so the portion of noise introduced by differential plasmid copy number per cell during transient transfection was eliminated, while averaging the effects from the cells' native regulatory elements across many integration sites for each allelic plasmid. In a previous study a traditional luciferase-based assay was used to demonstrate increased expression associated with the A allele of a poly-A tail variant in *GCG* [i]. In this experiment, the A and G alleles were cotransfected into a human hepatoma cell line (PLC/PRF) and cells with integrated plasmid were sorted into different phenotypic levels of expression for subsequent multiplexed sequencing. In comparison to traditional approaches, this novel method provided a more efficient approach to screening multiple variants in a region and to quantify their relative effects in gene expression.

2.5.1 Comparison to Other Recent Approaches

Over recent years, numerous studies have employed novel strategies to gain a global appreciation of the regulatory landscape in a variety of model systems. Although these strategies are higher throughput than traditional reporter assays, there remain significant limitations for testing specific human regulatory variants associated with common diseases. In several studies,

investigators attempted to identify the level of gene expression associated with large libraries of randomly mutated regulatory regions [77, 78]. Although this is useful for gaining a general appreciation of the sensitivity of enhancers [78] or post-transcriptional elements [77] to sequence variation, assessing phenotypic results of random variation (reverse genetics) is inefficient for interpreting the significance of specific risk alleles identified through association studies. The use of random mutagenesis restricts an investigator's control over the type of mutations that can be studied by preventing systematic introduction of insertions and deletions and removing the ability to design patterns of specific point mutations. Other programmable microarrays have been used to design mutant plasmids [76, 79] but these approaches are limited in the length of sequences that can be engineered [80]. Although site-directed mutagenesis remains relatively expensive and slow, the use of bioinformatics tools to select priority variants remains the most cost-effective solution for hypothesis-driven analysis of regulatory regions.

Recent studies have varied in their approaches for measuring phenotypic effects. Two recent studies have measured mRNA abundance rather than the levels of a reporter protein [78, 79]. Although this removes the need to use a reporter gene, it does not allow the quantification of phenotypic effects related to translational efficiency. Furthermore, these studies inserted a sequence tag into the 3'UTR to map back mRNAs to a library of mutations, which could significantly impact mRNA stability and translational efficiency between the alleles. Other studies have combined fluorescent activated cell sorting and multiplexed high throughput sequencing of the DNA to quantify the phenotypic effects of regulatory variants. However, one of these studies used random mutagenesis[77], while the other measured yellow fluorescent proteins in yeast [76]. Therefore, to my knowledge this is the first study to verify that combining FACs with high throughput sequencing can be used to reproduce known effects of a regulatory variant in a human cell line.

Although it has been argued that cell lines do not capture the milieu of regulatory influences from internal and external stimuli present in living systems, the use of model organisms such as transgenic mice is expensive and low throughput [78]. Furthermore tail vein injections are not cell type or tissue specific and do not necessarily reflect the mechanisms present in humans. Alternatively, yeast and bacteria are cheaper and more rapid models, but can only be used to test regulatory mechanisms that are conserved across kingdoms. In many circumstances this is an inappropriate assumption for human regulatory elements. Finally, each of these approaches used transient expression of the plasmid, which introduces too much interference to detect very modest effects.

2.5.2 Possible limitations of the method and additional clarifications

As a proof of principle, a variant thought to exert a moderate effect on post-transcriptional regulation was used. However, it is recognized that many of the regulatory effects discovered in GWAS have very modest effects in comparison. Therefore, when the effect size is suggested to be very small, it may be necessary to use a targeted insertion site in a region of the cellular genome with very little to no native regulatory elements. This will increase the precision of the measured difference by reducing the variance associated with a large number of local cellular regulatory effects at the very large number of random insertion sites. Creating a cell line with a single insertion (e.g. Flp-In method) takes longer, but the resulting cell line can be used to test a very large number of variants [81]. Therefore, this approach would still be higher-throughput and more robust than current strategies for detecting very small effects.

A second limitation of this strategy is that it does not allow analysis of the enhancer function in the context of the whole organism. It is acknowledged that cell lines are often transformed and are not in their normal physiological conditions, making it difficult to predict the effects of susceptibility loci. However, this limitation is true of all laboratory functional assays, and current

efforts to create more phenotypically 'normal' cell lines will continue to be of interest. Given restrictions on the potential size of the plasmid insert, it is also difficult to demonstrate how long distal enhancers may differ in their effects on gene expression. However, this is not seen as a major drawback to this method as it is presumed that long distance enhancers exert their effects by looping into the proximity of a promoter.

2.5.3 Conclusions and Future directions

FlowSeq is a novel reporter assay which combines fluorescent activated cell sorting with next generation sequencing technology to test a putative regulatory polymorphism for altered protein expression. This strategy provided several advantages over more traditional methods. First, the use of a semi-stable cotransfection averages the potential confounding effects of cellular regulatory elements across hundreds of thousands of cells. Although this approach increases the variation in the expression distribution across cells, the positive and negative effects should be balanced for each allele. Cotransfection of the alleles reduces the sensitivity of the assay to small errors in quantitation of allelic plasmid preparations. Second, quantifying relative differences in protein measured in each individual cell, in tandem with next generation sequencing, allowed a precise measurement of expression in thousands of individual cells rather than a single aggregate measure of expression in a population of cells. Third, this method could be applied to allow for many polymorphisms to be tested simultaneously, rather than one allele at a time, making this design higher throughput than previous approaches. Lastly, although this method tested a 3' UTR variant in GCG, this method could be adapted to test many other categories of regulatory elements, in addition to testing various combinations of sequence variations.

This study demonstrated that FlowSeq is a powerful method to detect relatively modest differences in expression, and can be modified to rapidly test several variants in a single experiment.

Adapting this strategy in future studies could help expedite functional discovery and biological interpretation of regulatory polymorphisms.

CHAPTER 3

Genotype-Driven Recruitment of Participants from the Database of Genotypes and Phenotypes (dbGaP): Analysis of Publicly Accessible Consent Documentation

3.1 ABSTRACT

Using individual-level genotype information from centralized data repositories to recruit participants that possess specific genetic variants could expedite the translation of genetic research results into functional knowledge about the human genome. However, uncertainty remains on whether genotype-driven recruitment would, in most cases, be permissible given the terms of informed consent under which participants were originally enrolled. This study performed a directed content analysis of 36 accessible consent documents in the database of Genotypes and Phenotypes (dbGaP) hosted by the National Center for Biotechnology Information (NCBI). The research studies represented by these documents included a diverse set of study designs, allowing the examination of many contexts in which GDR may occur. This analysis explored whether: (1) future contact was mentioned and if so for what purpose, (2) return of individual or aggregate research results was discussed and if so under what conditions, (3) processes for how re-contact or return of results would take place were described and if so, what details were provided, (4) re-contact or return of results were addressed in the discussion of study risks or benefits, and (5) participants were offered the opportunity to express preferences about future re-contact or return of individual genetic research results. The analysis identified considerable heterogeneity in the approach to re-contact and return of individual research results, suggesting that current research norms may hinder the widespread adoption of GDR as a recruitment strategy. These results challenge the feasibility of

GDR, at least for the majority of currently archived data. Future studies should, wherever possible, enroll participants under an expectation that GDR may occur.

3.2 INTRODUCTION

Considerable effort has been devoted to uncovering genetic variation that can be used to tailor prevention and treatment strategies to fit the particular needs of an individual. To date there are approximately 4,000 statistically significant ($p<=10^{-8}$) genetic associations with a disease or trait in the literature [68]. However, only a fraction of these will ultimately point to a functional variant that can be used to elucidate molecular mechanisms underlying disease susceptibility. Results from genome-wide association studies (GWAS) often implicate both functional variant(s) and many non-functional bystanders that are located near multiple genes without a clear relationship to the studied phenotype. Thus, uncovering the underlying biology of most GWAS findings remains a formidable challenge [42, 42].

To advance the pace of discovery, it has been recognized that sharing individual genetic and phenotypic data across studies is a particularly efficient approach for exploratory research and validation of existing findings. To facilitate data-sharing, the National Institutes of Health (NIH) developed the Database of Genotypes and Phenotypes (dbGaP) to provide a centralized repository for the secured storage of genetic data generated from federal research support. Recently, the NIH released a draft of a new Genomic Data Sharing Policy (http://grants.nih.gov/grants/guide/notice-files/NOT-OD-13-119.html). This draft describes the expectation of investigators to submit NIH-funded data to dbGaP and the controlled access usage of individual-level data. However, notably absent from the draft policy is guidance relevant to the return of individual research-generated genetic information. Also missing from the draft were guidelines on future contact of participants

from dbGaP to invite participation in additional studies. Given that the interpretation of disease associated variants is currently one of the most pressing issues in the post-GWAS period, the lack of guidance regarding these issues could become a major barrier for effective clinical translation.

Genotype-driven research recruitment is a potentially powerful tool for discovering the functional significance of a variant [82]. In this recruitment strategy, investigators use existing genomic data to identify participants who possess a variant of interest and then contact them to invite their participation in further research [83]. Individuals possessing specific genetic variants of interest could be recruited for collection of biological samples, additional genotyping, other molecular assays, more targeted phenotyping, or a genotype specific intervention. Genotype-driven recruitment (GDR) is a particularly efficient method for functional discovery because it eliminates the need to genotype many individuals from the general population to find those few who actually possess the variant(s) of interest [84]. As we continue to study rarer variants, screening the general population could be prohibitively expensive and time-intensive. However, GDR also stretches existing ethical norms for both initial re-contact and the consenting process for additional involvement in research.

There are several barriers preventing widespread use of GDR to explore functional hypotheses and advance discovery. In order to protect the privacy of participants, data contained in dbGaP has been removed of all personal identifiers such as name, social security number, address and phone number. Therefore, future contact of participants for additional research beyond the scope of the original study poses both logistical and ethical concerns. Logistical barriers could be remedied through the involvement of the submitting investigator or an entity from the original study, assuming the link between the identity of the participant and the data was maintained after repository submission and that sufficient resources to permit re-contact are available. Even if logistically feasible, a range of ethical concerns also pertain, many of which have previously been explored through both conceptual analysis and empirical data involving multiple stakeholders [83, 85-87].

The main concerns surrounding GDR involve the return of individual genetic research results and the less discussed issue of re-contacting research participants to invite enrollment in additional research.

The return of individual genetic research results has been acknowledged as one of the most significant ethical challenges facing current programs of research. GDR is intrinsically tied to the on-going debate about returning individual results from genetic research because recruitment is predicated on the possession of specific genotypes. In the context of GDR it remains unclear whether result return is a necessary element in explaining to participants why they are being invited for additional research. By design, result return would ordinarily be possible only after informed consent and enrollment into a research study. For GDR the ethical concerns are exacerbated because risks to the participant associated with result return are shifted to the recruitment stage.

The primary concern with individual research result return is that revealing findings in the course of GDR could inadvertently deliver unwanted information or mislead the participant into thinking something important to their wellbeing has been uncovered as a consequence of their participation in research. In non-GDR contexts, such risks are usually minimized by the recommendation that only medically actionable findings, of clear analytic validity, be offered for return [88]. As a result, if consent agreements mention return of individual genetic research results it is expected that it will be qualified with statements about improvement to health. With GDR, however, the clinical salience of genetic information is nearly always still under active investigation, so such assurances cannot be made. Consequently, there is greater potential for participants to misunderstand and/or misuse genetic research results that may be conveyed at the time of recruitment.

Ethical concerns with re-contact of research participants have been discussed less in the literature. However, it has been acknowledged that using protected health information (e.g. phone

number or address) for recruitment purposes could be seen as an invasion of an individual's privacy. Although participants have the choice to decline participation, it is possible that this intrusion may diminish their trust and interest in research agendas. It has previously been recommended that a trusted intermediary between new researchers and participants make the initial re-contact so that participants are not alarmed by contact from an unknown, secondary investigator [87]. Although this may be challenging in some circumstances, this approach should not be overtly prohibitive since, in the case of dbGaP, the original investigator or study staff must be involved to link genetic data back to an individual and their contact information.

Although it is ethically desirable for participants to have been informed at the time of enrollment to the initial study of possible future recruitment, current recommendations do not view the lack of warning as precluding re-contact. It is argued that lack of advance notice for recruitment is not a definite harm because participants can still express preferences and disinterest in additional research at the time of re-contact [83]. Conversely, the current consensus is that consent forms that explicitly state or imply that participants will *not* be re-contacted for recruitment should be honored. Empirical evidence suggests that the majority of participants feel there is a strong presumption that these types of statements will be respected [87]. In light of incomplete trust in databanks that already exists, maintaining public support for genetic research is critical for continued success [89, 90]. Therefore, recruitment by biobanks and other researchers unknown to research participants based on sensitive information, such as genetic information, is likely to be inappropriate in most circumstances.

In light of these concerns, a set of recommendations for GDR were made based on a national survey of IRB chairs and interviews with research participants in six studies where genotype-driven re-contact had previously occurred [86, 91-93]. Using this empirical data to discuss the issues associated with GDR, a comprehensive group of stakeholders, including researchers, study

coordinators, and participants from the aforementioned studies, as well as bioethicists, IRB chairs, clinicians, and federal officials then drafted a set of recommendations for ethical approaches to GDR (See table 3.1).

In the recommendations for GDR, there were seven summary statements derived from empirical evidence from various stake-holders [85-87]. It was recommended that (1) participants be informed of potential future re-contact for recruitment during the consenting process for the initial study; (2) participants be given a choice about further research recruitment during consent; (3) participants be contacted by a person or entity they know; (4) that the process for re-contact consider contextual issues pertaining to the persons conducting the study; (5) that thresholds for returning individual genetic research results should not be used for decisions about return of results in the context of GDR; (6) that in most cases, the opportunity to learn individual genetic research results (irrespective of clinical utility) should be offered as part of the recruitment process; and (7) for each study, appropriate approaches to GDR and individual result disclosure should be determined by researchers in consultation with the IRB. Many contextual issues were also mentioned. For instance, the procedure for GDR should consider whether the new research will be performed by new investigators, focus on different medical condition than the one for which participants were enrolled, or whether the initial research or new research study pertains to sensitive or potentially stigmatizing issues. However, the authors felt that these situations did not necessarily prohibit GDR, but would entail more tailored strategies. The one situation that was listed as a likely preclusion for contact about additional research would be when genetic information was derived from participants who never consented to storage and research use of their biospecimens and data. This situation can arise from research using biospecimens leftover from a clinical procedure, where the protocol was exempt from human subjects regulations (45 CFR §46.101(b)(4)).

Although these recommendations were based on empirical data and were designed to not be excessively restrictive, to my knowledge the feasibility of implementing these recommendations in the context of a standing repository such as dbGaP has yet to be systematically explored. To address this uncertainty, this study performed a content analysis of the publicly accessible consent forms for studies for which data are stored in dbGaP and compared the content to the recommendations by Beskow et al.[87] Specifically, this analysis explored whether or not consent forms (1) mentioned future contact and if so, to what end; (2) discussed return of individual or aggregate research results and if so, under what conditions;(3) described processes for how re-contact or return of results would take place and if so, what details were provided; (4) addressed re-contact or return of results under the categories of risk or benefit and (5) allowed participants to express preferences about future re-contact or return of individual genetic research results.

3.3 METHODS

To investigate the feasibility of pursuing GDR with retrospectively collected data currently housed in the dbGaP, a directed content analysis was conducted on study-specific consent forms made available by investigators via dbGaP. A provisional codebook, informed by the current ethical framework for GDR, was developed to identify consent language of relevance to re-contact, result return, and other GDR-related study details. Consent documents were coded and the coded text subsequently analyzed to identify major themes of relevance to GDR recommendations.

3.3.1 Identification and Conversion of Consent Forms

As of May 2013 there were 418 studies in dbGaP identified by searching the term "1s_discriminator" and 346 studies that had an NIH grant number. Of those studies, 312 had a

reported sample size greater than zero. A search of all documentation that was accessible in dbGaP for these studies identified 58 original consent documents or sample consent documents shared across project sites and one document with model consent language (generated from the Electronic Medical Records and Genomics [eMERGE] research network). Documents included in this study were taken from the study site page in dbGaP and were typically available under "study documents." In some circumstances, consent documentation was contained within larger protocol files. After removing 16 duplicate consent forms from this collection, 43 documents remained for analysis. PDF documents were converted to text files using the Python PDF parser and analyzer, "PDFminer" (https://pypi.python.org/pypi/pdfminer/). In this way, 36 of the 43 unique consent documents were successfully converted to text files and imported into the mixed methods analysis program "Dedoose" (www.dedoose.com) for subsequent content analysis (see additional description below).

Most of the documents included in this study were consent forms (n=33), meaning they likely represent the exact language given to the participant during the consenting process. Two documents consisted of model consent language shared across institutions, which in some cases may have been modified to reflect specific requirements of an IRB within a larger consortium study. One document was model consent language drafted by eMERGE [94]. This document was analyzed separately from the other consent documentation and used as a comparison since it remains unknown how many studies have followed this model.

3.3.2 Content Analysis

A provisional codebook was developed to explore all 36 consent forms using *a priori* domains based on recommendations for ethical approaches to GDR (see table 3.1). The domains investigated in this directed content analysis[95] pertained to whether: (1) future contact was mentioned and if so for what purpose; (2) return of individual results was mentioned and if so under what conditions; (3)

the concepts of risk and benefit incorporated statements about re-contact or return of individual genetic research results; (4) a choice about re-contact or return of results was given to participants; (5) retention and access to genomic data or biological samples for future studies was mentioned; and (6) contextual issues of the consented study would warrant concern for GDR. These domains (and the sub-issues within each) were chosen to provide insight into whether current recommendations for GDR would be feasible in light of the consent agreements used to for broad sharing of data currently stored in dbGaP.

Dedoose was used to assign these codes to excerpts of text from each of the 36 consent documents. Dedoose is a web-based application for managing, analyzing and presenting qualitative and mixed method research data [96]. This resource integrates both qualitative and quantitative information to explore interesting trends in data that would otherwise be difficult to extract by mere inspection. Code-specific excerpts were then compared and further subdivided or redefined to capture themes within each major coding domain. Thus, the final coding scheme was developed through an iterative process. In addition, a second investigator (ML) independently applied codes to a subset (~25%) of excerpts pertaining to return of results and re-contact of participants. Coding differences were identified and discrepancies reconciled through an iterative process until greater than 80% agreement was reached. The same second investigator then double-coded five additional consent forms using the full codebook and the codes were then further refined until agreement on inclusion and exclusion criteria for code application was reached (see figure 3.1 and table 3.2)

[i] CS Carlson, MO Goodarzi, SA Rosse, AP Reiner, UM Schick, PL Auer, O Kahsai, X Guo, LJ Raffel, TA Buchanan et al., in prep.

Table 3.1 Coding Domains and their Relationship to GDR Recommendations [87]

	RECOMMENDATIONS	DOMAIN
1	Researchers should disclose the possibility of future contact for further research recruitment during the informed consent process for the initial study	Re-contact: Purpose Content regarding the purpose of future contact.
2	Researchers should consider offering participants a choice at the time of initial consent about future contact for further research recruitment	Re-contact: Choice Content allowing participants to express their preferences about re-contact.
3	Contact about additional research should be made by a person or entity known to the participant	Re-contact: By whom Statements about who will make contact with participants in the future.
4	The process for contacting participants about additional research should be designed based on a range of considerations related to the research team and study question	Study: Purpose, Size, Design Various contextual issues that may reflect the nature of the relationship between the researcher and the participant.
5	Thresholds established for the return of individual genetic research results in general should not be used for decision-making about return of results in the context of GDR	Return of Results: Purpose, What, Language Statements about the return of aggregate or individual research results, the processes in place to do so and the use of positive/negative language suggesting the valence of return of individual genetic research results.
6	In most cases, individual genetic research results should be offered in the context of GDR. A careful series of steps should be used both to avoid leaving prospective participants uninformed about the purpose of the study and to maximize their right not to know unwanted genetic information	Return of Results: Purpose, What, Language, Choice In addition to recommendation 5 domain, content allowing participants to express their preferences about receiving individual genetic research results.
7	For each study, appropriate approaches to GDR and disclosure of individual genetic research results should be determined by researchers in consultation with their IRB.	Study: Purpose, Size, Design See recommendation 4 domain

Figure 3.1 *Hierarchical coding schematic*

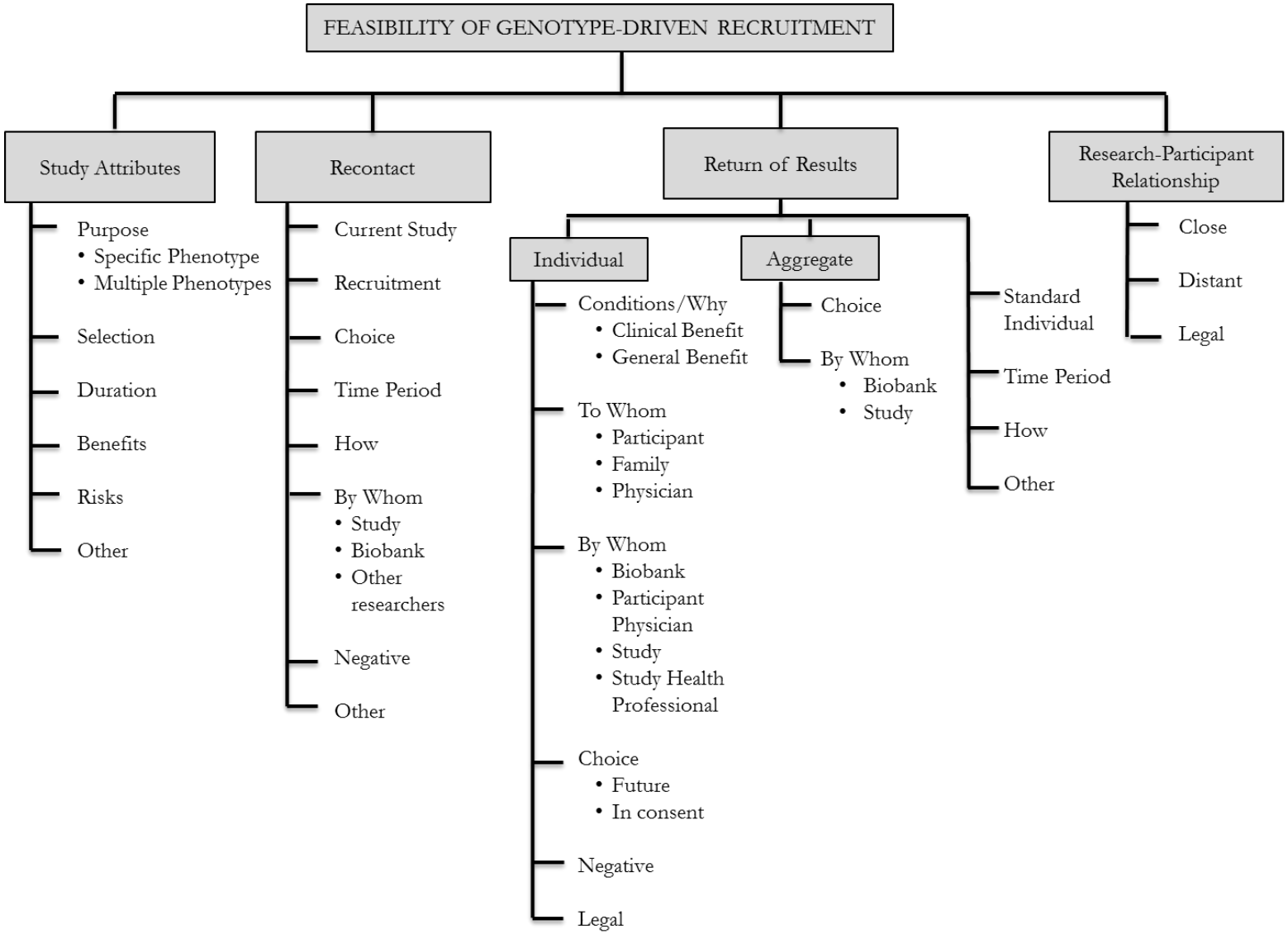


Table 3.2 *Code Book*

CODE	DEFINITION	EXEMPLAR(S)
Study Purpose> Disease Specific	• Statements that the purpose of the study is for a specific disease or trait • Statements that the expected outcomes pertain to a specific disease or trait • The purpose of the study should be at the beginning of the consent document	"The purpose of this study is to understand the genetic basis of psoriasis and psoriatic arthritis."
Study Purpose> Multiple Diseases	• Statements that the purpose of the study is to improve or learn about multiple phenotypes or to improve health/medicine/treatments etc.	"We hope that all of this research will eventually transform the very way we think about health and the way we prevent, diagnose, and treat many diseases—not only those related specifically to the gut, but others as well."
Study> Selection	• Statements about why the participant was chosen for recruitment • Statements about the cohort	"You qualify for this study because you have had a stroke."
Study> Duration	• Statement on how long the study will be conducted or participant is expected to partake • Statements about how long data/sample will be stored and used	"The study will continue for many years. Records, DNA and living cell lines may be kept *indefinitely*." [*emphasis added*]
Study> Benefits	• Statements about who the study will benefit- the participant versus society • Statements about a participant benefitting from received health information • Does not include monetary benefits • If RoR is listed under benefits it is coded as	"One benefit of participating in this project is an evaluation of certain aspects of your health at no cost to you. Information from the evaluation will be available to you. If a health condition is detected during the evaluation, you will be told..."

	both benefits and the applicable RoR code (individual or aggregate).	
Study> Risks	• These statements should generally fall under risk header • Risks pertaining to the storage/analysis/or use of genetic information (examples include IT security and privacy issues) • Risks that may contribute to a participant's concern in the future about how they were identified for re-contact • Does not include risks due to procedure or aspects of the study unrelated to genetic information or tissue/data storage • Apply this code with the applicable RoR or Re-contact code(s) when they are discussed under a risks category	"The only risk of future use or sharing the tissue and/or data is a breach of confidentiality..."
Study> Other	• Noteworthy statements about the study-important context for GDR	"In order to help advance future patient screening strategies, a new condition in our research study might require that your DNA sequence, diagnosis, and de-identified pedigree data be deposited into the NIH's (National Institute of Health) dbGAP database..."
Secondary Use	• Include statements about secondary/future use of genetic data/samples by the same or different researchers for purposes not explicitly defined in present consent form.	"Someone from the Biobank can use my medical record from time to time to get updated information about my health"
Re-contact> Current Study	• Statements that the purpose of re-contact will be to collect more information/samples in the future for the present study aim by	"We may also contact you in the future to ask if you would like to participate in studies addressing health issues other than stroke"

	current study staff. • This is often present in biobanking consent forms and statements that contact by the biobank to collect additional sample are included here unless it would require new consent. • Apply this code with "Re-contact>choice" if a choice is given for re-contacting the participant for follow-up of the current study. • Statements about re-contact to: "see how you are" should be included here.	
Re-contact> Recruitment	• Statements about re-contacting a participant for future studies. • Apply this code for re-contact about anything that will require additional consent. • Apply this code with "Re-contact-choice" if the participant is allowed to express preferences about recruitment into additional studies.	"May we contact you for future studies conducted as part of the HMP? ___Yes ___No If Yes, we will need to look at your Protected Health Information (PHI) to check for your study eligibility."
Re-contact> Choice	• Participants are given a choice about whether they can be re-contacted in the future. Sometimes this includes by whom. • This code should be applied with "Recontact-recruitment" or "re-contact-current study". • Apply this code even if the choice refers to participant preferences being considered previously or later in the consent form or if referring to a participant initiating the contact to make these preferences known.	"*This project will last many years*. Since we will be comparing genetic information to medical and questionnaire information, you should expect to be re-contacted *occasionally* to update the questionnaire information. This will occur at *different times for different people* and would not imply that anything has been learned about you specifically." [emphasis added]
Re-contact>	• Statements on the timing and duration of	"We also ask permission to contact you in the future

Time Period	when participants will be re-contacted	*by mail or phone* to tell you about new research studies." [emphasis added]
Re-contact> How	• Statements about what form or the means of re-contacting the participants. Examples, by phone, letter, email, through a provider etc.	*"Someone from the Biobank* can contact me about offers to take part in more research." [emphasis added]
Re-contact> By whom	• Statements about who will re-contact a participant. Example, Principal Investigator, study staff, other researchers. • "We" is also included as by whom.	"It is possible that findings from future studies will lead to new research studies and we may contact you and ask if you would like to be part of the new study"
Re-contact>By whom> Study	• Statements that the participant will be re-contacted by the study. • Child code of Re-contact-By whom	"Someone from the *Biobank* can use my medical record from time to time to get updated information about my health" [emphasis added]
Re-contact>By whom> Biobank	• Statements that the participant will be contacted by the biobank • Child code of Re-contact-By whom	"If any such study has a different design than this project, and would create risks not considered in this consent, the Institutional Review Board would require the *researchers* to inform you of these risks..." [emphasis added]
Re-contact>By whom> Other researchers	• Statements that the participant will be contacted by other researchers • Child code of Re-contact-By whom	"In the unlikely event that we find information that may be clinically relevant and clinically reliable, we will contact you directly; *otherwise you will not hear from us*."[emphasis added]
Re-contact> Negative	• Use of language to capture negative valence on re-contact • Discouraging statements avoiding re-contact of participants • Includes statements that the participant will not be re-contacted and why	

Re-contact> Other	Catchall for items that don't fall within a code – but are relevant to Re-contact	"There is a small chance that researchers could discover something that might be very important to your health or medical care right now. If this happens, we will contact you to see if you want to learn more."
RoR*> Individual	• Meant to capture any discussion of the return of individual genetic research results • Includes any statement about the return of Individual research results that could be genetic, including statements about results from blood tests • Does not include return of individual results from "standard tests" as genetic testing is not typically considered a standard or routine medical tests • Should be applied when the statement about the return of individual results is negative.	"There is a small chance that researchers could discover something that might be *very important to your health or medical care* right now…" [emphasis added]
RoR*>Individual> Conditions/Why	• Statements about why results will be returned	"If we have findings that might be of *medical importance*, we may re-contact you to tell you about the results." [emphasis added]
RoR*>Individual> Conditions/Why> Health Benefit	• Statements that the results will be returned if there is something important for their clinical care	"If as a result of participation in this study we obtain information that could significantly affect your health *or wellbeing*…" [emphasis added]
RoR*>Individual> Conditions/Why> General Benefit	• Broad statement of Benefit not necessarily in the clinic. • "Improve well-being" • When "or well-being" is included with clinical meaning/benefit just code well-being	"In the unlikely event that we find information that may be clinically relevant and clinically reliable, we will *contact you directly*…" [emphasis added]
RoR*>Individual>	• Statements about who will receive the	"you may be asked if you are willing to *share your genetic*

To Whom	participant's results.	*information with your family* members" [emphasis added]
RoR*>Individual> To Whom> Family	• Statements about the return of participant results to family members • Usually in pedigree/family studies • Child code of Individual genetic	"If we have findings that might be of medical importance, we may re-contact you to *tell you* about the results." [emphasis added]
RoR*>Individual> To Whom> Participant	• Statement that results will be returned to participant. • If the statement says results are returned to a physician to return to the participant this code still applies but will occur with the "by physician" code • Do not apply to statements that results will not be given to participant (that is the "RoR-Individual-negative" code) • Do not apply to statements where results are being given to the physician only and do not specify whether participant will be informed (that is the "RoR-Individual-To Physician only". • Child code of RoR-Individual	"Restrictions on release of results *to participant's physician*?" [emphasis added]
RoR*>Individual> To Whom> Physician only	• This is applicable when the results are only given to the participant's physician • If the statement also says the results will then be given to the participant apply the codes "RoR-Individual-to participant and "RoR- By Whom-By Participant's Phsician" • Child code of RoR-Individual	"If clinically relevant results are found from your blood work, *we* will come to your home to discuss these results with you in person." [emphasis added]
RoR*>Individual> By Whom	• Statements about who will deliver	"In the rare case that information from your sample is determined to be clinically significant, the Research

	information. Example, Principle Investigator, study staff, other researchers, their provider, genetic counselor etc.	Ethics Board may authorize *your physician* to contact you." [emphasis added]
RoR>Individual> **>By Whom>** **Participant's Physician**	• Statements that the participant's physician will return results	"If we have findings that might be of medical importance, *we* may re-contact you to tell you about the results." [emphasis added]
RoR*>Individual> **>By Whom>** **Study**	• Statements that the study staff will return results • Includes "we" assuming that this means someone affiliated with the current study	"If these researchers use the sample for future research and decide that a test result may be useful for your health care, they may contact the Mayo Clinic and *Mayo would then contact you* to offer you the choice to learn the test results." [emphasis added]
RoR*>Individual> **By Whom>** **Study Health Professional**	• By a genetic counselor or other health care provider that is hired by the study or affiliated with the study. • Apply this when it is clear the researcher is also a physician	"… I agree to allow the FHS to notify me, and then with my permission to notify my physician."
RoR*>Individual> **Choice**	• A choice is given to the participant to learn individual or aggregate research results • This code should be applied with RoR-individual Refers to both choice in the consent document and choice in the future	"…we will contact you to see if you want to learn more."
RoR*>Individual>Choice> **Future**	• Statements that the participant will be contacted and then can choose to whether or not to receive individual results	"Would you like to be contacted by Dr. XXXX or one of their associates if in the future a genetic testing result is found from your blood/genetic sample that may be helpful in diagnosing…"
RoR*>Individual>Choice> **In consent**	• Statements enabling the participant to express	"No information resulting from the analysis of your DNA or about your genetic status, such as the

	preferences on the RoR in the current consent document	probability of developing a specific disease, will be provided to you. Research results are often preliminary, inconclusive, and not necessarily valid for decisions concerning patient care and treatment. Preliminary information, if given to you, could create false conclusions and significant risks."
RoR*>Individual> Negative	• Use Positive/Negative language to capture valence of Returning individual research results that could be genetic. • Does not include statements that a meant to address the protection participant privacy when returning aggregate research results (such as you will not be individually identified in presentation, publications, or newsletters) • Apply code with aggregate return if the statement says you will not receive individual results but… • Includes reasons or statements for NOT returning results • Should be applied with either individual or aggregate RoR codes	"You will not be provided with your personal research or genetic information obtained during this study, since the research will be done by a lab that is not authorized by the Clinical Laboratory Improvement Act (CLIA) to provide clinical information."
RoR*>Individual> Legal	• Comments on legal obligation to re-contact or to policy to prevent return of results (examples include references to not being CLIA certified)	"With your permission, a summary letter of routine test results from this exam will be sent to you and your physician."
RoR*> Standard Individual	• Statements about the return of standard non-genetic medical test results.	"If clinically relevant results are found from your blood work, we will come to your home to discuss these results with you in person. This will usually occur within one month of your last clinic visit."
RoR*>	• Statements on the timing and duration of	"Results of this study may be published in scientific

Time Period	when participants can expect to receive study results	journals or presented at medical meetings."
RoR*> Aggregate	• Return of aggregate research results. • Should not be applied to statements that only state individual research results will not be returned without mention of returning aggregate research results • RoR aggregate should be applied with RoR-negative if the statement says individual results will not be returned but aggregate research results will. RoR-negative is defined for individual and not aggregate	"If all of the results from the blood work are normal, you will be notified by mail."
RoR*> How	• Statements about what form or the means of delivery. Example, by phone, letter, email, through a provider etc.	
RoR*> Other	Catchall for items that don't fall within a code – but are relevant to Return of results	"In this study, investigators will not tell you what they find out about you, nor will the investigators, your doctor, or the University contact you if a test becomes available to diagnose a condition you might have or later develop. This is unlikely to have a negative impact on your health because both psoriasis and psoriatic arthritis are multifactorial diseases, meaning that multiple genes and the environment determine risk of disease."
Great quote	Catchall for noteworthy – surprising or emblematic statements	"The NINDS Repository at Coriell does not perform genetic testing services and therefore will not return genetic research results to any individual research participant. Since other researchers do not know your identity, you will never receive any genetic information determined using your

		sample."
Relationship> Distant	• Statements/words that reflect a distant relationship between the researcher and participant	*"Dr. XXXX and Dr. XXXX invite you* to be in a study to learn more about dental health in families, including behaviors, genes (those factors that determine a person's physical characteristics"
Relationship> Close	• Statements/words the reflect a close relationship between the researcher and participant	"The consent given here to collect and process information about my health is irrevocable. I have already been informed that I can terminate my participation in the clinical study at any time. If I should do so, the anonymous data stored up to that time point will be further used, so far as is required: To detect effects of the drug being tested."
Relationship> Legal	• Statements/words the reflect a predominantly legal relationship between the researcher and participant	

***** RoR=Return of Results

3.4 RESULTS

Although the recently developed ethical framework for GDR is based on empirical data representing a variety of stakeholders, it remains unclear whether these recommendations align with the content of consent forms from studies currently housing data in dbGaP. As such, this study performed a content analysis of the publicly accessible consent forms available in dbGaP and compared the content to current recommendations [87]. Three broad domains derived from these recommendations guided this analysis: (1) contextual issues, such as participant characteristics, study–duration, size, purpose, and restrictions for secondary use of generated data; (2) future contact–described purpose, process, and inclusion of participant preferences; and (3) return of results– described purpose, process and inclusion of participant preferences.

Across these domains we found that there was considerable heterogeneity. The contextual issues examined revealed that studies ranged from a few participants with ongoing contact to very large studies involving a single contribution of a blood sample and completion of questionnaire(s). Some studies only enrolled healthy participants, some involved research of potentially stigmatizing conditions, some enrolled families, others included minors, and a few involved an intervention for participants with a particular condition. As such, the participant relationship with the original investigator, and the corresponding implications for GDR, varied considerably. Content pertaining to re-contact and return of results also had considerable variation. Many consent documents failed to mention either of these issues. Discussion of risks and benefits sometimes included return of individual research results, though discussion of result return under benefits was often ambiguous as to whether this would include genetic results. Although some consent documents did allow participants to express preferences about these topics through tiered consent or binary yes/no responses, most offered no choice at all. No consent document included language addressing all of

the recommendations made in the current ethical framework. Of particular note, none of the documents in this analysis mentioned the possibility of future recruitment based on genetic research results.

3.4.1 Contextual Issues- Study Duration, Participant Characteristics, Study Size, and Purpose

Current recommendations for GDR advise that the process for contacting participants and decisions around offering the return of individual results should be based, in part, on characteristics of the new research and the study for which the participant originally enrolled. Therefore, this study examined linked study descriptors, as reported in dbGaP, as well as consent language describing study characteristics.

Table 3.3 shows the study characteristics and participant demographics of the 36 studies for which consent documentation was analyzed. It was posited that the duration of study requiring active involvement of a participant might reflect the likely strength of relationship between the researcher and the participant. Of the 36 studies, seven were longitudinal and three involved an intervention. Thus, based on study design, as well as descriptions of involvement within consent forms, just under a third of the studies analyzed involved on-going contact between participants and the study staff. Alternatively, case-control, case-only, control, and biobanking studies indicated minimal clinic visits. As such approximately 60% of the consent documents analyzed were from studies that required minimal active participant involvement, such as on-going clinic visits, completion of questionnaires, adherence to an intervention, etc. The remaining consent documents were for family studies, which varied in duration, but pose important challenges for re-contact and potential return of genetic research results. Although there were consent documents from 16 different funding initiatives in this analysis, 38% of consent documents were made available in

dbGaP by the Gene Environment Association Studies (GENEVA) (n=7) and Human Microbiome Project (HMP)(n=7).

It was hypothesized that smaller studies might enable more dialogue or more extensive interactions between a research participant and the investigator in comparison to very large studies. Furthermore, the size of the study may have implications for how participants feel about their particular role in the research study. The median study size was 1911 participants; however, the study populations ranged from 23 to 12,771 participants. Twenty studies stated in the consent document the expected number of participants to be enrolled. Review of statements in the consent forms revealed that the studies varied greatly both in the number of participants and in the number of institutions involved. As such, when many institutions were involved, identifying the "original" investigator or a study staff member familiar to a participant could be challenging.

"The plan is to have 3000 people take part in this study at Mayo Clinic Rochester, with up to 3200 people being screened to find enough eligible people for this study."

"The study will be conducted at approximately fifty-six medical centers in North America including [ENTER CENTER NAME]."

It was found that 24 of the 36 (66.7 %) consent documents stated that the purpose of the study was to explore a particular disease as opposed to many diseases to improve health more broadly. Although the majority of consent documents did not state that the purpose of the study was to investigate multiple diseases, all but one discussed secondary use of stored data or samples. In fact the length of excerpts describing secondary use of data was typically much longer than the discussion of the aims of the study. Therefore, at least in principle, participants should have been well-informed of the potential for additional studies to look at conditions unrelated to the aims of the study they enrolled in.

Current recommendations suggest that there are important considerations for tailoring GDR based on particular features of the study population. All but four of the studies considered here enrolled at least some participants with a specific condition or at higher risk for a condition. Although the majority of studies enrolled adults (n=29), a smaller proportion involved minors (n=7), for which return of research results likely requires additional precaution.

The issue of mistrust in research is particularly relevant to GDR because it requires linking genetic data with personal identifiers that were purposefully removed to protect the privacy of the participant. Although these concerns are relevant for all participants, past abuse of marginalized populations by government funded research may exacerbate these concerns. For this reason, this analysis considered ethnicity to be an additional consideration, though this was not made explicit in the recommendations for GDR. Nearly half of the studies represented in this analysis (n=15) were restricted to White participants, 10 (28%) enrolled participants from more than one ethnicity, two (5%) were restricted to African Americans, one (3%) was restricted to Hispanics, one (3%) was restricted to Caucasians (Pakastani, Arabic, or White), and the remaining seven (20%) did not report a race variable nor recruitment criteria based on race. Of the studies defined as multiethnic, half (n=5) enrolled participants that were predominantly either Hispanic or White, and the other half (n=5) enrolled participants that were either Black or White. Thus, the consent documents included in this analysis reflect the current overrepresentation of Whites in GWAS.

Table 3.3 Study Characteristics for 36 studies contributing consent documentation

Document Type	
Consent Form	33
Sample Consent Form	2
Model Consent Language	1
Study Design	
Intervention	3
Case-Control	16
Case-only	4
Control	1
Biobanking	1
Longitudinal	7
Parent-Offspring Trios or Twin	4
Size of Study	
23-1844	16
1845-3666	12
3667-5488	2
5489-7310	4
7311-9132	1
9133-12776	1
Disease Status	
Healthy volunteers	4
Case subjects	7
Cases and Healthy volunteers	24
Healthy volunteers with HD testing	1
Type of Study Sample	
Adults	29
Children	4
Parents and pediatric subjects	3
Ethnicity	
White	15
Multiethnic	10
Black	2
Hispanic	1
Caucasian	1
Undefined	7
Data Use Restrictions	
Health Research Use (HRU)	4
General Research Use (GRU)	12
Disease Specific Use (DSU)	11
GRU & DSU	1
Human Microbiome Project Use (HMU)	3
HRU & DSU	2
Nonprofit research use (NPU)	1
Data now unavailable	2

Data housed in the dbGaP are organized into groups that consist of all the data from study participants who have agreed to the same data use as specified in the informed consent document for each specific study. As such, data access is only approved in the unit of consent group. Data use restrictions were summarized into the following categories: Health Research Use (HRU), General Research Use (GRU), Disease Specific Use (DSU), both GRU & DSU, Human Microbiome Project Use (HMU), both HRU & DSU, Nonprofit research use (NPU) or "data now unavailable". HRU restricts usage to health related research purposes, GRU permits data to be used for all research purposes, DSU is a category defined here to describe data usage restrictions to only certain types of diseases, HMU restricts usage to studies under the human microbiome project, NPU prohibits use of data in for-profit studies and "data now unavailable" means data has been removed from dbGaP and can no longer be used. The vast majority of studies were designated GRU and DSU. Interestingly, approximately one third of these studies would not allow, on the basis of these data use restrictions, GDR for a study pertaining to a phenotype other than the study for which the participant was enrolled. However, given that additional consent would be required for a new research study, data use restrictions would not necessarily preclude GDR for studies defined outside of these categories.

Data use restrictions are institutional agreements that may or may not coincide closely with consent language describing the original purpose of research, a contextual feature relevant to the feasibility of GDR. To obtain a better understanding of the type of research projects participants originally consented to, this anlaysis examined consent language describing study purpose. It was found that the majority of the analyzed consent documents stated that the purpose of the study was related to a specific disease or trait without mention of other diseases. However, 12 documents, five of which were from the Human Microbiome Project (HMP), also stated that that the purpose of the study was, at least in part, designed to examine a variety of conditions influencing human health.

The purpose of studies reflected three general themes, the first of which was very specific and related to a particular disease, which may not necessarily prohibit GDR, but may entail additional steps to ensure that participants are not alarmed by recontact for additional studies. The second type of purpose mentioned a specific condition of interest, but also stated that a secondary aim was to consider other conditions. The last purpose statement, which was the least frequent and the most appropriate for GDR context, was to collect information and samples for general research that could improve health. It should be noted that consent documents from the HMP stated that the purpose of the research was to learn about multiple diseases to improve health.

Table 3.4 General themes observed in the description of study purpose

1. Purposes related to a specific disease or trait
"The goal of this study is to identify the specific genetic defect underlying the condition in your family."
2. Statements that a specific condition is of interest, but other conditions will be studied
"The overall goal of the National Institute for Neurological Disorders and Stroke (NINDS) Repository is to establish a bank of samples from individuals with neurological diseases and controls. This bank will allow for distribution of DNA to scientists to help learn more about neurological diseases and many other diseases."
3. Statements that the study is to improve health in general
"The purpose of this research project is to collect and store human samples (such as blood) and health information. Researchers can then use the stored materials in future studies. Through such studies, they hope to find new ways to detect, treat, and maybe even prevent or cure health problems."

There were only two documents that did not explicitly state that biological samples or data obtained in the study could be used for additional studies in the future. Statements about secondary use most often stated that the samples could be used in the future for medical research on a wide

range of diseases. However, statements varied from restrictive to very broad use. Some documents allowed participant preferences about how and for what purposes data and samples could be used in the future.

Table 3.5 General themes on data use restrictions included in consent documentation

1. Broad statements:
"Once information is de-identified, it may be used and shared for other purposes not discussed in this consent form."

2. Medical Research:
"All de-identified samples and data acquired from this research may be shared with other scientists for medical research."

3. Participant Preferences about usage:
"Type of restriction on use/storage of DNA?
CVD research C
ARIC only ..... A
No use/storage of DNA ..... N
Other 0
Specify details of DNA restrictions:……"

In the subset of studies represented by the consent documents analyzed in this study, there was considerable variability in factors that would impact participant-researcher relationships. The studies varied in size, duration, involvement of clinic visits or completion of questionnaires, and most involved the evaluation of a medical condition. Although most studies involved adults, approximately 20% involved at least some pediatric subjects. The over-representation of participants of European descent in genetic research was also reflected in the studies represented in this analysis. There was also great deal of variation in the description of the purpose of the study and discussion of dbGaP submission for secondary use. The majority of consent documents stated that the purpose of the research was to study a specific disease. However, the majority of consent forms discussed secondary use of data in great detail. Very few studies offered any options to participants about what type of research or which investigators could use generated data in the future. As such, the concerns

for risk to participants by GDR are likely to vary a great deal, even in this small subset of studies that have submitted data to dbGaP.

3.4.2 Future contact: Purpose, choice, and process

GDR requires direct re-contact of study participants and current recommendations suggest that participants should receive advance notice during the initial informed consent process about the possibility of future recruitment. Therefore, this analysis also explored whether re-contact was mentioned in existing consent documents and, if so, how re-contact was discussed. It was found that only 22 of the 36 (61.1 %) consent documents mentioned future contact. When future contact was mentioned in consent documents it was either for purposes related to the aims of the original study, such as obtaining more samples or filling out a questionnaire, or re-contact was discussed in the context of recruitment to additional studies. When re-contact was mentioned it was usually (n=15) for the purposes of recruitment, rather than for follow-up (n=7) pertaining to the aims of the original study. Future contact for recruitment was usually discussed as invitation into secondary research outside of the aims of the current study. However, in six documents re-contact for recruitment was discussed in the context of collecting more information for the current study or for returning research results from the study in which the participant enrolled. In some documents, mention of re-contact included follow-up, recruitment, and return of research results in the same sentence as seen in the following example:

"We may re-contact you in years to come to request a *follow-up visit*, request *more information*, or *discuss research findings*" [*emphasis added*]

Re-contact was usually discussed at the beginning or at the end of the consent form and was only discussed under study risks once. In this document the risk appeared to pertain more to confidentiality than to re-contact itself:

"Your blood sample will be stored under a code number that will be connected to your name or other unique identifier. The key to this code will be accessible only to the Principal Investigator, the Co-Investigator and the Study Coordinator. *If there is a medical reason to seek specific information from you* in the future for purposes of this study, you may be contacted by the Investigators or your doctor" [*emphasis added*]

Language disallowing future re-contact was only present in two consent documents. The first was a document for parent assent of their child to partake in a study on the genetics of psoriasis. The document specifically stated that in agreeing to partake in the study, participants "waive their right" to be consented or contacted about future studies using their de-identified data (genetic or other health information). Only one additional document explicitly stated to participants that they would not be contacted in the future by the study investigators:

"In the unlikely event that we find information that may be clinically relevant and clinically reliable, we will contact you directly; otherwise *you will not hear from us.*" [emphasis added]

Generally, these results suggest that future contact was not seen as posing undue risk to the participant. This is presumably because upon invitation into additional research participants are able to decline participation [82,86]. Likewise, recommendations do not prohibit GDR in the absence of content pertaining to future contact [87]. However, when documents specifically state or imply the participants will not be contacted in the future, empirical evidence suggests that these statements should be honored [87].

In addition to providing advance notice of potential future contact, recommendations urged consent documents to allow participants the opportunity to decline future contact. Approximately a third (n=10) of the consent documents provided participants with the ability to express preferences

about re-contact, by providing a tiered or binary yes/no option. The majority of consent documents (n=6) asked if the participants would be willing to be re-contacted by the study (often indicated by the use of "we") for future research. Two of the documents did not indicate who would re-contact the participant. Of the remaining documents, one was model consent language for participation in a biobank study drafted by eMERGE and the other was for the informed consent of participation in the Personalized Medicine Research Database. The eMERGE model consent language was the only document mentioning re-contact by a biobank unaffiliated with the study. The consent document for the Personalized Medicine Research Database asked if a participant would allow contact by an entity unrelated to the study in which they enrolled:

"This database will be used for many studies. New information or knowledge will cause new research questions to be asked, and new and different studies designed to answer them. If any such study has a different design than this project, and would create risks not considered in this consent, the Institutional Review Board would require **the researchers** to inform you of these risks and ask you to sign a separate consent for this separate study. You would be under no obligation to participate."

Two documents asked if participants would be willing to be re-contacted for more samples or information. One document asked if participants would be willing to be contacted for any reason. The remaining document asked if the participant would provide their social security number so that they could be found in the future and did not allow any preferences to be specified about future contact explicitly; although allowing the option to refrain from providing a social security number is arguably a means for expressing their disinterest in future contact. Therefore, the vast majority of consent documents analyzed did not allow participant preferences about re-contact to be expressed. No documents mentioned choice about re-contact for the purpose of recruitment on the basis of genetic research results.

This study also looked for content describing processes for future contact, such as how participants could expect to be contacted (by mail, phone, email, etc.) and whether there was a time frame on when this contact could occur. As indicated above, the majority of studies that discussed

future contact for recruitment stated that if the participant was re-contacted it would be by someone from the current study. Surprisingly, no documents actually provided a time frame or frequency that a participant could be re-contacted for recruitment, as the eMERGE sample consent language had recommended. These results suggest that there are many differences in the way re-contact is discussed across studies and that very few have discussed the procedures in place for future contact with participants. However, when participants were told they may be contacted in the future, content in the documents analyzed suggest participants are most likely to expect that contact would be from someone related to the study in which they enrolled.

3.4.3 Return of Results: Why, when, how, choice

Since the current ethical recommendations suggest that in most circumstances participants should be given the choice to learn their genetic research results in the context of GDR, this study also examined whether consent documents discussed the return of results and, if so, how return of results were mentioned. In the coding, a distinction was made between return of aggregate research results versus return of individual research results. This study explored whether the return of individual results was discussed negatively or as a potential benefit to participation. Given that most genetic research is conducted in laboratories that are not CLIA certified for genetic testing, analysis included whether consent forms discussed legal barriers to returning research results. If result return was discussed, discussion of processes in place to facilitate the return of individual results was also of interest.

Just under a third of the consent documents (n=10) mentioned the return of aggregate research results and almost all stated that the aggregate results would be returned through a newsletter. Even fewer consent documents mentioned the possible return of individual genetic research results (n=9). Discussion of return of individual research results was almost always

discussed negatively or under a study risk heading. When return of individual results was discussed under benefits, it was either with regards to standard non-genetic tests or with the statement that results would only be returned if they could significantly improve the participant's health:

"BENEFITS …The results of standard medical tests will be given to you. Study results from your DNA or white blood cells will not."

"This study may be of no direct benefit to you or members of your family. If as a result of participation in this study we obtain information that could significantly affect your health or well-being, we will attempt to inform you of the existence of this information. You may then decide if you wish to know what we have learned."

Thus, discussion on the return of individual research results was almost always in the context of incidental findings that could improve the health of the participant. There was only one exception to this:

"If you choose, you will be notified of results obtained through this study. We will not disclose non-maternity or non-paternity information."

It is possible this statement would allow a participant to inquire about individual genetic research results. However, the brevity of the statement makes this unclear.

Return of individual research results was almost always discussed as return by the investigator or research staff to the participant. However, there were a few exceptions. One study discussed the possibility of releasing research results to the physician with the permission of the participant and another consent document mentioned return of individual results to family members if they agreed to partake in a pedigree study. Four studies discussed that individual research results would be returned by a health care provider (including a genetic counselor) affiliated with the study and one stated that results would be returned to the individual's physician to return to the

participant. Only one study had systems in place to return results by coming to the participant's home and discussing the results of "blood work." It was unclear from this consent document whether this would include genetic results. Surprisingly one study stated:

"If we identify genes in the future that show a definite involvement in CVD and which could affect your medical treatment of CVD, you will be notified and have the opportunity to receive your *genetic testing results* if you wish." [Emphasis added]

This type of language could be particularly misleading to a participant during GDR and extra precaution would likely be necessary to explain to the participant that they are not being re-contacted because of genetic results of relevance to medical treatment.

There were three consent documents that mentioned that research results are not from CLIA certified laboratories, and as such are not considered to be clinical results.

"The results of this study will only be used for research purposes. The laboratory studies that we plan to conduct are not approved for clinical use and cannot be used for clinical care. At your request, we will provide you with a summary of the combined research results from the study."

"…these results are preliminary and if possible would need to be confirmed by a medically certified laboratory (CLIA) in consultation with your physician and or a Clinical Medical Geneticist. We will share the results with the clinician who helped us obtain your sample."

"You will not be provided with your personal research or genetic information obtained during this study, since the research will be done by a lab that is not authorized by the Clinical Laboratory Improvement Act (CLIA) to provide clinical information."

Although each of these statements reflects that research results are not clinical information none state any legal restrictions to prevent result return.

As was expected, fewer documents offered options about the return of results than re-contact. There were only five documents that asked participant preferences about the return of

results. Four of these allowed the participant to choose whether incidental findings important for their health could be returned. The remaining document only allowed a choice to be made about aggregate research results. One statement specifically said that the choice to find out what the researchers had learned would happen after re-contact rather than the choice being provided in the actual consent document.

3.5 DISCUSSION

As we attempt to interpret the vast catalog of GWAS results we will inevitably need to explore the effect of particular variants in research participants. GDR is a particularly efficient strategy for exploring functional associations but confronts significant ethical concerns around the return of individual research results and fostering participant trust in research. Performing a directed content analysis of 36 consent documents from dbGaP revealed significant variability in how various studies have addressed issues central to GDR. This analysis observed substantial variability in contextual issues across the studies considered in this analysis suggesting that tailoring GDR to address the associated concerns will not be a trivial task for researchers and IRBs. Furthermore, the discussion of re-contact found in these documents usually pertained to recruitment but never mentioned recruitment based on genetic research results. Consent language about return of research results also varied considerably, but explicit discussion of the return of individual research results was invariably only when result return could improve health. The language used to discuss each of these topics was diverse. The substantial variability in discussion of these topics may, in part, reflect uncertainty on the approaches to take in response to mandatory submission of de-identified data.

Recommendations for GDR (see table 3.1) state that researchers should disclose the possibility of future contact for invitation into new research studies during the initial informed consent process. In this analysis, the majority of consent documents did not mention future contact.

However, when re-contact was discussed, it was typically for recruitment purposes. Empirical evidence suggests that participants view advance notice about future contact as a demonstration of respect, and that this may influence their decision about research participation [86]. Thus, although lack of content in consent documents about GDR doesn't necessarily restrict its use, advanced notice about the possibility of re-contact may increase enrollment and foster trust between research participants and researchers. It has also been recommended that consent language implying or explicitly stating that future contact will *not* occur should be honored[87].

This study found that two of the 36 consent documents analyzed included statements that participants would not be contacted in the future. Although the intention of these statements was to permit future use of research results without obtaining additional consent or to remove participant expectations for return of individual genetic research results, these statements also inadvertently preclude the possibility of future recruitment. Overall, the statements about future contact varied considerably in language and in length. Some consent documents included statements that recruitment would be for other studies in the same consortium or by the same investigators, some statements were for additional research pertaining to the same condition, while others were very broad. Sometimes statements included processes for how the participants would be contacted (example by phone or by mail). Usually statements reflected that the participant would be contacted by someone affiliated with the study they enrolled in, but not always. It remains unclear whether this variability would preclude recruitment for certain types of future studies.

It was also recommended that researchers should consider offering participants a choice at the time of initial consent about future contact for further research recruitment. This study found that 11 of the 15 documents stating that re-contact may occur also provided participants a choice about such re-contact. Again, these statements varied a great deal in the language used and the

permissions received. Very few provided multiple options or tiered consent about re-contact, yet many included statements that were for specific diseases or contact for research by the same study or institution. As such, these statements may prevent GDR for certain conditions or by outside investigators.

Recommendations also stated that contact about additional research should be made by a person or entity known to the participant. In the model consent language drafted by eMERGE it was recommended that participants be asked whether they could be re-contacted by someone from the biobank or by other researchers. When a participant gives explicit permission for recruitment by someone outside of the original study, it is not clear whether the incorporation of genetic research results would make it necessary to include someone familiar to the research participant. In many circumstances the original investigator may be unavailable to make this contact. If the participant gave permission for another entity to contact them, requiring contact be made by someone from the original study could be unnecessarily prohibitive. However, based on the actual consent documents analyzed here, it is likely that participants were given the opportunity to permit someone outside of the current study to contact them.

It has also been recommended that the process for contacting participants about additional research should consider a range of contextual issues related to the research team and study question [87]. Although it is not possible to predict all of the study questions that may take place in the future, this analysis attempted to investigate the diversity of studies housing data in dbGaP. In doing so, it was established that there were many contextual issues that would need to be considered, making this recommendation a necessary, but non-trivial task. The appropriateness of re-contact, particularly in the context of GDR, is partially related to the relationship between the researcher and the participant. To this end, the duration of the study, the level of involvement by the participant, and the number of participants enrolled in the study may all be indicative of the nature of the

relationship between the participant and study investigators. Using these descriptors and content pertaining to these subjects, the nature of the relationship varied greatly across studies. However, as large studies also typically have a large number of visits, in the absence of input from the original investigator, it is difficult to determine whether participants would be familiar with staff from the original study.

It was also found that more than half of the studies placed some restrictions on data usage and most studies involved a particular condition or trait. If research is performed by new investigators, or focuses on different medical condition than the one for which participants were enrolled, or if the initial research or new research study pertains to sensitive or potentially stigmatizing issues, there are likely additional concerns for IRBs. If data use restrictions are used to guide decisions about GDR, then this analysis indicates that there are many limitations.

It was also acknowledges that thresholds established for the ethically appropriate return of individual genetic research results should not be used as the basis for offering return of results in the context of GDR, per se, and that in most cases, such results should be offered. It was found that when return of results was discussed there was a very strong emphasis that a choice would only be given if the information had clinical utility. The strong emphasis of individual result return occurring only under these circumstances is troublesome for GDR because it could mislead participants into placing undue weight on uncertain information received as part of the recruitment process.

In light of the significant challenges for interpreting GWAS results, GDR will be an essential tool in the translation of genetic associations for precision medicine. However, little is known about whether our current research infrastructure will support recruitment-by-genotype. The diversity of approaches taken by studies, regarding both result return, future contact, and the role of participant preferences, is notable. It is reassuring that almost all of the consent forms stated that genetic information generated from consent documents is provisional and has uncertain significance.

Additionally, these statements typically alluded to the fact that it would take additional work to disentangle the functional meaning.

3.5.1 Limitations

Certain limitations of the study affect the likely generalizability of these findings. This study assessed consent forms from only a small proportion of the total number of studies in dbGaP and only included those that were made available by the investigators. In addition, this analysis could not assess how these forms were used in practice. Informed consent is an ongoing process that involves considerable in-person explanation and communication and often more than the document. Therefore, a more comprehensive analysis of consent understandings would have involved the original study staff, who are the best informed about what expectations regarding re-contact and return of research results were in play.

Many additional questions about the practicable feasibility of GDR also remain unanswered. For example, this study does not provide empirical evidence regarding whether submitting researchers are interested in partaking in future research for GDR and if they are not the ones conducting the research it is unclear how they would be compensated for their help identifying and recruiting participants to follow-on research. It also remains unclear how long links between participants and data housed in dbGaP are maintained. It is possible that post submission these links are destroyed by many studies to protect participant confidentiality.

3.5.2 Future Research

To better address the feasibility of GDR using archived data sources such as dbGaP, future research will need to collect additional evidence on whether GDR is not only ethically permissible

(according to the language of the consent agreement) but also logistically feasible. For instance, researchers should be contacted by dbGaP to address both their interest in helping to recruit participants into a genotype-driven study and to ascertain whether links have been maintained between research participants contact information and their genetic data.

Furthermore, this study only considered consent forms for studies housed in dbGaP. However, there are a number of other databases, both public and private, that were not considered. As such, it will be important to understand how studies and consenting processes may be the same or different among different resources.

Another issue that has not been explored in great detail is whether there is need for a tracking system to be put in place to make sure that participants are not over-burdened by frequent GDR related requests. Since there are differences in allele frequencies between populations and some groups are less represented in genetic research than others, it is possible that participants from otherwise underrepresented groups might be contacted more frequently than others. There should be additional discussion on the implementation on tracking systems to maintain participant preferences and to avoid harassment.

3.5.3 Conclusions

Overall, the findings from this work suggest considerable heterogeneity in study approaches to re-contact and return of individual research results. Although relevant to the widespread feasibility of GDR as a general approach to post-GWAS functional confirmation, this should not be seen as discouraging or disallowing the practice of genotype driven recruitment altogether. Instead, they emphasize that diverse contextual issues are in play and will need to be negotiated as GDR and related approaches are considered. Moving forward, new studies would do well to consider the

recommendations by Beskow et al.[87] and design their consent and sample governance processes to facilitate genotype-driven recruitment. Future studies should, wherever possible, enroll participants under an expectation that GDR may occur.

REFERENCES

1. Manolio TA. Bringing genome-wide association findings into clinical use. Nat Rev Genet 2013 Aug;14(8):549-58.

2. Knowler WC, Williams RC, Pettitt DJ, Steinberg AG. Gm3;5,13,14 and type 2 diabetes mellitus: An association in American Indians with genetic admixture. Am J Hum Genet 1988 Oct;43(4):520-6.

3. Need AC, Goldstein DB. Whole genome association studies in complex diseases: Where do we stand? Dialogues Clin Neurosci 2010;12(1):37-46.

4. Rosenbloom KR, Sloan CA, Malladi VS, Dreszer TR, Learned K, Kirkup VM, Wong MC, Maddren M, Fang R, Heitner SG, et al. ENCODE data in the UCSC genome browser: Year 5 update. Nucleic Acids Res 2013 Jan;41(Database issue):D56-63.

5. Wittkowski KM, Sonakya V, Song T, Seybold MP, Keddache M, Durner M. From single-SNP to wide-locus: Genome-wide association studies identifying functionally related genes and intragenic regions in small sample studies. Pharmacogenomics 2013 Mar;14(4):391-401.

6. Han B, Chen XW, Talebizadeh Z, Xu H. Genetic studies of complex human diseases: Characterizing SNP-disease associations using Bayesian networks. BMC Syst Biol 2012;6 Suppl 3:S14,0509-6-S3-S14. Epub 2012 Dec 17.

7. Slattery ML, Lundgreen A, Herrick JS, Wolff RK. Genetic variation in RPS6KA1, RPS6KA2, RPS6KB1, RPS6KB2, and PDK1 and risk of colon or rectal cancer. Mutat Res 2011 Jan 10;706(1-2):13-20.

8. Fu Z, Shrubsole MJ, Li G, Smalley WE, Hein DW, Chen Z, Shyr Y, Cai Q, Ness RM, Zheng W. Using gene-environment interaction analyses to clarify the role of well-done meat and heterocyclic amine exposure in the etiology of colorectal polyps. Am J Clin Nutr 2012 Nov;96(5):1119-28.

9. Campos-Vega R, Garcia-Gasca T, Guevara-Gonzalez R, Ramos-Gomez M, Oomah BD, Loarca-Pina G. Human gut flora-fermented nondigestible fraction from cooked bean (phaseolus vulgaris L.) modifies protein expression associated with apoptosis, cell cycle arrest, and proliferation in human adenocarcinoma colon cancer cells. J Agric Food Chem 2012 Dec 26;60(51):12443-50.

10. Hutter CM, Slattery ML, Duggan DJ, Muehling J, Curtin K, Hsu L, Beresford SA, Rajkovic A, Sarto GE, Marshall JR, et al. Characterization of the association between 8q24 and colon cancer: Gene-environment exploration and meta-analysis. BMC Cancer 2010 Dec 4;10:670,2407-10-670.

11. Patwardhan RP, Hiatt JB, Witten DM, Kim MJ, Smith RP, May D, Lee C, Andrie JM, Lee SI, Cooper GM, et al. Massively parallel functional dissection of mammalian enhancers *in vivo*. Nat Biotechnol 2012 Feb 26;30(3):265-70.

12. Miller DT, Adam MP, Aradhya S, Biesecker LG, Brothman AR, Carter NP, Church DM, Crolla JA, Eichler EE, Epstein CJ, et al. Consensus statement: Chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. Am J Hum Genet 2010 May 14;86(5):749-64.

13. Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, Yamrom B, Yoon S, Krasnitz A, Kendall J, et al. Strong association of de novo copy number mutations with autism. Science 2007 Apr 20;316(5823):445-9.

14. Walsh T, McClellan JM, McCarthy SE, Addington AM, Pierce SB, Cooper GM, Nord AS, Kusenda M, Malhotra D, Bhandari A, et al. Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. Science 2008 Apr 25;320(5875):539-43.

15. Fialho AM, Chakrabarty AM. Patent controversies and court cases: Cancer diagnosis, therapy and prevention. Cancer Biol Ther 2012 Nov;13(13):1229-34.

16. Whiting DR, Guariguata L, Weil C, Shaw J. IDF diabetes atlas: Global estimates of the prevalence of diabetes for 2011 and 2030. Diabetes Res Clin Pract 2011 Dec;94(3):311-21.

17. Akey JM, Eberle MA, Rieder MJ, Carlson CS, Shriver MD, Nickerson DA, Kruglyak L. Population history and natural selection shape patterns of genetic variation in 132 genes. PLoS Biol 2004 Oct;2(10):e286.

18. Chermet J, Kieffer E, Taboury J, Monnier JP, Chalut J. Atypical angiographic appearances of aneurisms of the abdominal aorta (author's transl). J Radiol Electrol Med Nucl 1976 Oct;10(57):733-45.

19. Helgason A, Palsson S, Thorleifsson G, Grant SF, Emilsson V, Gunnarsdottir S, Adeyemo A, Chen Y, Chen G, Reynisdottir I, et al. Refining the impact of TCF7L2 gene variants on type 2 diabetes and adaptive evolution. Nat Genet 2007 Feb;39(2):218-25.

20. Sanna S, Li B, Mulas A, Sidore C, Kang HM, Jackson AU, Piras MG, Usala G, Maninchedda G, Sassu A, et al. Fine mapping of five loci associated with low-density lipoprotein cholesterol detects variants that double the explained heritability. PLoS Genet 2011 Jul;7(7):e1002198.

21. Peters U, North KE, Sethupathy P, Buyske S, Haessler J, Jiao S, Fesinmeyer MD, Jackson RD, Kuller LH, Rajkovic A, et al. A systematic mapping approach of 16q12.2/FTO and BMI in more than 20,000 African Americans narrows in on the underlying functional variation: Results from the population architecture using genomics and epidemiology (PAGE) study. PLoS Genet 2013;9(1):e1003171.

22. Gong J, Schumacher F, Lim U, Hindorff LA, Haessler J, Buyske S, Carlson CS, Rosse S, Buzkova P, Fornage M, et al. Fine mapping and identification of BMI loci in African Americans. Am J Hum Genet 2013 Oct 3;93(4):661-71.

23. Wu Y, Waite LL, Jackson AU, Sheu WH, Buyske S, Absher D, Arnett DK, Boerwinkle E, Bonnycastle LL, Carty CL, et al. Trans-ethnic fine-mapping of lipid loci identifies population-

specific signals and allelic heterogeneity that increases the trait variance explained. PLoS Genet 2013 Mar;9(3):e1003379.

24. Buyske S, Wu Y, Carty CL, Cheng I, Assimes TL, Dumitrescu L, Hindorff LA, Mitchell S, Ambite JL, Boerwinkle E, et al. Evaluation of the metabochip genotyping array in African Americans and implications for fine mapping of GWAS-identified loci: The PAGE study. PLoS One 2012;7(4):e35651.

25. Voight BF, Scott LJ, Steinthorsdottir V, Morris AP, Dina C, Welch RP, Zeggini E, Huth C, Aulchenko YS, Thorleifsson G, et al. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. Nat Genet 2010 Jul;42(7):579-89.

26. Teo YY, Small KS, Kwiatkowski DP. Methodological challenges of genome-wide association analysis in Africa. Nat Rev Genet 2010 Feb;11(2):149-60.

27. Sim X, Ong RT, Suo C, Tay WT, Liu J, Ng DP, Boehnke M, Chia KS, Wong TY, Seielstad M, et al. Transferability of type 2 diabetes implicated loci in multi-ethnic cohorts from southeast Asia. PLoS Genet 2011 Apr;7(4):e1001363.

28. Matise TC, Ambite JL, Buyske S, Carlson CS, Cole SA, Crawford DC, Haiman CA, Heiss G, Kooperberg C, Marchand LL, et al. The next PAGE in understanding complex traits: Design for the analysis of Population Architecture using Genetics and Epidemiology (PAGE) study. Am J Epidemiol 2011 Oct 1;174(7):849-59.

29. Patterson N, Price AL, Reich D. Population structure and Eigenanalysis. PLoS Genet 2006 Dec;2(12):e190.

30. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 2006 Aug;38(8):904-9.

31. Weale ME. Quality control for genome-wide association studies. Methods Mol Biol 2010;628:341-72.

32. Willer CJ, Li Y, Abecasis GR. METAL: Fast and efficient meta-analysis of genome-wide association scans. Bioinformatics 2010 Sep 1;26(17):2190-1.

33. 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. An integrated map of genetic variation from 1,092 human genomes. Nature 2012 Nov 1;491(7422):56-65.

34. Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, Boehnke M, Abecasis GR, Willer CJ. LocusZoom: Regional visualization of genome-wide association scan results. Bioinformatics 2010 Sep 15;26(18):2336-7.

35. Manning AK, Hivert MF, Scott RA, Grimsby JL, Bouatia-Naji N, Chen H, Rybin D, Liu CT, Bielak LF, Prokopenko I, et al. A genome-wide approach accounting for body mass index

identifies genetic variants influencing fasting glycemic traits and insulin resistance. Nat Genet 2012 May 13;44(6):659-69.

36. Ward LD, Kellis M. HaploReg: A resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res 2012 Jan;40(Database issue):D930-4.

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

38. Natarajan A, Yardimci GG, Sheffield NC, Crawford GE, Ohler U. Predicting cell-type-specific gene expression from regions of open chromatin. Genome Res 2012 Sep;22(9):1711-22.

39. ENCODE Project Consortium, Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M. An integrated encyclopedia of DNA elements in the human genome. Nature 2012 Sep 6;489(7414):57-74.

40. Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, Bernstein BE, Bickel P, Brown JB, Cayting P, et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res 2012 Sep;22(9):1813-31.

41. Cartegni L, Wang J, Zhu Z, Zhang MQ, Krainer AR. ESEfinder: A web resource to identify exonic splicing enhancers. Nucleic Acids Res 2003 Jul 1;31(13):3568-71.

42. Freedman ML, Monteiro AN, Gayther SA, Coetzee GA, Risch A, Plass C, Casey G, De Biasi M, Carlson C, Duggan D, et al. Principles for the post-GWAS functional characterization of cancer risk loci. Nat Genet 2011 Jun;43(6):513-8.

43. Sanyal A, Lajoie BR, Jain G, Dekker J. The long-range interaction landscape of gene promoters. Nature 2012 Sep 6;489(7414):109-13.

44. Schadt EE, Molony C, Chudin E, Hao K, Yang X, Lum PY, Kasarskis A, Zhang B, Wang S, Suver C, et al. Mapping the genetic architecture of gene expression in human liver. PLoS Biol 2008 May 6;6(5):e107.

45. Gibbs JR, van der Brug MP, Hernandez DG, Traynor BJ, Nalls MA, Lai SL, Arepalli S, Dillman A, Rafferty IP, Troncoso J, et al. Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. PLoS Genet 2010 May 13;6(5):e1000952.

46. Karolchik D, Hinrichs AS, Kent WJ. The UCSC genome browser. Curr Protoc Bioinformatics 2012 Dec;Chapter 1:Unit1.4.

47. Pang KC, Stephen S, Dinger ME, Engstrom PG, Lenhard B, Mattick JS. RNAdb 2.0--an expanded database of mammalian non-coding RNAs. Nucleic Acids Res 2007 Jan;35(Database issue):D178-82.

48. Kozomara A, Griffiths-Jones S. miRBase: Integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res 2011 Jan;39(Database issue):D152-7.

49. Friedman RC, Farh KK, Burge CB, Bartel DP. Most mammalian mRNAs are conserved targets of microRNAs. Genome Res 2009 Jan;19(1):92-105.

50. GTEx Consortium. The genotype-tissue expression (GTEx) project. Nat Genet 2013 Jun;45(6):580-5.

51. Felsenstein J, Churchill GA. A hidden markov model approach to variation among sites in rate of evolution. Mol Biol Evol 1996 Jan;13(1):93-104.

52. Wang H, Liu L, Zhao J, Cui G, Chen C, Ding H, Wang DW. Large scale meta-analyses of fasting plasma glucose raising variants in GCK, GCKR, MTNR1B and G6PC2 and their impacts on type 2 diabetes mellitus risk. PLoS One 2013 Jun 28;8(6):e67665.

53. Scott RA, Lagou V, Welch RP, Wheeler E, Montasser ME, Luan J, Magi R, Strawbridge RJ, Rehnberg E, Gustafsson S, et al. Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. Nat Genet 2012 Sep;44(9):991-1005.

54. Saxena R, Hivert MF, Langenberg C, Tanaka T, Pankow JS, Vollenweider P, Lyssenko V, Bouatia-Naji N, Dupuis J, Jackson AU, et al. Genetic variation in GIPR influences the glucose and insulin responses to an oral glucose challenge. Nat Genet 2010 Feb;42(2):142-8.

55. Drucker DJ. The role of gut hormones in glucose homeostasis. J Clin Invest 2007 Jan;117(1):24-32.

56. Drucker DJ, Philippe J, Mojsov S, Chick WL, Habener JF. Glucagon-like peptide I stimulates insulin gene expression and increases cyclic AMP levels in a rat islet cell line. Proc Natl Acad Sci U S A 1987 May;84(10):3434-8.

57. Pare G, Chasman DI, Parker AN, Nathan DM, Miletich JP, Zee RY, Ridker PM. Novel association of HK1 with glycated hemoglobin in a non-diabetic population: A genome-wide evaluation of 14,618 participants in the women's genome health study. PLoS Genet 2008 Dec;4(12):e1000312.

58. Saxena R, Saleheen D, Been LF, Garavito ML, Braun T, Bjonnes A, Young R, Ho WK, Rasheed A, Frossard P, et al. Genome-wide association study identifies a novel locus contributing to type 2 diabetes susceptibility in Sikhs of Punjabi origin from India. Diabetes 2013 May;62(5):1746-55.

59. Sparso T, Andersen G, Nielsen T, Burgdorf KS, Gjesing AP, Nielsen AL, Albrechtsen A, Rasmussen SS, Jorgensen T, Borch-Johnsen K, et al. The GCKR rs780094 polymorphism is associated with elevated fasting serum triacylglycerol, reduced fasting and OGTT-related insulinaemia, and reduced risk of type 2 diabetes. Diabetologia 2008 Jan;51(1):70-5.

60. Lyssenko V, Nagorny CL, Erdos MR, Wierup N, Jonsson A, Spegel P, Bugliani M, Saxena R, Fex M, Pulizzi N, et al. Common variant in MTNR1B associated with increased risk of type 2 diabetes and impaired early insulin secretion. Nat Genet 2009 Jan;41(1):82-8.

61. Sakai K, Imamura M, Tanaka Y, Iwata M, Hirose H, Kaku K, Maegawa H, Watada H, Tobe K, Kashiwagi A, et al. Replication study for the association of 9 east asian GWAS-derived loci with susceptibility to type 2 diabetes in a Japanese population. PLoS One 2013 Sep 25;8(9):e76317.

62. Warner JP, Leek JP, Intody S, Markham AF, Bonthron DT. Human glucokinase regulatory protein (GCKR): CDNA and genomic cloning, complete primary structure, and chromosomal localization. Mamm Genome 1995 Aug;6(8):532-6.

63. Martin CC, Bischof LJ, Bergman B, Hornbuckle LA, Hilliker C, Frigeri C, Wahl D, Svitek CA, Wong R, Goldman JK, et al. Cloning and characterization of the human and rat islet-specific glucose-6-phosphatase catalytic subunit-related protein (IGRP) genes. J Biol Chem 2001 Jul 6;276(27):25197-207.

64. Rose CS, Ek J, Urhammer SA, Glumer C, Borch-Johnsen K, Jorgensen T, Pedersen O, Hansen T. A -30G>A polymorphism of the beta-cell-specific glucokinase promoter associates with hyperglycemia in the general population of whites. Diabetes 2005 Oct;54(10):3026-31.

65. Dupuis J, Langenberg C, Prokopenko I, Saxena R, Soranzo N, Jackson AU, Wheeler E, Glazer NL, Bouatia-Naji N, Gloyn AL, et al. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. Nat Genet 2010 Feb;42(2):105-16.

66. Baerenwald DA, Bonnefond A, Bouatia-Naji N, Flemming BP, Umunakwe OC, Oeser JK, Pound LD, Conley NL, Cauchi S, Lobbens S, et al. Multiple functional polymorphisms in the G6PC2 gene contribute to the association with higher fasting plasma glucose levels. Diabetologia 2013 Jun;56(6):1306-16.

67. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A 2009 Jun 9;106(23):9362-7.

68. A Catalog of Published Genome-Wide Association Studies, Available at: www.genome.gov/gwastudies. [Internet] [cited 2013 .

69. Carlson CS, Heagerty PJ, Nord AS, Pritchard DK, Ranchalis J, Boguch JM, Duan H, Hatsukami TS, Schwartz SM, Rieder MJ, et al. TagSNP evaluation for the association of 42 inflammation loci and vascular disease: Evidence of IL6, FGB, ALOX5, NFKBIA, and IL4R loci effects. Hum Genet 2007 Mar;121(1):65-75.

70. Peters U, North KE, Sethupathy P, Buyske S, Haessler J, Jiao S, Fesinmeyer MD, Jackson RD, Kuller LH, Rajkovic A, et al. A systematic mapping approach of 16q12.2/FTO and BMI in more than 20,000 african americans narrows in on the underlying functional variation: Results from the population architecture using genomics and epidemiology (PAGE) study. PLoS Genet 2013;9(1):e1003171.

71. Tomlinson IP, Carvajal-Carmona LG, Dobbins SE, Tenesa A, Jones AM, Howarth K, Palles C, Broderick P, Jaeger EE, Farrington S, et al. Multiple common susceptibility variants near BMP pathway loci GREM1, BMP4, and BMP2 explain part of the missing heritability of colorectal cancer. PLoS Genet 2011 Jun;7(6):e1002105.

72. Ong CT, Corces VG. Enhancer function: New insights into the regulation of tissue-specific gene expression. Nat Rev Genet 2011 Apr;12(4):283-93.

73. McLean CY, Reno PL, Pollen AA, Bassan AI, Capellini TD, Guenther C, Indjeian VB, Lim X, Menke DB, Schaar BT, et al. Human-specific loss of regulatory DNA and the evolution of human-specific traits. Nature 2011 Mar 10;471(7337):216-9.

74. Moore MJ. From birth to death: The complex lives of eukaryotic mRNAs. Science 2005 Sep 2;309(5740):1514-8.

75. Garneau NL, Wilusz J, Wilusz CJ. The highways and byways of mRNA decay. Nat Rev Mol Cell Biol 2007 Feb;8(2):113-26.

76. Sharon E, Kalma Y, Sharp A, Raveh-Sadka T, Levo M, Zeevi D, Keren L, Yakhini Z, Weinberger A, Segal E. Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. Nat Biotechnol 2012 May 20;30(6):521-30.

77. Holmqvist E, Reimegard J, Wagner EG. Massive functional mapping of a 5'-UTR by saturation mutagenesis, phenotypic sorting and deep sequencing. Nucleic Acids Res 2013 Jul 1;41(12):e122.

78. Patwardhan RP, Hiatt JB, Witten DM, Kim MJ, Smith RP, May D, Lee C, Andrie JM, Lee SI, Cooper GM, et al. Massively parallel functional dissection of mammalian enhancers in vivo. Nat Biotechnol 2012 Feb 26;30(3):265-70.

79. Melnikov A, Murugan A, Zhang X, Tesileanu T, Wang L, Rogov P, Feizi S, Gnirke A, Callan CG,Jr, Kinney JB, et al. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. Nat Biotechnol 2012 Feb 26;30(3):271-7.

80. Haberle V, Lenhard B. Dissecting genomic regulatory elements in vivo. Nat Biotechnol 2012 Jun 7;30(6):504-6.

81. Ward RJ, Alvarez-Curto E, Milligan G. Using the flp-in T-rex system to regulate GPCR expression. Methods Mol Biol 2011;746:21-37.

82. McGuire SE, McGuire AL. Don't throw the baby out with the bathwater: Enabling a bottom-up approach in genome-wide association studies. Genome Res 2008 Nov;18(11):1683-5.

83. Beskow LM, Linney KN, Radtke RA, Heinzen EL, Goldstein DB. Ethical challenges in genotype-driven research recruitment. Genome Res 2010 Jun;20(6):705-9.

84. Chulada PC, Vainorius E, Garantziotis S, Burch LH, Blackshear PJ, Zeldin DC. The environmental polymorphism registry: A unique resource that facilitates translational research of environmental disease. Environ Health Perspect 2011 Nov;119(11):1523-7.

85. Beskow LM, Namey EE, Miller PR, Nelson DK, Cooper A. IRB chairs' perspectives on genotype-driven research recruitment. IRB 2012 May-Jun;34(3):1-10.

86. Beskow LM, Namey EE, Cadigan RJ, Brazg T, Crouch J, Henderson GE, Michie M, Nelson DK, Tabor HK, Wilfond BS. Research participants' perspectives on genotype-driven research recruitment. J Empir Res Hum Res Ethics 2011 Dec;6(4):3-20.

87. Beskow LM, Fullerton SM, Namey EE, Nelson DK, Davis AM, Wilfond BS. Recommendations for ethical approaches to genotype-driven research recruitment. Hum Genet 2012 Sep;131(9):1423-31.

88. National Heart, Lung, and Blood Institute working group, Fabsitz RR, McGuire A, Sharp RR, Puggal M, Beskow LM, Biesecker LG, Bookman E, Burke W, Burchard EG, et al. Ethical and practical guidelines for reporting genetic research results to study participants: Updated guidelines from a national heart, lung, and blood institute working group. Circ Cardiovasc Genet 2010 Dec;3(6):574-80.

89. Lemke AA, Halverson C, Ross LF. Biobank participation and returning research results: Perspectives from a deliberative engagement in south side chicago. Am J Med Genet A 2012 May;158A(5):1029-37.

90. Greely HT. The uneasy ethical and legal underpinnings of large-scale genomic biobanks. Annu Rev Genomics Hum Genet 2007;8:343-64.

91. Cadigan RJ, Michie M, Henderson G, Davis AM, Beskow LM. The meaning of genetic research results: Reflections from individuals with and without a known genetic disorder. J Empir Res Hum Res Ethics 2011 Dec;6(4):30-40.

92. Tabor HK, Brazg T, Crouch J, Namey EE, Fullerton SM, Beskow LM, Wilfond BS. Parent perspectives on pediatric genetic research and implications for genotype-driven research recruitment. J Empir Res Hum Res Ethics 2011 Dec;6(4):41-52.

93. Namey EE, Beskow LM. Epilepsy patient-participants and genetic research results as "answers". J Empir Res Hum Res Ethics 2011 Dec;6(4):21-9.

94. Beskow LM, Smolek SJ. Prospective biorepository participants' perspectives on access to research results. J Empir Res Hum Res Ethics 2009 Sep;4(3):99-111.

95. Hsieh HF, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005 Nov;15(9):1277-88.

96. Dedoose: web application for managing, analyzing, and presenting qualitative and mixed method research data [Internet] Version 4.5. Los Angeles, CA: SocioCultural Research Consultants, LLC [cited 2013]. Available from: www.dedoose.com.