8-2015

Detection Of Genes Influencing Chronic And Mendelian Disease Via Loss-Of-Function Variation

Alexander H. Li

DETECTION OF GENES INFLUENCING CHRONIC AND MENDELIAN DISEASE

VIA LOSS-OF-FUNCTION VARIATION

by

Alexander Hung Li, M.S.

APPROVED:

_____

Eric Boerwinkle, Ph.D.

Supervisory Professor

_____

Hope Northrup, M.D.

_____

Laura Mitchell, Ph.D.

_____

Paul Scheet, Ph.D.

_____

Wenyi Wang, Ph.D.

APPROVED:

_____

Dean, The University of Texas

Graduate School of Biomedical Sciences at Houston

DETECTION OF GENES INFLUCING CHRONIC AND MENDELIAN DISEASE VIA LOSS-OF-FUNCTION VARIATION

A

DISSERTATION

Presented to the Faculty of

The University of Texas

Health Science Center at Houston

and

The University of Texas

MD Anderson Cancer Center

Graduate School of Biomedical Sciences

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

by

Alexander Hung Li, M.S.

Houston, Texas

August, 2015

Acknowledgements

I would like to acknowledge my wife, family, and friends for their personal support. I also thank my advisor, committee members, faculty, and students at the Human Genetics Center and Human and Molecular Genetics program for lending their expertise to this work. Finally, I would like to thank the participants in the Atherosclerosis Risk in Communities study, patients recruited from Texas Children's Hospital and other study participants included in this dissertation without whom none of this work would have been possible.

DETECTION OF GENES INFLUCING CHRONIC AND MENDELIAN DISEASE VIA LOSS-OF-FUNCTION VARIATION

Alexander Hung Li, M.S.

Advisory Professor: Eric Boerwinkle, Ph.D.

A typical human exome harbors dozens of loss-of-function (LOF) variants predicted to severely disrupt or abolish gene function. These variants are enriched at the extremely rare end of the allele frequency spectrum (< 0.1%), suggesting purifying selection against these sites. However, most previous population-based sequencing studies have not included analysis of genotype-phenotype relationships with LOF variants. Thus, the contribution of LOF variation to health and disease within the general population remains largely uncharacterized.

Using whole exome sequence from 8,554 participants in the Atherosclerosis Risk in Communities (ARIC) study, we explored the impact of LOF variation on a broad spectrum of human phenotypes. First, we selected 20 common chronic disease risk factor phenotypes and performed gene-based association tests. Analysis of this sample verified two relationships in well-studied genes (*PCSK9* and *APOC3*) and identified eight new loci. Novel relationships included elevated fasting glucose in heterozygous carriers of LOF variation in *TXNDC5*, which encodes a biomarker for type 1 diabetes progression, and apparent recessive effects of *C1QTNF8* on serum magnesium levels. Next, we explored the effect of LOF variation on 308 small molecular metabolites, observing 8 significant genotype-phenotype associations. We highlight the relationship between serum histidine and *HAL,* a gene essential to histidine

catabolism, demonstrating the biologically interpretability of associations with molecular metabolite targets. Finally, we explore the impact of LOF variation on a rare birth defect by comparing sequence from 342 unrelated left ventricular outflow tract obstruction (LVOTO) cases to ARIC sequence, identifying genes harboring case-exclusive LOF mutations. Comparison to an *a priori* list of cardiac candidate genes revealed 28 genes potentially related to LVOTO, including 22 not previously associated with a human disorder. Genotype validation in these samples revealed diverse inheritance patterns, including 9 confirmed de novo variants (*ACVR1*, *JARID2*, *KMT2D*, *NF1*, *NR2F2*, *PLRG1, SMURF1, TBX20*, and *ZEB2*).

The analytical strategy presented here highlights the role of biologically-informed annotation on large-scale human genetic studies. The genes identified by these methods may have applications in disease prediction and drug development, and future genome studies will continue to refine our understanding of the scope of genetic variation affecting human health and disease.

Table of Contents

List of Figures

List of Tables

Abbreviations

LOF – Loss of Function

WES – Whole exome sequence

WGS – Whole genome sequence

MAF – Minor allele frequency

MAC – Minor allele count

ARIC – Atherosclerosis Risk in Communities

DNS – Damaging nonsynonymous

LVOTO – Left ventricular outflow tract obstruction

CVM – Cardiovascular malformation

Chapter 1: Introduction

Genetics, health and disease

Genomic information is emerging as a valuable addition to traditional health care and an essential component of the National Institutes of Health's precision medicine initiative[1]. Knowing and understanding the content of an individual's genome can inform a personalized disease risk evaluation, since one's DNA sequence is generally stable throughout a lifetime, with the exception of somatic mutation and epigenetic modifications. For example, the content of a genome may serve as biomarkers for late-onset conditions such as Alzheimer's disease[2], common chronic conditions such as coronary artery disease[3], or sensitivity to the anticoagulant warfarin[4]. The genetic profile of an individual can be used to develop individualized management strategies, such as informing clinicians which drugs may be most effective for patients[5]. At the rare end of the disease spectrum, gene sequencing has also been used to identify specific individuals having increased risk to develop clinical disease, especially when informed by family history for a Mendelian disorder[6].

Genetic studies may also point to attractive drug targets, especially when associated with a beneficial health effect. For example, six pharmaceutical companies are actively developing *PCSK9* inhibitors[7], two of which have entered phase III clinical trials, showing great promise for lowering lipids and preventing major cardiovascular events[8,9]. The common paradigm of these drugs is to downregulate levels of single gene product within a patient. Recent studies have suggested that other genes, such as APOC3[10], may also make suitable targets for similar drugs which act via a downregulation pathways to achieve a protective cardiovascular effect.

It is therefore of great interest to both clinicians and pharmaceutical companies to understand the major genes that contribute to human health. This demand drives researchers to discover novel associations and understand the contribution of these individual genes to human

phenotypes. Next generation sequencing technology, especially whole-exome sequencing, has emerged as a powerful and efficient tool to capture gene sequence which is a crucial step in this discovery process.

Sequencing platforms for gene discovery

Whole exome sequencing (WES) has emerged as a cost-effective platform suitable for both clinical diagnoses and research discovery. Over a dozen companies offer commercial whole exome sequencing services (https://www.scienceexchange.com/services/whole-exome-seq) with costs ranging from$445 to $1,535 per sample. Per-sample costs within high-throughput sequencing centers may be even lower. Another advantage of WES is the ability to capture sequence variation from across the allelic frequency spectrum, from common to rare or even private sites. This contrasts with less expensive chip-based genotyping for gene discovery, which may only provide information in a limited allele frequency spectrum or a slice of the human genome (eg, loci previously characterized such as those ascertained by GWAS). These chip-based platforms also target known variant sites and thus cannot be used for novel variant discovery, an important aspect as we focus more on ethnically-diverse or isolated populations. Additionally, the near-comprehensive gene coverage provided by WES is an advantage over targeted sequencing methods, since these methods rely on *a priori* selection of gene candidates which limits the potential for novel discovery. Finally, while whole genome sequencing provides both comprehensive coverage and high-resolution detection of mutations, this platform also has disadvantages compared to WES. For many applications whole genome sequence data remains cost-prohibitive to generate and store. In addition, the interpretation of intergenic genomic regions remains a challenge in the context of human disease studies, whereas the sequence of exons is more readily interpretable

Within limits, exonic variants can be interpreted in the context of established molecular biology paradigms. WES reliably ascertains and genotypes both single-nucleotide substitutions and small insertions/deletions (up to 50bp) within the protein-encoding regions of genes. Since these variants are in the exons of genes, the functional effects on mRNA splicing and protein translation are predictable. A number of tools exist to predict the effect of protein-altering nucleotide substitutions (e.g. Polyphen2[11], SIFT[12]), incorporating information from sequence conservation, local amino acid context, and predicted protein structures that arise from these sequence changes. However, there is no single consensus on which method performs best at predicting pathogenicity, and there is poor correlation between the results of these tools[13]. Given the difficulties of predicting the pathogenicity of protein-altering variation, we decided to prioritize other functional classes of human variation.

Loss-of-function mechanism

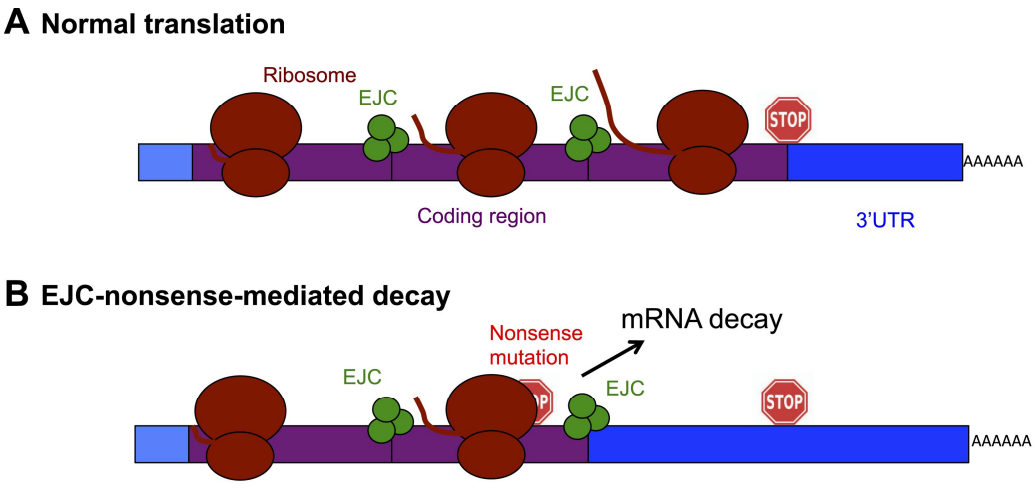
Loss-of-function (LOF) variants are sequence changes that are predicted to severely disrupt or even completely prevent the formation of protein from gene templates. True LOF variants should be contrasted to predicted hypomorphs, which may influence function by changes in amino acid context rather than overall gene levels. Several functional categories of LOF variation exist: premature stop, splice, and frameshift indel. These diverse functional categories of mutation converge on their potential to introduce premature stop (nonsense) signals into mRNA transcripts, which are targeted for degradation by highly conserved molecular pathways[14].

The various types of LOF variants can all trigger specific reactions from eukaryotic mRNA surveillance mechanisms, specifically the nonsense-mediated mRNA decay pathway (NMD). This pathway monitors gene expression and identifies transcripts containing premature

termination codons. During normal transcription in the nucleus of cells, introns are spliced out and adjacent exons are joined into a single contiguous mRNA transcript. Exon-junction complex (EJC) factors contribute to this process and remain bound to the mature mRNA transcript as it is exported into the cytoplasm, marking the junction where two exons have been joined[15] (**Figure 1.1a**). Since a transcript is expected to only contain one stop codon to indicate the end of protein translation, no EJC factors should be bound to the transcript downstream of this termination signal. The mammalian NMD pathway consists of core machinery which interacts with EJC factors[16]. When DNA mutations alter gene sequence to encode for premature stop codons in the non-terminal gene exon, mRNA-bound EJC factors will be detected downstream of these termination signals. The presence of these aberrant EJC complexes factors act as a "second signal" (ie, extra stop codon) when mRNA is proofread during translation which destabilize the transcript and elicit NMD[17] (**Figure 1.1b**).

LOF variants predicted to trigger NMD can be detected and annotated within WES data. Single nucleotide substitutions giving rise to premature stop codons can be directly annotated, and are likely to trigger NMD if they are not in the terminal exon[18]. Similarly, frameshift indels disrupting the 3-bp reading frame of all downstream codons often lead to low levels of gene transcript[19]. Mutations disrupting essential splice motifs also trigger NMD, as errors in RNA splicing often trigger NMD[20]. However, certain protein-coding genes may resist NMD, and these must be taken into account during variant annotation. As expected, intronless genes are resistant to EJC-mediated NMD[21,22]. Human beta-globin genes may escape NMD if a

Figure 1.1: Loss-of-Function mechanism. Figure depicts the role of the EJC in a form of NMD surveillance by which mRNA transcripts with premature stop codons are targets for degradation by the cell. (A) Normal mRNA transcript which will be translated into protein, with a single stop signal in the terminal exon. (B) Mutant mRNA transcript with LOF mutation (exon 2 and exon 3), whereby the premature stop signal redirects this transcript to be degraded before protein translation can occur. Image adapted from eLife 2014;3:e04300.



premature stop signal arises very early in the transcript (within exon 1) and sequences allowing re-initiation of translation reside at a downstream alternate start codon[23].

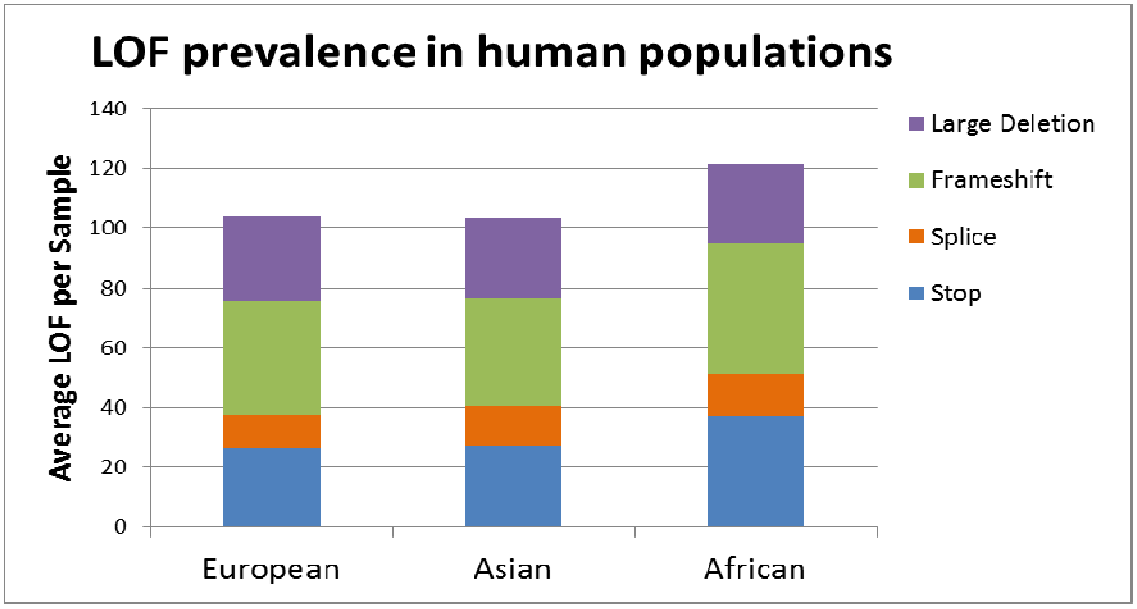
Loss of Function variation in health and disease

The prevalence of LOF variation in general human populations has been described[24]. A normal, healthy individual carries approximately 100 genes with variants predicted to have an LOF effect (20 homozygous), with varying rates between ancestry groups (**Figure 1.2**). Interestingly, these variants are enriched at the rare end of the population allele frequency spectrum, suggesting selection against these sites[24]. The relative frequency of these sites across genes can also provide insights into genetic architecture of human traits. For example, at the population level, these sites are observed less frequently in genes known to cause autosomal dominant Mendelian disorders[25].

LOF mutations have also been implicated in human disease. In studies of rare Mendelian disorders, disease pathology is generally attributed to a single rare (often unique) mutation within a patient. Human mutation databases such as the Human Gene Mutation Database (HGMD, http://www.hgmd.cf.ac.uk/) and ClinVar (http://www.ncbi.nlm.nih.gov/clinvar/) curate and catalog the specific mutations reported to cause rare disease. HGMD (2015.1) describes 17,562 premature stop variants and 34,466 frameshift indels classified as disease-causing mutations. Similarly, the NCBI ClinVar database includes description of 10,222 pathogenic LOF variants (4,892 premature stop, 1,767 splice, and 3,561 frameshift indel).

The role of LOF variation in common disease has been less extensively studied, but early results have been encouraging. African American individuals with variation introducing premature stop codons in *PCSK9* present lower serum LDL cholesterol and decreased risk for coronary heart disease than non-LOF individuals[26]. These results have sparked an interest in

Figure 1.2: Prevalence of Loss-of-Function variation in human populations. This figure depicts the average number of function variation per sample, representing three different ancestry groups from the 1000 genomes project. Image adapted from the 1000 genomes project[24].



PCSK9 as a drug target to lower cholesterol, with multiple drugs targeting this gene in development. Similarly, individuals with LOF variation in *APOC3*, including splice and premature stop, present lower serum triglycerides a reduced risk for coronary heart disease compared to others individuals without similar variation in this gene[10].

Despite these encouraging results in population-based studies and the ubiquitous frequency of these variants within human populations, the effect of LOF variation on human phenotypes in the general population remains poorly characterized. In addition, novel LOF variants that contribute to rare disease are an important tool for novel gene discovery. Therefore, we set out to characterize the effect of LOF variation on broad spectrum of human phenotypes including common complex disease biomarkers, small molecule metabolite levels, and cohorts of patients with rare Mendelian disorders.

Chapter 2: Common chronic disease biomarkers

This chapter is based on: Li, A. H., Morrison, A. C., Kovar, C., Cupples, L. A., Brody, J. a, Polfus, L. M., Yu, B., Metcalf, G., Muzny, D., Veerereghavan, N., Liu, X., Lumley, T., Mosley, T. H., Gibbs, R. A., Boerwinkle, E. (2015). Analysis of loss-of-function variants and 20 risk factor phenotypes in 8,554 individuals identifies loci influencing chronic disease. *Nature genetics*, *47*(6), 640–642. doi:10.1038/ng.3270.

Introduction

Investigations of genotype-phenotype associations leading to novel gene discovery have traditionally been facilitated by focusing on the most severe or earliest age of onset cases[4]. An alternative approach would be to identify variants with the most severe functional effects in a sample of deeply-phenotyped individuals, and then investigate the role of these variants in health and disease. To test this approach, we sequenced the exomes of 8,554 individuals who have been measured for many phenotypes related to common chronic diseases, such as diabetes and coronary heart disease. We annotated predicted loss-of-function (LOF) variants in these individuals and investigated their impact on 20 chronic disease risk factor phenotypes. Gene-based analyses identified and replicated 10 genetic loci associated with these measured traits. These results demonstrate the importance of detailed biological annotation to inform large-scale sequencing studies, and the utility of deeply-phenotyped cohort studies to further elucidate the genetic architecture of human health and disease.

Whole exome sequencing was performed on 2,836 African-American and 5,718 European-American individuals from the Atherosclerosis Risk in Communities (ARIC) study (**Table 2.1**). Ninety percent of target sites were covered at 20x or greater (mean depth 110.1 per sample), revealing 1,911,892 total single nucleotide variants (SNV) with an average Ti/Tv of 3.3 per sample, and 38,219 small insertions and deletions (indels). Indel sizes ranged from -51 base pairs (bp) to +27 bp, with a mode of -1 bp. We defined LOF variation as sequence changes predicted to abolish protein formation from all RefSeq isoforms for a given gene and identified a total of 36,561 candidate LOF sites (13,783 frameshift indels, 8,772 splice, 14,006 premature stop, **Table 2.2**) in 11,260 protein-coding genes. Not surprisingly[7], LOF variants were enriched

Table 2.1: Overview of individuals included in this study. Baseline characteristics of African American (AA) and European American (EA) participants from the ARIC cohort are shown, including the total number of individuals undergoing whole exome sequence, age at enrollment, gender distribution and body mass index (BMI).

	Discovery		**Replication**	
Characteristic	**AA**	**EA**	**AA**	**EA**
Individuals	1,418	2,859	1,418	2,859
Age	53.06 $\pm$ 5.75	54.29 $\pm$ 5.68	53.3 $\pm$ 5.81	54.47 $\pm$ 5.65
Males (%)	505 (35.61%)	1,378 (48.19%)	520 (36.67%)	1,338 (46.79%)
BMI	29.87 $\pm$ 6.28	26.89 $\pm$ 4.71	29.71 $\pm$ 6.31	26.88 $\pm$ 4.74

Table 2.2: Number of LOF sites in the study sample and per individual. This table describes the total number of LOF sites observed, and the average number of heterozygous (homozygous in parentheses) LOF sites per individual.

	LOF sites			Average per individual	
	AA	EA	Combined	AA	EA
Stop	5,837	9,312	14,006	27.3 (2.1)	21.1 (2.2)
Splice	3,789	5,731	8,772	16.7 (1.9)	9.6 (1.8)
Frameshift	6,575	8,264	13,783	36.1 (4.4)	22.6 (3.1)
Total LOF	16,201	23,307	36,561	80.1(8.4)	53.3 (7.1)

in the very rare range of the site frequency spectrum (MAF < 0.1%) compared to other functional categories (**Figure 2.1**).

LOF OP ratio

We next characterized the prevalence of LOF variation by gene. Because mutations may arise more frequently in larger genes and codon usage influences the chance of premature stops, we exhaustively simulated every single nucleotide substitution within each gene transcript to determine the maximum number of potential LOF substitution sites within that gene, which we then compared to the observed number of LOF sites within our sample (observed number/potential number = OP ratio)[11,7]. Almost half the genes in our capture regions presented no LOF alleles (n = 7,115, OP ratio = 0). The OP ratios of the remaining genes form a distribution with a peak near 0.003 with a skewed right tail (**Figure 2.2a**), underscoring the role of purifying selection against these sites. Genes known to influence human phenotypes in a dominant manner[8] present smaller average OP ratios (**Figure 2.2b**), while known recessive disease genes[7] have larger OP ratios (**Figure 2.2c**). The relationship between the OP ratio and the effects of LOF variants on the 20 risk factor phenotypes analyzed here is complex. Clearly, genes lacking LOF variants (i.e. OP ratio = 0) are not contributing to the analysis. Conversely, genes that tolerate a large number of LOF variants and have a high OP ratio (e.g. OP ratio > 0.1) did not significantly contribute to phenotypic variation. Genes contributing to the genetic architecture of health and disease in the population are likely to be important, by virtue of having an above average OP ratio, but not so critical such that LOF variants lead to debilitating disease or are inconsistent with life. To this point, we observed that homologs of essential mouse genes[9] (lethal phenotypes) have smaller average OP ratios compared to non-essential phenotype-changing genes ($p < 10^{-6}$, Wilcoxon), and the latter have smaller OP ratios compared

Figure 2.1: Site frequency spectrum of four categories of exome variation.

The relative proportion of these functional categories is shown binned by allele frequency. LCP = low-confidence or partial LOF criteria are described in the Methods Summary; Non-syn = nonsynonymous; Syn = synonymous.

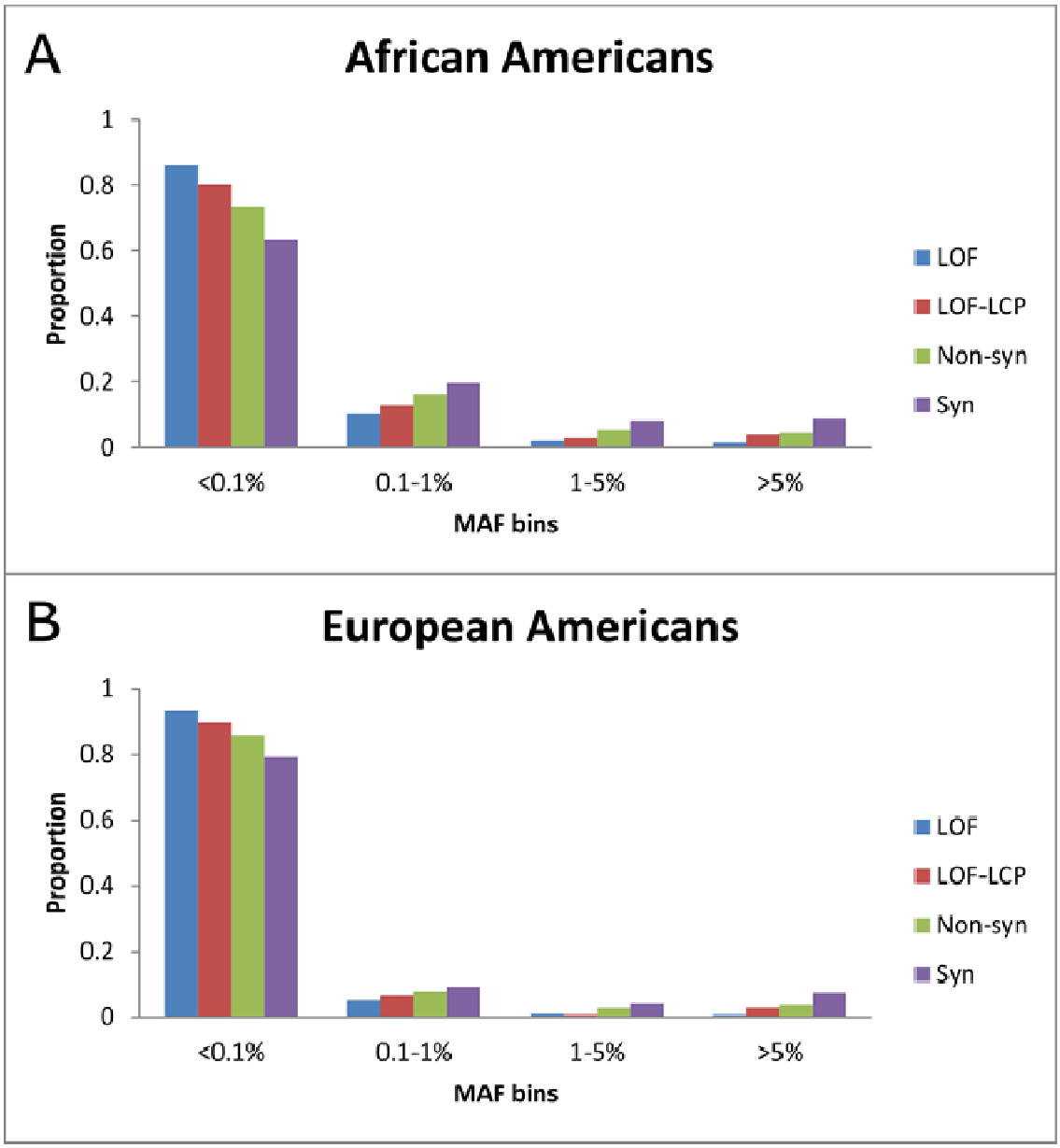
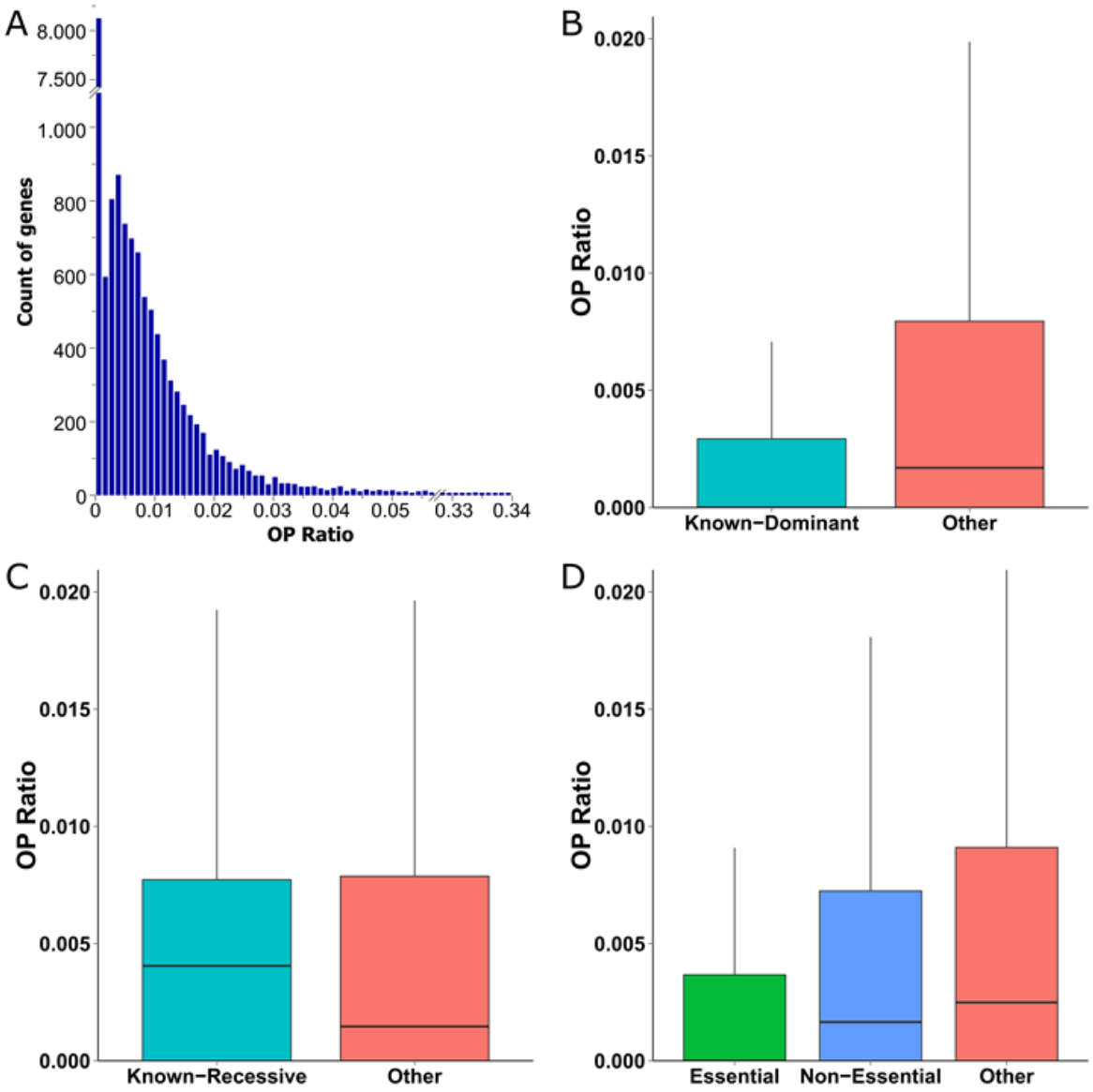


Figure 2.2: OP ratio trends across gene groups. (A) histogram of OP ratio for each gene; (B) lower OP ratio in genes causing dominant disorders (n = 248) vs other genes (n = 16,435), (C) higher OP ratio in genes causing recessive disorders(n = 652) vs other (n = 16,031), (D) lower OP ratio in human paralogs of essential mouse genes (Essential = embryonic lethal phenotype, n = 2,356; Nonessential = non-lethal phenotype, n = 3,520; Other = no phenotype reported, n = 10,807). Panels for B, C and D are boxplots denoting the median value, hinges at the 25th and 75th percentile, and whiskers extending to 1.5x inter-quartile range.



to all other genes ($p < 10^{-6}$, Wilcoxon) (**Figure 2.2d**). Genes with smaller OP ratios also tend to be stably expressed in more tissues and interact with more proteins (**Figure 2.3**).

Phenotype associations

To detect associations between LOF variation and common chronic disease phenotypes, we divided our sample into two non-overlapping discovery and replication groups each containing 4,277 individuals (**Table 2.1, Table 2.3**). Since LOF annotation enriches for variation with a similar predicted functional effect, namely the reduction or abolishment of protein formation, we grouped LOF variants by gene and performed a burden test for sites with MAF < 5% (T5 test)[10]. A summary of the most significant replicating results is shown in **Table 2.4**. As expected, LOF variants in *PCSK9* were associated with lower total cholesterol levels[11], and LOF variation in *APOC3* were associated with lower triglyceride levels[12]. We observed 8 novel relationships with compelling statistical evidence that replicated between the two samples (**Table 2.4**). Except for *PCSK9* and *APOC3*, the effects were in the direction thought to be increasing risk of disease.

Highlighting two examples, nine individuals with LOF variation in Thioredoxin Domain Containing 5 (*TXNDC5*) had elevated fasting blood glucose levels compared to non-LOF individuals (**Figure 2.4a**), and this gene has recently been suggested as a candidate for type 1 diabetes(T1D) risk[3]. In follow-up analyses, we observed a weak association between *TXNDC5* variation and fasting insulin levels within the ARIC study cohort ($p = 0.047$) (data not shown). In addition, five European-American study participants had a LOF mutation in *SEPT10* and these individuals had significantly reduced lung function (ratio of forced expiratory volume to

Figure 2.3: OP ratio trends across additional gene groups. (A) We selected the top 5% of genes from the egenetics database expressed in the most tissues, denoted as "Universally-expressed" (n=834) which we compared to all other genes (n=15,849). (B) We selected the top 5% of genes from the ConsensusPathDB database with the most protein interactions (n=834) which we compared to all other genes (n=15,849).

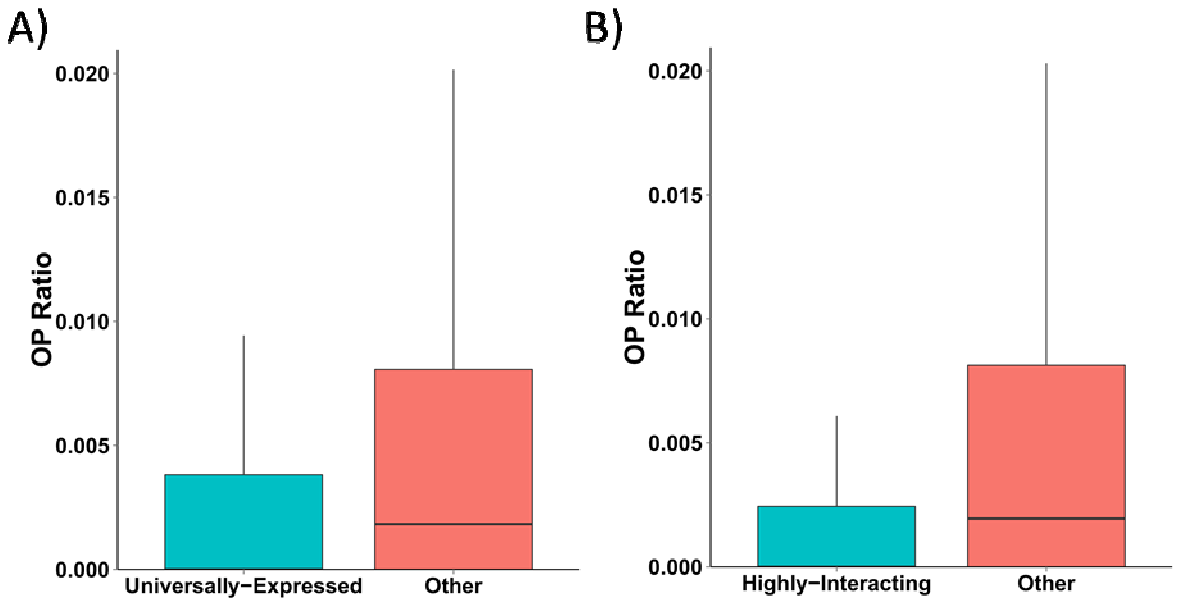


Figure 2.4: Distribution of phenotypes in LOF carriers. (A) Elevated fasting glucose in TXNDC5 LOF heterozygotes (n = 9) compared to individuals with no LOF variation in this gene (n = 8,545); (B) Elevated serum magnesium in C1QTNF8 homozygous individuals (n = 4) compared to LOF heterozygotes (n = 62) and non-LOF samples (n = 8,488). Both panels are boxplots denoting the median value, hinges at the 25th and 75th percentile, and whiskers extending to 1.5x inter-quartile range.

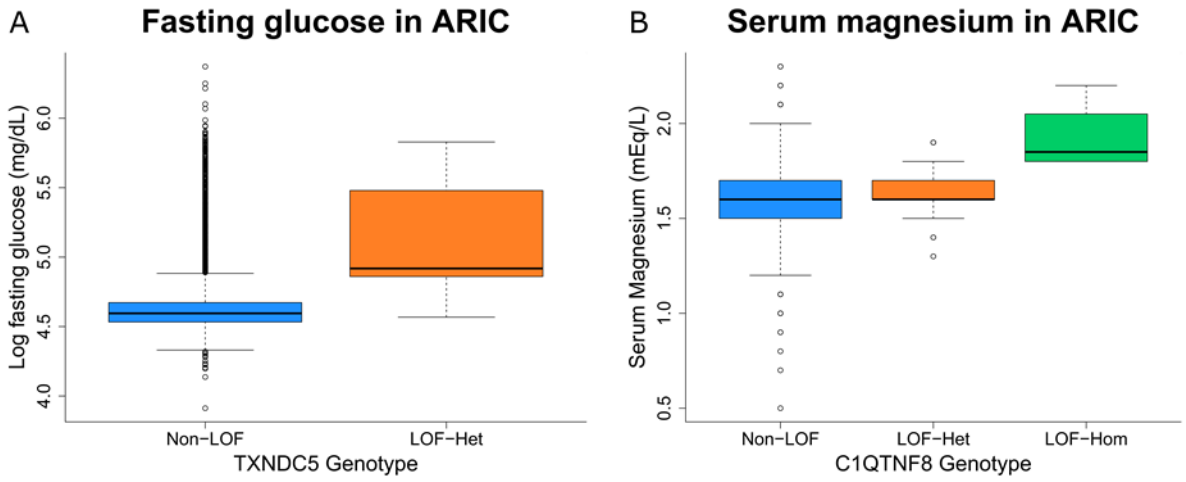


Table 2.3: List of phenotypes analyzed. This table describes with number of individuals (AA = African American; EA = European American) who were measured for each trait within the discovery and replication strata

			Discovery		Replication	
Category	Trait	Symbol	EA	AA	EA	AA
Electrolytes	Serum magnesium	Mg	2,858	1,369	2,859	1,377
	Serum phosphorus	P	2,858	1,369	2,859	1,377
	Serum calcium	Ca	2,858	1,369	2,859	1,377
	Serum potassium	K	2,858	1,369	2,859	1,377
	Serum sodium	Na	2,858	1,369	2,859	1,377
Liver enzymes	Aspartate transaminase	AST	2,315	934	2,306	926
	Alanine aminotransferase	ALT	1,411	807	1,327	797
	Gamma-glutamyl transpeptidase	GGT	1,413	807	1,329	797
Blood Pressure	Diastolic blood pressure	DBP	2,857	1,417	2,859	1,418
	Systolic blood pressure	SBP	2,858	1,417	2,859	1,418
Lung function	Forced vital capacity	FVC	2,850	1,397	2,855	1,397
	Forced expiratory volume/	FEV1FVC	2,850	1,397	2,852	1,396
Fatty acids	Serum triglycerides	TRG	2,854	1,337	2,856	1,359
	Total cholesterol	TCH	2,853	1,337	2,855	1,359
Diabetes	Fasting insulin	FI	2,784	1,273	2,804	1,284
	Fasting glucose	FG	2,784	1,273	2,804	1,283
Kidney	Creatinine	CRE	2,330	948	2,314	935
	Uric acid	UA	2,858	1,369	2,859	1,377
Other	White blood cell count	WBC	1,411	811	1,326	801
	Lactate	LAC	2,855	1,351	2,848	1,364

Table 2.4: Top gene-based phenotype associations which replicated. This table describes 10 significant associations which replicated and $\geq$ 3 individuals contributed to the genotype-phenotype association. "Genotype" denotes the heterozygous ("Het") or homozygous ("Hom") state of LOF individuals. "LOF sites" (snv, indel) describes the number of variants included for the T5 analyses. T5 betas were standardized ("Std Beta") by calculating the ratio of beta over the standard error. AA = African American; EA = European American; Disc = Discovery strata; Rep = Replication strata; Total = Discovery + Replication pooled.

Genotype	Trait	Gene	LOF sites	Ethnicity		T5 p-value			Std. Beta
				AA	EA	Disc.	Rep.	Total	
Het	Creatinine	*LHCGR*	2 (1,1)	0	3	6.71×10^{-6}	0.01	2.71×10^{-6}	4.69
		PLEKHG1	3 (1,2)	1	2	9.06×10^{-6}	3.0×10^{-3}	8.70×10^{-8}	5.35
	Fasting glucose	*GLIPR1*	3 (2,1)	1	2	6.14×10^{-4}	2.48×10^{-6}	9.38×10^{-9}	5.74
		TXNDC5	7 (4,3)	6	3	6.82×10^{-4}	5.75×10^{-5}	5.62×10^{-7}	5.00
	FEV1/FVC ratio	*SEPT10*	5 (1,4)	0	5	6.26×10^{-6}	1.21×10^{-4}	3.07×10^{-6}	–4.67
	Lactate	*WDR62*	3 (1,2)	3	0	8.0×10^{-3}	5.52×10^{-6}	1.91×10^{-6}	4.76
	Tot. cholesterol	*PCSK9*	6 (3,3)	24	2	8.27×10^{-5}	4.44×10^{-4}	5.25×10^{-8}	–5.44
	Triglycerides	*APOC3*	4 (3,1)	13	24	1.25×10^{-9}	1.38×10^{-8}	7.98×10^{-17}	–8.33
		TIGIT	2 (1,1)	2	1	2.74×10^{-4}	3.88×10^{-3}	4.11×10^{-6}	4.61
Homo	Magnesium	*C1QTNF8*	1 (1,0)	0	4	0.02	1.31x10–5	5.20×10^{-5}	4.08

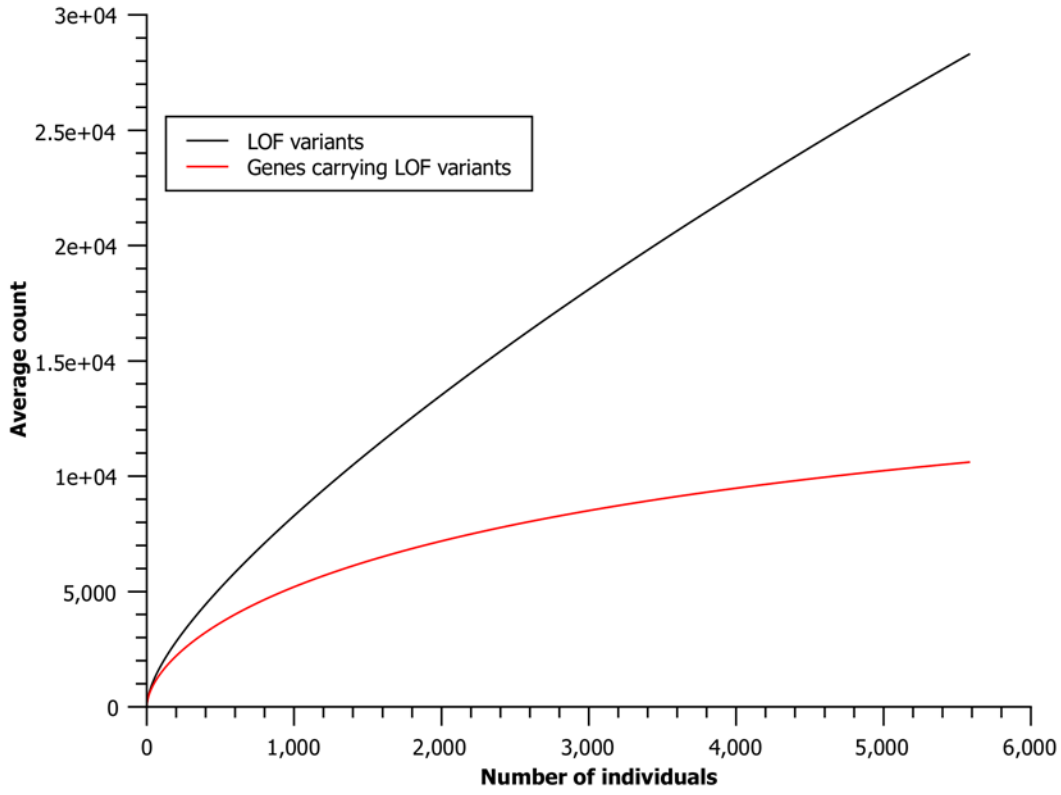
the forced vital capacity, FEV/FVC, $p=3.07 \times 10^{-6}$). *SEPT10* is contained with a known linkage peak for nicotine dependence[4], and three of the five LOF carriers were self-reported to be former smokers.

Considering that LOF alleles may primarily influence phenotype in the homozygous state[13], we separately analyzed 1,156 homozygous LOF sites representing 921 genes. Similar gene-based T5 tests were performed to compare the phenotype levels in LOF homozygotes to other individuals within the sample. We observed one homozygous association which replicated (**Table 2.4**). Four individuals were homozygous for LOF mutations in C1q and Tumor Necrosis Factor Related Protein 8 (*C1QTNF8*) and these individuals had elevated serum magnesium levels (**Figure 2.4b**). This diverse family of genes, including adiponectin, is linked to both metabolism and inflammatory processes[14], although this particular member is not well-characterized.

Discussion

We identified 10 LOF mutation-phenotype relationships that were both significant and replicated, but it is important to more broadly consider the concept of replication in the context of rare variant studies. In this study, 101 genotype-phenotype relationships with compelling statistical evidence ($p < 4.4 \times 10^{-6}$) were exclusive to either the discovery or replication group; in these cases, LOF mutations were only present in one or the other sample, but not both. These "absent" replications are not directly supported nor discredited, as they represent chance absence of the appropriate rare event (**Figure 2.5**).

Figure 2.5: LOF variants and genes carrying LOF variants with increasing sample size. For each sample size n, we randomly chose n ARIC individuals and observed the number of LOF variants and genes harboring them. This process was repeated 1,000 times to calculate the average numbers of LOF variants and genes carrying LOF variants for sample size n.



Identification of LOF variation influencing chronic disease risk factors represents a new and diverse paradigm in genomic medicine. LOF variation in certain genes, such as *TXNDC5*, may predispose individuals to develop disease. Further characterization of these risk loci will allow researchers and clinicians to better understand the pathways and mechanisms underlying disease risk, and to develop prevention strategies for at-risk patients as DNA sequencing moves inevitably towards common clinical practice. LOF variation can also have a protective, risk-lowering effect on their carriers. When coupled with knowledge about lack of other adverse effects, such LOF mutations may translate into novel drug targets. For example, LOF variants in *PCSK9* are associated with reduced LDL-cholesterol levels and incident coronary heart disease, fueling a burgeoning and successful effort to identify *PCSK9* inhibitors[15].

Discovery of novel gene associations via exome sequencing has many challenges, and represents a classic problem related to the signal-to-noise ratio. This study employed three ways to increase signal in whole exome sequence analyses. First, by including biochemical measures of risk factor levels, we optimize the size of a gene's effect relative to the corresponding disease endpoint. Second, the data presented here reinforces the need to have ethnic diversity in sequence-based gene discovery studies because the sentinel signals may be race-specific. And third, by careful annotation of the sequence motifs and variation, in this case by focusing on LOF variation, we increase the likelihood of detecting a functional effect. As we make the transition from whole exome sequencing to whole genome sequencing[16] careful annotation of variants with functional effect will become even more important and challenging.

Methods

Sample Ascertainment

Whole exome sequence data was derived from 8,554 individuals (5,718 EA; 2,836 AA) sampled from the Atherosclerosis Risk in Communities (ARIC) study cohort. Each ancestry group was then randomly divided in half to create two non-overlapping and identically-sized groups of 1,418 AA and 2,859 EA individuals for discovery and replication. EA individuals were selected as part of a large cohort random sample or had extreme values for at least one of the following phenotypes: age at menopause, electrocardiogram QT interval, fasting blood glucose, fibrinogen level, renal function, Stamler-Kannel-like extremes of risk factors selected by principal components, and waist-to-hip ratio. ARIC AA samples were randomly selected within the ARIC cohort for whole exome sequencing. A detailed description of the ARIC study is provided elsewhere[24].

Phenotying

For these analyses, we selected heart/lung/blood phenotypes related to cardiovascular outcomes that were (1) specifically not included in the sampling design to reduce potential bias and (2) measured across the entire cohort to maximize sample size. The full set of phenotypes included in these analyses is listed in **Table 2.3**. Serum magnesium (Mg) was measured using the metallochromic dye, Calmagite. Phosphorus (P), calcium (Ca), and creatinine (CRE) levels were measured using methods based on ammonium molybdate, *o*-cresolphthalein complexone, and modified kinetic Jaffe-picric acid, respectively. Serum potassium (K) and sodium (Na) levels were measured with a direct electrochemical technique. The liver enzymes, aspartate transaminase (AST), alanine aminotransferase (ALT) and gamma-glutamyl transpeptidase (GGT) were measured using standard methods. Blood pressure was measured using a standardized Hawskley random-zero mercury column sphygmomanometer with participants in a

sitting position after a resting period of 5 minutes. The size of the cuff was chosen according to the arm circumference. Three sequential recordings for systolic (SBP) and diastolic blood pressure (DBP) were obtained; the mean of the last two measurements was used in this analysis, discarding the first reading. Forced vital capacity (FVC) and the ratio of forced expiratory volume in one second (FEV1) to FVC were measured using a spirometer and the Pulmo-Screen II software. Triglycerides (TRG) and total cholesterol (TCH) were measured after an overnight fast using enzymatic methods. Fasting insulin (FI) was measured via radioimmunoassay. Glucose (FG) was measured with the hexokinase method on individuals having fasted > 8 hours prior to obtain fasting glucose. Uric acid (UA) was measured by the Uricase method. White blood cell (WBC) count was determined by an automated particle counter. Lactate (LAC) was measured using an enzymatic reaction that converts lactate to pyruvate.

Whole exome sequencing

DNA sequencing was performed on Illumina HiSeq instruments (San Diego, CA) after exome capture with VChrome2.1 (NimbleGen, Inc., Madison, WI) using chemistry recommended by the manufacturer. Sequence alignment and variant calling were performed using the Mercury pipeline in the DNA Nexus[41].

Variant calling and quality control

Mapping against Genome Reference Consortium Human Build 37 (GRCh37) was done using Burrows-Wheeler Alignment (BWA)[42] and allele calling and variant call file construction was performed with the Atlas2 suite (Atlas-SNP and Atlas-Indel) to produce a variant call file (VCF)[43]. The VCF includes filters for low-quality sites which were omitted from analyses,

including low-quality single nucleotide variants with a SNP posterior probability less than 0.95, total depth of coverage less than 10x, an allelic fraction < 0.1, 99% reads in a single direction and homozygous reference alleles with < 6x coverage. Similar, but stricter filters were applied to identify low-quality indels with the following differences: (1) minimum total depth < 60, (2) allelic fraction < 0.2 for heterozygous variants (< 0.8 for homozygous variants) and (3) variant reads < 30.

Validation

We validated a subset of LOF candidate genotypes using independent platforms with an emphasis on indels**.** We used targeted sequencing methods (Sequenom and Sanger) and observed a validation rate of 97.4% for SNV and 92.5% for LOF indel sites (**Table 2.5**).

This study took advantage of two opportunities to validate LOF variants detected by the Illumina HiSeq instrument and the Mercury data processing pipeline. First, 2,649 SNV LOF sites observed within our sample were also targeted on the Illumina exome chip[44]. Within this overlap, 98% of genotypes were identical between these two platforms. Second, we selected 263 LOF genotypes (176 indel, 87 snv) to validate on independent platforms (**Table 2.5**). These variants were a mixture of our top phenotype association results presented in **Table 2.4** and convenience sample of other sites, with oversampling of indels because of previous experience with their validation rates. These genotypes represent 147 unique LOF sites (126 indel, 21 SNV). Validation genotypes were re-genotyped via both Sanger sequencing and a targeted Sequenom panel. Twenty-four genotypes failed both assays. Concordant genotypes were observed for 225 LOF genotypes (148 indel, 77 snv), and at least one platform was discordant for 14 genotypes. Of note, none of the 14 discordant genotypes failed to validate on both

Table 2.5: Summary of Sanger/Sequenom validation rate by LOF class. Validation procedures are described in detail in the Methods Summary and Supplement IIb. "Selected genotypes" = number of LOF HiSeq genotypes submitted for validation via Sanger/Sequenom; "Failed genotypes" = no results from validation assay; "Remaining genotypes" = Submitted - Failed; "Conflicting genotypes" = Sanger/Sequenom genotypes do not match HiSeq; "Validated genotypes" = Sanger/Sequenom genotypes match HiSeq; "Validation Rate" = "Validated genotypes" / "Remaining genotypes".

Validation Status	SNP	INDEL	Total
Selected genotypes	87	176	263
Failed genotypes	8	16	24
Remaining genotypes	79	160	239
Conflicting genotypes	2	12	14
Validated genotypes	77	148	225
Validation Rate	0.974	0.925	0.941

platforms and represent inconsistencies between the validation platforms. Thus, using definitions that are common in the field, the observed validation rate for sites was 100%, and the observed validation rate for genotypes was 94.1% (225/(263-24)). More specifically the observed rate for genotypes was 97.4% for SNV and 92.5% for indel sites, and this may be a conservative underestimate of the true validation rate of our Illumina HiSeq data.

Annotation

We defined loss-of-function variation as sequence changes predicted to trigger nonsense-mediated decay of mRNA transcripts derived from all isoforms of a given gene. Thus, the basic annotation[29] categories of variation analyzed were premature stop codons, essential splice site disrupting, and indels predicted to disrupting the downstream reading frame. We further enriched for variants likely to abolish protein formation by identifying and excluding (1) stop-gain mutations occurring in the terminal gene exon, and (2) LOF candidates which did not map to chromosomal coordinates used by all gene isoforms for a given gene ("low-confidence-partial" (LCP)). Finally, we excluded candidate LOF sites with a MAF > 0.5 and genes lacking introns or designated non-protein-coding by RefSeq.

We used resampling methods to determine the relationship between sample size and LOF variant ascertainment. For each sample size *n*, we randomly chose *n* samples from the total samples and counted both the number of LOF variants observed and the number of genes carrying LOF variants. We repeated the process 1,000 times and calculated the average numbers of LOF variants and genes carrying LOF variants for sample size *n*. **Table 2.5** shows the average numbers of LOF variants and genes carrying LOF variants with increasing sample size.

Genotype-phenotype association

A gene-based burden test (T5)[32] was used to evaluate the association between aggregated rare LOF variants and phenotypes. We chose to employ this test due its interpretable detection of unidirectional phenotype mean shifts between LOF carriers and non-carriers. The following phenotype transformations were performed for T5 analyses: ALT, AST, CRE, FI, and LAC underwent natural log transformation; FEV1FVC, GGT and MG underwent power transformation; CA was corrected using the following formula: total calcium (mmol/l) + ([40 - serum albumin (g/dl)] * 0.025). TCH was adjusted (TCH/0.8) only among statin users; measured SBP and DBP were respectively adjusted by +15 mmHg and +10 mmHg for individuals taking anti-hypertensive medication; all other traits did not require transformation. T5 tests were implemented using the SeqMeta package available in Cran R (http://cran.r-project.org/web/packages/seqMeta/), and only associations that were 1) independently detected in both sample strata, 2) persisted with the inclusion of all samples, and 3) driven by ≥ 3 individuals are presented in **Table 2.4**. Allele frequencies were calculated separately for each ancestry group and only variants with an observed MAF < 5% were included in ancestry-specific analyses. Based on a Bonferroni correction procedure for the number of genes in our sample presenting LOF variation (n = 11,260), a p-value of 4.4×10^{-6} was considered statistically significant. Similarly, a p-value of 5.42×10^{-5} was considered significant for associations driven by homozygous individuals, adjusting for the number of genes presenting homozygous LOF genotypes (n=921). Quantile-quantile plots of all T5 p-values are provided in **Figure 2.6** and **Figure 2.7.**

OP Ratio

We developed the OP ratio ("Observed" /"Potential") as a gene-based metric to quantify LOF variation while accounting for transcript size, and as a useful tool to compare the rate of LOF variation in different gene groups. This metric is the ratio of the number of observed LOF sites in a gene to the number of possible LOF sites that could arise due to single-nucleotide substitutions. We compared the OP ratio to other measures of gene variability. We used the egenetics database[30] to rank all genes by the number of tissues where they are stably expressed, calling the top 5% of this list "universally expressed". On average, we observe a smaller OP ratio within stably expressed genes compared to all others (**Figure 2.3a**). Similarly, we sorted the genes according to the number of known protein interactions according to ConsensusPathDB[31], and categorized the top 5% of these genes as "highly-interacting genes". This gene group also has a smaller OP ratio on average compared to other genes (**Figure 2.3b**).

We compared our OP ratio with the Residual Variation Intolerance Score (RVIS)[32] for 15,053 genes having both OP ratio and RVIS available. RVIS is based on the ratio of common non-synonymous and splicing-site SNPs to the total numbers of coding SNPs using the ESP6500 dataset. Both the OP ratio and RVIS are designed to measure the gene tolerance of damaging amino acid changes, but are different as to the way of measurement and the databases they are based on. Both the Pearson's correlation coefficient (0.204) and the Spearman's rank correlation coefficient (0.229) between the two scores are highly statistically significant ($p \sim= 0$), although we do not see clear linear relationship between them (**Figure 2.8**).

Figure 2.6: Quantile-quantile plots of p-values from T5 associations with 20 phenotypes. The 95% confidence intervals are depicted and each dot represents one gene. Phenotype symbols are defined in Supplementary table 2.

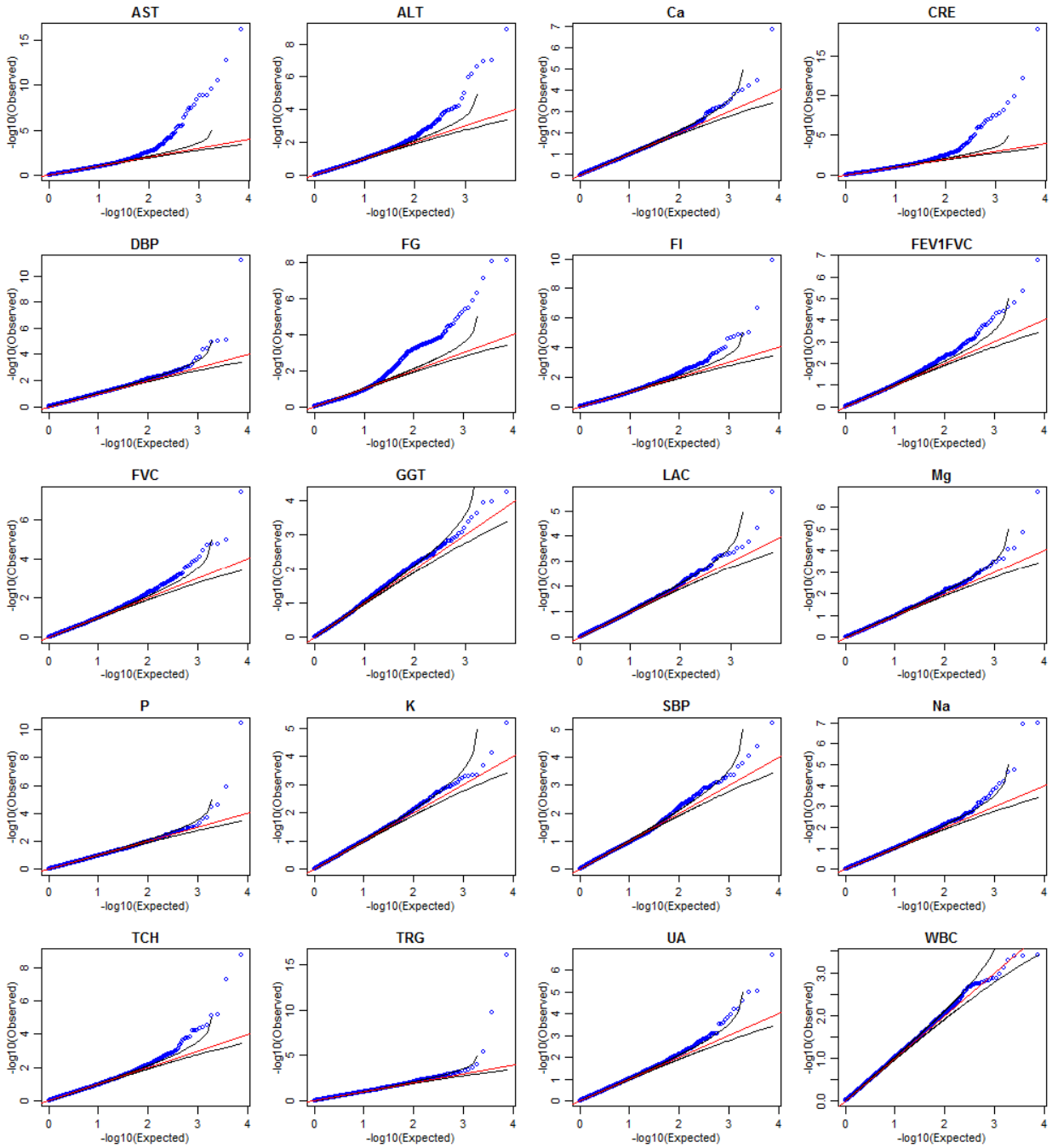


Figure 2.7: Quantile-quantile plots of p-values from T5 homozygous associations with 20 phenotypes. The 95% confidence intervals are depicted and each dot represents one gene. Phenotype symbols are defined in Supplementary table 2.

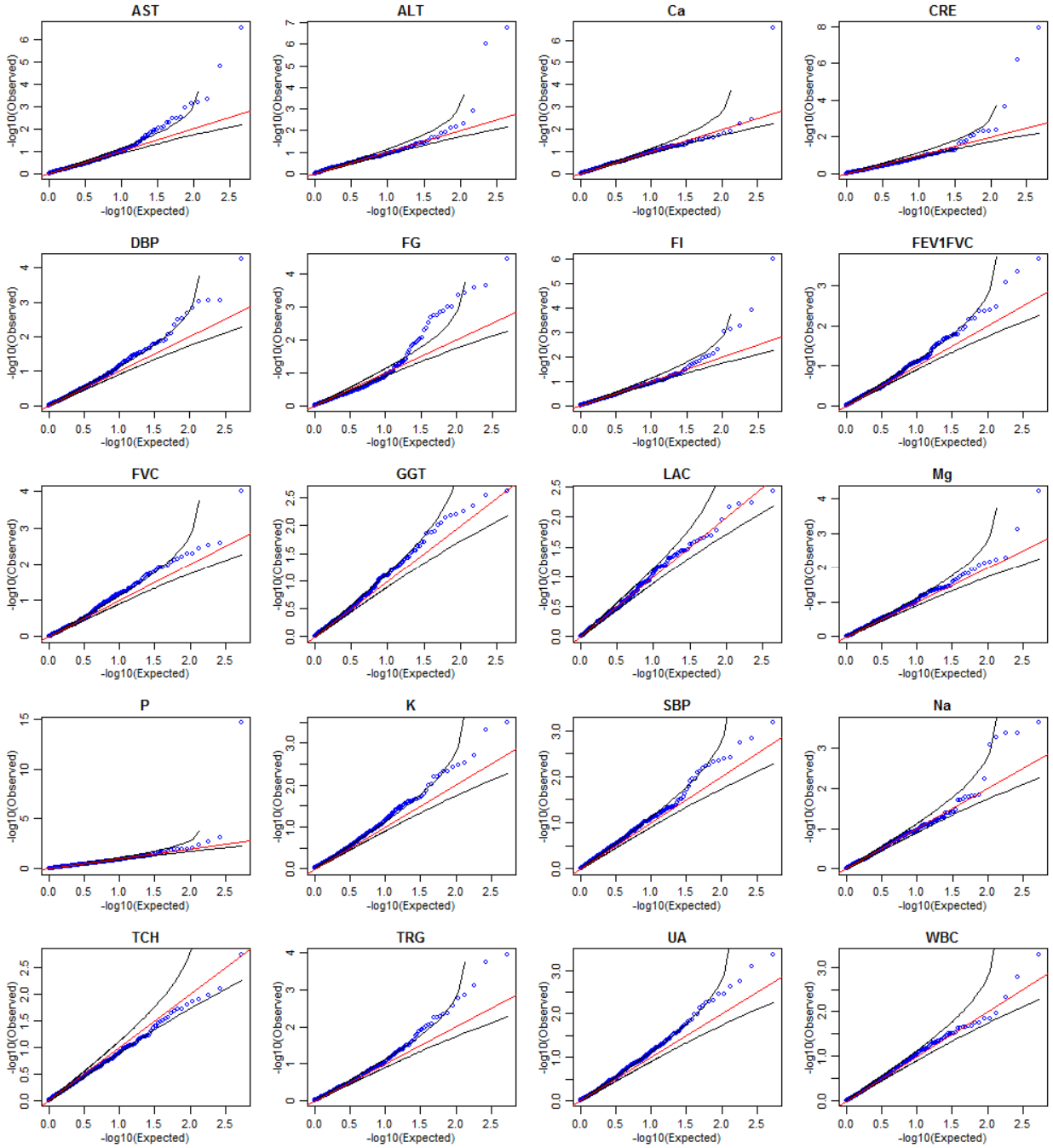
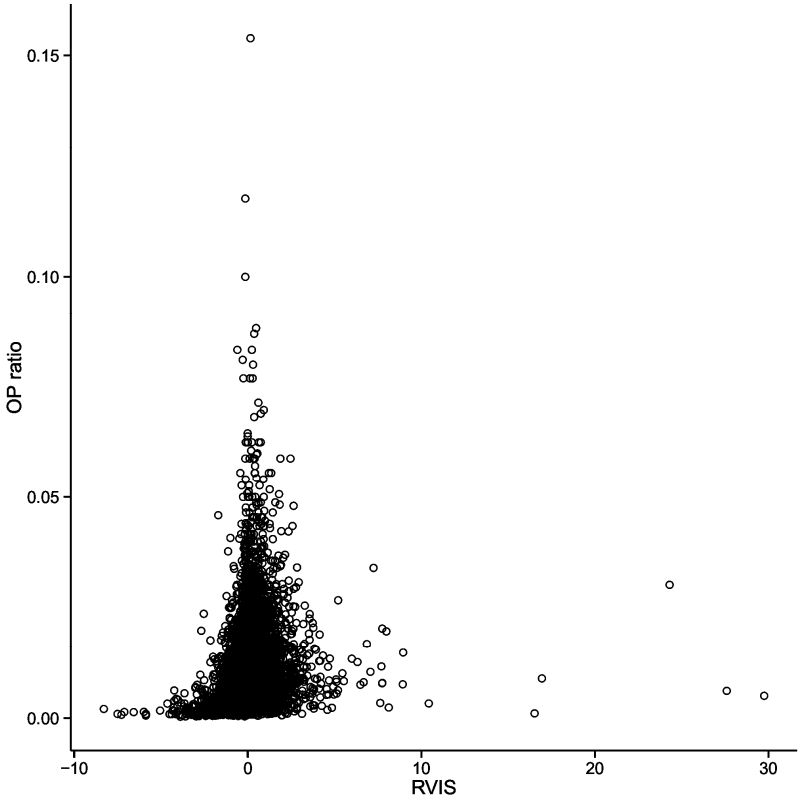


Figure 2.8: Relationship between OP ratio and RVIS for 15,053 genes. The y-axis depicts the OP ratio and the x-axis shows RVIS scores as described by Petrovski et al[48]. Each circle represents one gene.



Chapter 3: Metabolite intermediate phenotypes

Some of the figures and tables in this chapter are based on: Yu, B., Li, A. H., Muzny, D., Veeraraghavan, N., de Vries, P. S., Bis, J. C., Musani, S. K., Alexander, D., Morrison, A. C., Franco, O. H., Uitterlinden, A., Hofman, A., Dehghan, A., Wilson, J. G., Psaty, B. M., , Gibbs, R., Peng Wei, P., Boerwinkle, E. (2015). Association of Rare Loss-Of-Function Alleles in HAL, Serum Histidine: Levels and Incident Coronary Heart Disease. Circulation. Cardiovascular genetics, 8(2), 351–5. doi:10.1161/CIRCGENETICS.114.000697

Introduction

Complex molecular interactions between the products of gene function and external environmental influences underlie and culminate in the gross anatomic, metabolic, and physiologic traits analyzed in the other chapters of this thesis (*e.g.*, lung function, blood pressure, cardiac birth defects, etc). Many of these molecular interactors are mediated by small molecule metabolites, and the totality of these small molecules in a system is referred to as the human metabolome.

Many features of the metabolome are appealing for genetic association studies. First, metabolites may be directly encoded by genes or may function in close biological proximity to gene products. For example, they can serve as substrates for enzymes that are directly encoded by genes[49]. This proximity increases the interpretability of associations, and may also reduce the possibility for confounding by interaction with networks of other small molecules. Next, the metabolites may serve as biomarkers for disease prediction before the onset of clinically recognized symptoms. For example, amino acid profiles within individuals have recently been used to predict coronary artery disease[50], myocardial infarction[51], and other cardiovascular outcomes[52]. Finally, metabolites are readily quantifiable. Recent technological advances allow high-throughput quantification of these small molecules comprising the human metabolome. Bioanalytical techniques emerging from advances in high-performance liquid-phase chromatography (HP-LPC) allow for simultaneous measurement of hundreds of small molecules within a single blood sample.

Metabolites have been studied at the population level through genome wide association studies (GWAS). The loci identified in GWAS of the metabolome have presented with large effect sizes and provide insight into the biological actions of the associated regions leading to an

effect on the trait of interest[53,54]. For example, genetic markers may map near genes that encode enzymes or transporters with a biological function relevant to the associated small molecule. However, the majority of the genetic markers driving these GWAS signals do not directly map to protein-encoding genes. One explanation is that the polymorphisms genotyped in GWAS are not causal; rather they may be linked to and in linkage disequilibrium with a deleterious variant that is directly driving the observed associations.

LOF allelic variation has great potential to influence the human metabolome but has not yet been studied. First, this type of variation is predicted to have direct effects on gene action, including metabolite levels and the proteins that regulate those levels. Thus, associations driven by LOF variation may be very interpretable, especially if this variation disrupts the function of a gene whose action may influence the metabolite of interest. In addition, this functional class of variation is enriched at the rare end of the allele frequency spectrum compared to protein-altering sequences. Thus, it seems likely that a significant portion of loci associated with metabolome phenotypes may harbor rare LOF variation contributing to these associations.

Methods

Study Population

The Atherosclerosis Risk in Communities (ARIC) study is a prospective epidemiological study designed to investigate the etiology and predictors of cardiovascular disease (CVD). A detailed description of the ARIC study design and methods is published elsewhere[40]. Basic cardiovascular risk factors were measured at each visit, and cardiovascular endpoints, such as heart failure were ascertained annually using telephone interviews and hospital medical record review. Detailed demographics are provided in **Table 3.1**.

Table 3.1: Baseline Characteristics of African Americans in ARIC for whole exome sequencing analyses (N=1,361). For continuous variables, mean values ± standard deviations are shown. Circulation. Cardiovascular genetics: 8 January 2015 - Volume 8 - Issue 17 - p 351-355. American Heart Association, Inc ©.

Characteristic	n (%) or Mean (SD)
Age, years	52.5 ± 5.6
Body mass index, kg/m^2	29.8 ± 6.2
Male, n (%)	464 (34.1)
Hypertension, n (%)	705 (51.8)
Diabetes, n (%)	190 (14.0)
Current smoker, n (%)	377 (27.7)
Prevalent coronary heart disease, n	49 (3.6)
Systolic blood pressure, mm Hg	126.9 ± 19.0
Diastolic blood pressure, mm Hg	80.2 ± 11.4
HDL cholesterol, mg/dL	56.1 ± 17.4
LDL cholesterol, mg/dL	135.2 ± 38.0
Triglycerides, mg/dL	106.2 ± 55.9
Total cholesterol, mg/dL	212.5 ± 40.0

.

Metabolome Measurements

Metabolite profiling was measured using fasting serum samples collected from the baseline visit (1987-1989) of ARIC African-Americans study participants. A total of 602 metabolites were detected and quantified by Metabolon Inc. (Durham, USA) using an untargeted, gas chromatography-mass spectrometry (GC-MS) and liquid chromatography-mass spectrometry (LC-MS)-based metabolomic quantification protocols[55,56]. Metabolites were excluded if: 1) more than 50% of the samples presented values below the platform detection limit; or 2) the chemical structure was unknown. After filtering based on these criteria, a total of 308 named metabolites were included in the present study; 83 amino acids, 16 carbohydrates, 9 cofactors and vitamins, 7 energies, 136 lipids, 12 nucleotides, 25 peptides and 20 xenobiotics (**Table 3.2**).

Whole Exome Sequencing and Variant Validation

Whole exome sequence (WES) was performed using Illumina HiSeq 2000 (Illumina, San Diego, CA, USA) and VCRome2.1 capture reagents (Roche NimbleGen, Madison, WI, USA). Sequences were aligned to the hg19 reference genome using Burrows–Wheeler Aligner[42]. Allele calling and variant call file (VCF) construction was performed using the Atlas2 suite[43] (Atlas-SNP and Atlas-Indel) to produce a VCF file. Single nucleotide variants were excluded if they had a posterior probability less than 0.95, total depth of coverage less than 6x, allelic fraction < 0.1, 99% of reads in a single direction and homozygous reference alleles with < 6x coverage. Low-quality indels were excluded if they had minimum total depth < 30, allelic fraction < 0.2 for heterozygous variants and < 0.8 for homozygous variants and variant reads < 10.

Table 3.2: List of 308 metabolites included in this study. This table includes the category, number of metabolites (n), and the names of specific metabolites included in these analyses.

Category	n	Metabolites
Amino acid	83	2-aminobutyrate; 2-hydroxybutyrate (AHB); 2-hydroxyisobutyrate; 2-methylbutyroylcarnitine; 3-(4-hydroxyphenyl)lactate; 3-hydroxy-2-ethylpropionate; 3-hydroxyisobutyrate; 3-indoxyl sulfate; 3-methoxytyrosine; 3-methyl-2-oxovalerate; 3-methylhistidine; 3-phenylpropionate (hydrocinnamate); 4-acetamidobutanoate; 4-guanidinobutanoate; 4-methyl-2-oxopentanoate; 5-oxoproline; alanine; alpha-hydroxyisocaproate; alpha-hydroxyisovalerate; anthranilate; arginine; asparagine; aspartate; beta-hydroxyisovalerate; betaine; C-glycosyltryptophan; citrulline; creatine; creatinine; cysteine; dimethylarginine (SDMA + ADMA); dimethylglycine; glutamate; glutarate (pentanedioate); glutaroyl carnitine; glycine; histidine; homocitrulline; homostachydrine; hydroxyisovaleroyl carnitine; indoleacetate; indolelactate; indolepropionate; isobutyrylcarnitine; isoleucine; isovalerylcarnitine; kynurenine; leucine; lysine; methionine; methionine sulfoxide; N6-acetyllysine; N-acetylalanine; N-acetyl-beta-alanine; N-acetylglycine; N-acetylornithine; N-acetylphenylalanine; N-acetylserine; N-acetylthreonine; N-methyl proline; o-cresol sulfate; ornithine; p-cresol sulfate; phenol sulfate; phenylacetate; phenylacetylglutamine; phenylalanine; phenyllactate (PLA); pipecolate; proline; pyroglutamine; serine; serotonin (5HT); stachydrine; threonine; tiglyl carnitine; trans-4-hydroxyproline; tryptophan; tryptophan betaine; tyrosine; urea; urocanate; valine
Carbohydrate	16	1,5-anhydroglucitol (1,5-AG); 1,6-anhydroglucose; arabinose; erythronate; erythrose; fructose; gluconate; glucose; glucuronate; glycerate; lactate; mannitol; mannose; pyruvate; threitol; trehalose
Cofactors and vitamins	9	alpha-tocopherol; arabonate; bilirubin (E,E); bilirubin (Z,Z); biliverdin; gamma-tocopherol; pantothenate; pyridoxate; threonate
Energy	7	acetylphosphate; cis-aconitate; citrate; malate; phosphate; succinate; succinylcarnitine
Lipid	136	1,2 propanediol; 10-heptadecenoate (17:1n7); 10-nonadecenoate (19:1n9); 13-HODE + 9-HODE; 1-arachidonoylglycerophosphocholine; 1-arachidonoylglycerophosphoethanolamine; 1-arachidonoylglycerophosphoinositol; 1-docosahexaenoylglycerophosphocholine; 1-docosapentaenoylglycerophosphocholine; 1-eicosadienoylglycerophosphocholine; 1-eicosatrienoylglycerophosphocholine; 1-heptadecanoylglycerophosphocholine; 1-linoleoylglycerophosphocholine; 1-linoleoylglycerophosphoethanolamine; 1-myristoylglycerophosphocholine; 1-O-hexadecylglycerophosphocholine; 1-oleoylglycerol (1-monoolein); 1-oleoylglycerophosphocholine; 1-oleoylglycerophosphoethanolamine; 1-palmitoleoylglycerophosphocholine; 1-palmitoylglycerol (1-monopalmitin); 1-palmitoylglycerophosphocholine; 1-palmitoylglycerophosphoethanolamine; 1-palmitoylglycerophosphoinositol; 1-pentadecanoylglycerophosphocholine; 1-stearoylglycerol (1-monostearin); 1-stearoylglycerophosphocholine; 1-stearoylglycerophosphoethanolamine; 1-stearoylglycerophosphoinositol; 21-hydroxypregnenolone disulfate; 2-arachidonoylglycerophosphocholine; 2-arachidonoylglycerophosphoethanolamine; 2-hydroxyglutarate; 2-hydroxyoctanoate; 2-hydroxypalmitate; 2-hydroxystearate; 2-linoleoylglycerophosphocholine; 2-linoleoylglycerophosphoethanolamine; 2-oleoylglycerophosphocholine; 2-oleoylglycerophosphoethanolamine; 2-palmitoylglycerophosphocholine; 2-palmitoylglycerophosphoethanolamine; 2-

		stearoylglycerophosphocholine; 3-carboxy-4-methyl-5-propyl-2-furanpropanoate (CMPF); 3-dehydrocarnitine; 3-hydroxybutyrate (BHBA); 3-hydroxydecanoate; 4-androsten-3beta,17beta-diol disulfate 1; 4-androsten-3beta,17beta-diol disulfate 2; 5alpha-androstan-3beta,17beta-diol disulfate; 5alpha-pregnan-3beta,20alpha-diol disulfate; 5-dodecenoate (12:1n7); 5-HETE; 7-alpha-hydroxy-3-oxo-4-cholestenoate (7-Hoca); 7-beta-hydroxycholesterol; acetylcarnitine; adipate; adrenate (22:4n6); andro steroid monosulfate 2; androsterone sulfate; arachidonate (20:4n6); azelate (nonanedioate); caprate (10:0); caproate (6:0); caprylate (8:0); carnitine; cholate; cholesterol; choline; cis-vaccenate (18:1n7); cortisol; cortisone; decanoylcarnitine; dehydroisoandrosterone sulfate (DHEA-S); deoxycarnitine; deoxycholate; dihomo-linoleate (20:2n6); dihomo-linolenate (20:3n3 or n6); docosahexaenoate (DHA; 22:6n3); docosapentaenoate (n3 DPA; 22:5n3); docosapentaenoate (n6 DPA; 22:5n6); dodecanedioate; eicosapentaenoate (EPA; 20:5n3); eicosenoate (20:1n9 or 11); epiandrosterone sulfate; glycerol; glycerol 3-phosphate (G3P); glycerophosphorylcholine (GPC); glycochenodeoxycholate; glycocholate; glycocholenate sulfate; glycodeoxycholate; glycolithocholate sulfate; heptanoate (7:0); hexadecanedioate; hexanoylcarnitine; hyodeoxycholate; inositol 1-phosphate (I1P); isovalerate; laurate (12:0); laurylcarnitine; linoleate (18:2n6); linolenate[alpha or gamma; (18:3n3 or 6)]; margarate (17:0); methyl palmitate; myo-inositol; myristate (14:0); myristoleate (14:1n5); nonadecanoate (19:0); octadecanedioate; octanoylcarnitine; oleate (18:1n9); oleoylcarnitine; palmitate (16:0); palmitoleate (16:1n7); palmitoyl sphingomyelin; palmitoylcarnitine; pelargonate (9:0); pregn steroid monosulfate; pregnen-diol disulfate; propionylcarnitine; scyllo-inositol; sebacate (decanedioate); stearate (18:0); stearidonate (18:4n3); stearoyl sphingomyelin; stearoylcarnitine; suberate (octanedioate); taurochenodeoxycholate; taurocholate; taurocholenate sulfate; taurolithocholate 3-sulfate; tetradecanedioate; undecanedioate; undecanoate (11:0); valerate
Nucleotide	12	5-methyluridine (ribothymidine); 7-methylguanine; adenosine; allantoin; guanosine; hypoxanthine; inosine; N1-methyladenosine; pseudouridine; urate; uridine; xanthine
Peptide	25	[H]HWESASLLR[OH]; alanylleucine; alpha-glutamylglutamate; aspartylphenylalanine; bradykinin, des-arg(9); DSGEGDFXAEGGGVR; gamma-glutamylalanine; gamma-glutamylglutamate; gamma-glutamylisoleucine; gamma-glutamylleucine; gamma-glutamylphenylalanine; gamma-glutamylthreonine; gamma-glutamyltyrosine; gamma-glutamylvaline; glycylleucine; glycylphenylalanine; glycyltyrosine; glycylvaline; HWESASXX; HXGXA; leucylleucine; leucylphenylalanine; pro-hydroxy-pro; pyroglutamylglycine; threonylphenylalanine
Xenobiotics	20	1,7-dimethylurate; 1-methylurate; 2-hydroxyhippurate (salicylurate); 3-ethylphenylsulfate; 4-ethylphenylsulfate; 4-hydroxyhippurate; 4-vinylphenol sulfate; 5-acetylamino-6-amino-3-methyluracil; benzoate; caffeine; catechol sulfate; erythritol; glycerol 2-phosphate; hippurate; paraxanthine; piperine; salicylate; theobromine; theophylline; thymol sulfate

Following statistical analysis, all significantly associated variants were validated using an orthogonal laboratory technology (i.e. Array-based genotyping[44], Sequenom genotyping, or Sanger sequencing). All reported genotypes driving associations were successfully validated.

Statistical Analyses

LOF variants included in this study were defined as premature stop codons occurring in the non-terminal gene exon, essential splice site disrupting ($\pm$ 2bp), and indels predicted to disrupt downstream reading frame. T5 tests[32] were performed to evaluate the joint effects of rare alleles (MAF<5%) in a gene, and were conducted on each metabolite after adjusting for age, gender, estimate glomerular filtration rate calibrated (eGFR)[57] and population structure. Statistical significance was defined as a p-value $< 1.3 \times 10^{-7}$ for T5 tests (Bonferroni correction of 395,780 tests: 1,285 genes × 308 metabolites). Metabolite levels were natural log-transformed prior to the analysis. All the analyses were performed using R (www.r-project.org).

Results

Metabolite associations

WES of the 1,361 African-American ARIC participants also measured for these metabolites revealed 12,522 polymorphic LOF variants (5,060 stopgains, 2,599 splice and 4,863 frameshift indels) representing 7,038 genes. Each sample contained an average of 111.7 heterozygous and 14.5 homozygous LOF variants.

The effect of LOF variation on 308 metabolites was analyzed by gene-based aggregation of these variant sites. Eight genes harboring 17 LoF variants (7 stopgains, 3 splice and 7 frameshift indels) were identified to be significantly associated with eight metabolites levels (p

$< 1.3 \times 10^{-7}$, **Table 3.3**), and these variants were related to 19-50% change of the geometric mean for metabolite levels depending on the particular metabolite. As expected[26], we observed that LOF variants in PCSK9 had lower cholesterol levels compared to the non-carriers ($p = 5.4\times10^{-9}$). However, we also highlight several novel associations with compelling underlying biology.

We observed 3 LOF variants in the gene Histidine Ammonia Lyase (*HAL*) that were strongly associated with decreased levels of Histidine (published, 2015)[58]. Histidine is an antioxidant and anti-inflammatory factor (**Figure 3.1a**). In addition to these features, HAL plays an essential role in the catabolism of Histidine (**Figure 3.2b**). The 24 carriers of LOF variation in this gene collectively presented a 29.7% increase in histidine's geometric mean and explained 4.8% of its variance (**Figure 3.1c**), and the direction of this effect (increase) is consistent with the expectation given the role of HAL in Histidine biology. The association between R322X was replicated in an independent samples of 718 ARIC study participants with both exome chip data and serum histidine levels ($p=1.2x10^{-4}$)[58].

We also observed a LOF splice variant in *SLCO1B1* which was associated with high levels of hexadecanedioate ($p = 2.2\times10^{-9}$), a C16 dicarboxylic acid (manuscript submitted). *SLCO1B1* is an organic ion transporter expressed at high levels in the liver[59], and follow-up studies showed this variant has pleiotropic effects on tetradecanedioate, a C14 dicarboxylic acid ($p = 9.0\times10^{-5}$).

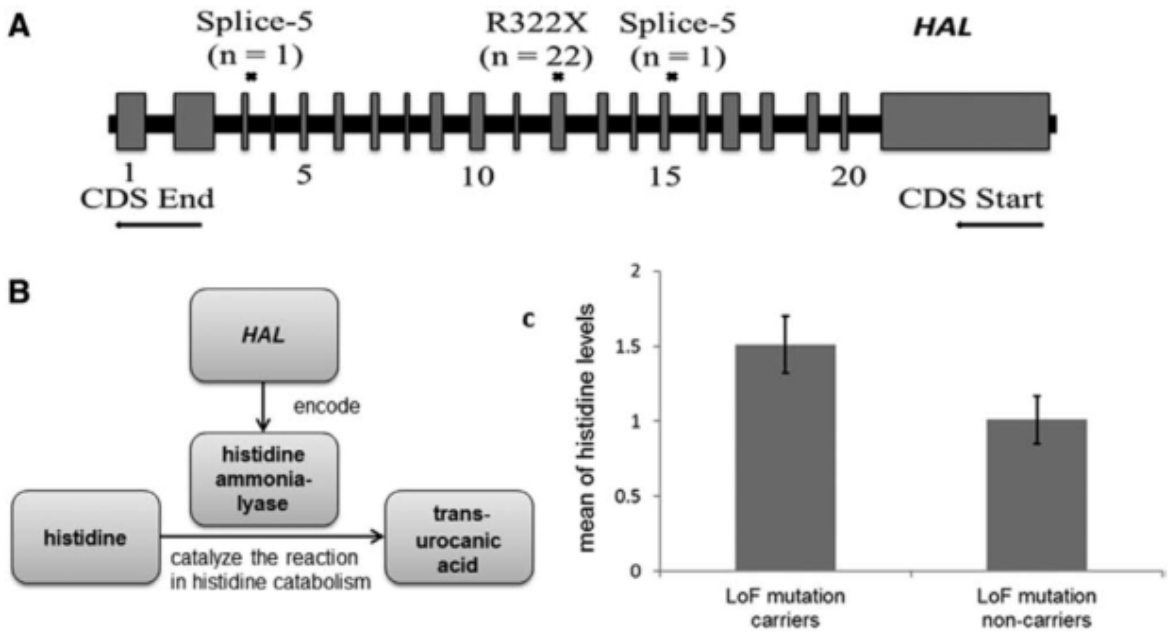
Considering that LOF may act recessively, we also observed samples with homozygous stop mutations in two genes (*LRRC69*, *SLCO1B1*) who also presented extreme values for metabolites. The presentation of hexadecanedioate within *SLCO1B1* heterozygotes is described above, but the single homozygous sample presented levels of this metabolite in the tail of this

Table 3.3. Eight significant gene-metabolite associations identified among African Americans in ARIC. SE = standard error; cMAC = cumulative minor allele count.

Metabolite	Pathway	Gene	P	Beta (SE)	cMAC
Histidine	Amino acid	*HAL*	2.3×10^{-13}	0.23 (0.03)	26
Methionine sulfoxide	Amino acid	*C6orf25*	1.3×10^{-8}	-0.45 (0.08)	9
Mannose	Carbohydrate	*TEX15*	7.9×10^{-9}	-0.70 (0.12)	10
Cholesterol	Lipid	*PCSK9*	5.4×10^{-9}	-0.21 (0.04)	30
Deoxycarnitine	Lipid	*LRRC69*	8.5×10^{-16}	-0.42 (0.05)	17
Hexadecanedioate	Lipid	*SLCO1B1*	2.2×10^{-9}	0.38 (0.06)	67
5-HETE	Lipid	*FAM198B*	4.5×10^{-9}	-0.38 (0.07)	14
Urate	Nucleotide	*LRRC46*	1.1×10^{-7}	-0.44 (0.08)	10

.

Figure 3.1: HAL LOF alleles and their association with histidine levels. (A) Three LOF variants in HAL among African Americans in ARIC; (B) Flow chart of HAL gene function; (C) Histidine levels in HAL LOF carriers and noncarriers, the error bard indicate standard deviation. Image via: Circulation. Cardiovascular genetics: 8 January 2015 - Volume 8 - Issue 17 - p 351-355. American Heart Association, Inc ©.



distribution (**Figure 3.2b**). Similarly, we observe extremely low levels of deoxycarnitine in a single sample with homozygous LOF variation in *LRRC69* (**Figure 3.2a**).

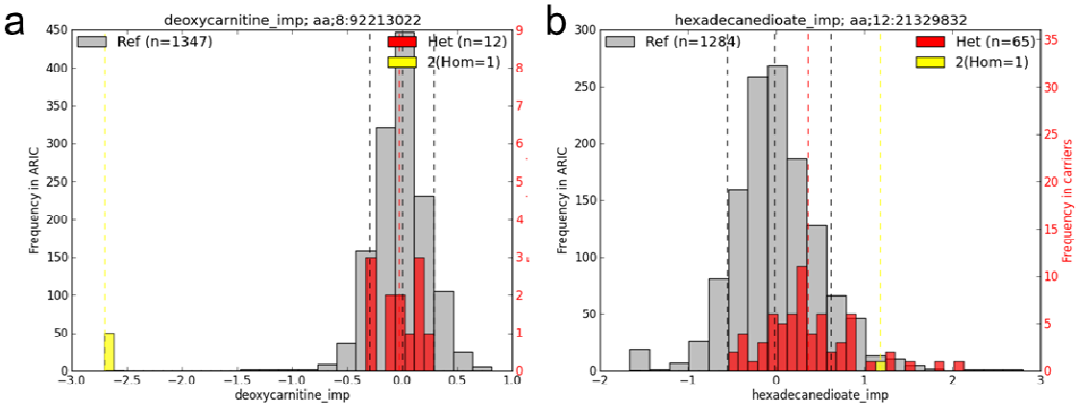
Discussion

Using high-throughput metabolomic and genomic sequence technologies, we carried out extensive analysis to evaluate the effect of predicted LOF mutations on the human serum metabolome. This study identified nine genes that were associated with ten metabolite levels among African-Americans, and seven out of the nine genes represent novel findings (**Table 3.3**). Our results provide new insights into the genetic architecture of the human metabolome among African-Americans, including the possible identification of LOF variation as a biomarker for a protective effect against coronary heart disease[58].

Inborn errors formed the earliest understanding of genetic architecture on human metabolism[60]. Over a hundred genes with multiple variants were reported to be associated with multiple metabolites by GWASs of the metabolome[53,61], but only in a few cases has the underlying functional variant(s) been identified that are responsible for the observed association. Thus far, hundreds of causal genes for severe inherited disorders of metabolism have been discovered, and the majority of the causal variants identified have been shown to be rare. Sequencing large numbers of samples and annotating clear functional categories of the detected variations provides one path leading toward identifying novel genes and variants contributing to complex phenotypes. Here, we report here that LOF mutations observed in the general population are related to human metabolome, and this approach provides a strategy for identifying novel disease genes.

Histidine is an essential amino acid in humans and other mammals, and a precursor for histamine and carnosine biosynthesis. The enzyme encoded by *HAL* catalyzes the first reaction

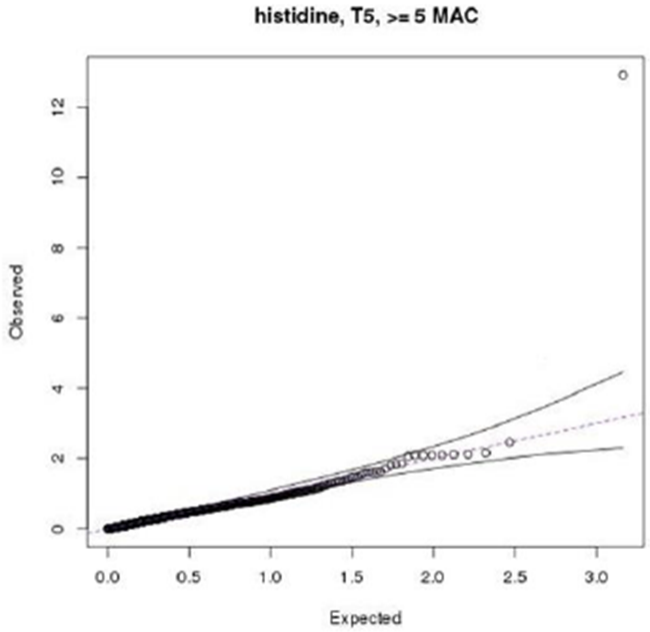
Figure 3.2. Distribution of metabolite levels among LoF mutation carriers in ARIC. (a) Low serum deoxycarnitine level in a single *LRRC69* LoF homozygote; (b) high serum hexadecanedioate levels in *SLCO1B1* LoF heterozygotes and a single homozygote.



in Histidine catabolism, and missense mutations in this gene cause autosomal recessive histidinemia (MIM:609457). Histidine levels in the blood exceeding 6 mg/dL are one marker for the diagnosis of this disorder. In this study, 24 study participants with LOF variation in *HAL* present elevated serum histidine (mean 1.5 mg/dL; **Figure 3.1**; **Figure 3.3**), suggesting a milder subclinical form may be present in these carriers. Histidine has anti-inflammatory properties[62] and low levels are associated with inflammatory diseases such as chronic kidney disease[63] and rheumatoid arthritis[64]. Inflammatory cytokines are also related to the early stages of atherosclerosis and coronary heart disease (CHD)[65]. Follow-up analyses within ARIC and other cohorts revealed an association between high levels of histidine and a reduced risk for CHD[58]. These results suggest that *HAL* and other members of the nuclear factor-κB pathway[66] may represent therapeutic targets for CHD and other inflammatory disorders.

Common non-functional variants tagging *SLCO1B1* have been reported in GWAS to be associated with fatty acids levels, including tetradecanedioate and hexadecanedioate[53]. *SLCO1B1* encodes a protein that mediates the cellular uptake of numerous endogenous compounds and is involved in clearing many drug compounds, including the statin drug class which lowers cholesterol[67]. In this study, we observed associations between LOF variation in *SLCO1B1* with tetradecanedioate and hexadecanedioate. Little is known about these two medium chain fatty acids, except a recent study showed that their levels are increased in the lung tissue of patients with pulmonary arterial hypertension (PAH)[68]. Follow up analyses reveal that LOF alleles in *SLCO1B1* are associated with high risk of heart failure in an extended sample of ARIC participants (data not shown). The findings indicate a possible causal effect of *SLCO1B1* on HF via altered fatty acids metabolism and uptake.

Figure 3.3: Quantile-quantile plots of T5 test on histidine levels in ARIC African Americans. Circulation. Image adapted from: Cardiovascular genetics: 8 January 2015 - Volume 8 - Issue 17 - p 351-355. American Heart Association, Inc ©.



The mechanisms potentially underlying other associations we detected are less clear. As the collection of data (genetic and phenotypic) becomes increasingly larger, granular, and higher quality, interpretation of findings rather than data collection become rate limiting. Cross-referencing analytical results with large public databases cataloguing metabolite function, such as the Human Metabolome Database (HMDB, http://www.hmdb.ca/) is a one approach to interpret associations with the greatest potential to impact healthcare. This data-driven approach does not preclude the collaboration between genetic analysts, clinicians, and phenotype specialists; rather it serves as additional tool for facilitating discovery. In addition, model organisms remain an essential tool for the study of specific mutational activity *in vivo*, although they are not high-throughput and rely on the *a priori* detection of strong analytical candidates.

In summary, we report here the first whole exome LOF study of the untargeted metabolome in African-Americans. Our findings illuminate the value of utilizing deep phenotype collection methods ("-omic") studies in cohort studies to provide new insights and generate new hypotheses into gene function and disease etiology. We identify LOF variation in eight genes which may make promising drug targets, especially *HAL* which regulates histidine and is linked to lowered risk of coronary heart diseae[58]. While the analysis of large "-omic" data sets may seems like an insurmountable challenge for researchers, these findings reveal that functional and biological paradigms (such as LOF variation) can be used to inform association studies and make significant and clinically interpretable findings.

Chapter 4: Rare Congenital Cardiovascular Malformation

Introduction

Severe congenital cardiovascular malformations (CVMs) occur in 5-8/1000 live births and have a high mortality rate compared to other birth defects[69,70]. Left Ventricular Outflow Tract Obstructions (LVOTO) comprise 15-20% of severe CVMs[71,72] , and include Hypoplastic Left Heart Syndrome (HLHS), Aortic Valve Stenosis (AS), Coarctation of the Aorta (CoA), Interrupted Aortic Arch Type A (IAAA), Mitral Stenosis and Shone Complex (**Figure 4.1**). This diverse family of cardiac conditions share underlying mechanisms driven by altered or obstructed blood through the left side of the heart during development[73].

Genetic contributions to the development of LVOTO are complex and include single nucleotide substitutions, chromosome abnormalities[74], genomic disorders[75] and oligogenic inheritance[76]. Over 30 genes have been previously implicated in human syndromes including LVOTO-type malformations. These loci include HLHS (*ZIC3*, *TBX5*, *CREBBP*, *ACVR2B*, *LEFTY2*, *DTNA*, *DHCR7*, *EVC1-2*, *FOXF1-FOXC2-FOXL1*, and *PEX* genes), AS (*NOTCH1, FOXC1*, *FGD1*), CoA (*JAG1*, *NOTCH2*, *NF1*, *PTPN11*, *KRAS*, *SOS1*, *RAF1*, *NRAS*, *BRAF*, *SHOC2*, *CBL*, *ZIC3*, *CREBBP*, *MLL2*, *FGD1*, *DHCR7*, *NSDHL*, *KCNJ2*, *MKS1*) (**Table 4.1**). Familial clustering of cases[77] and an increased risk of LVOTO in first-degree relatives[78] are consistent with single gene or oligogenic inheritance. However, the fact that many cases are sporadic also suggests a role for *de novo* mutations and other rare chance events. Zaidi et al[79] report the occurrence of *de novo* mutations in a cohort of congenital heart defect cases, albeit without respect to a specific CVM type.

To gain a deeper understanding of the spectrum of genetic variation associated with LVOTO, we performed whole-exome sequencing of a cohort of 342 LVOTO patients without extra-cardiac features. Variant frequencies were compared to multiple population data

Figure 4.1. Representation of Hypoplastic Left Heart Syndrome features. This figure depicts the spectrum of cardiovascular malformation associated with HLHS, a feature of LVOTO. Other LVOTO features include aortic valves stenosis (AS), coarctation of the aorta (CoA), interrupted aortic arch type A (IAAA), mitral stenosis and Shone Complex (Supravalvular mitral membrane, parachute mitral valve, subaortic stenosis, & CoA).

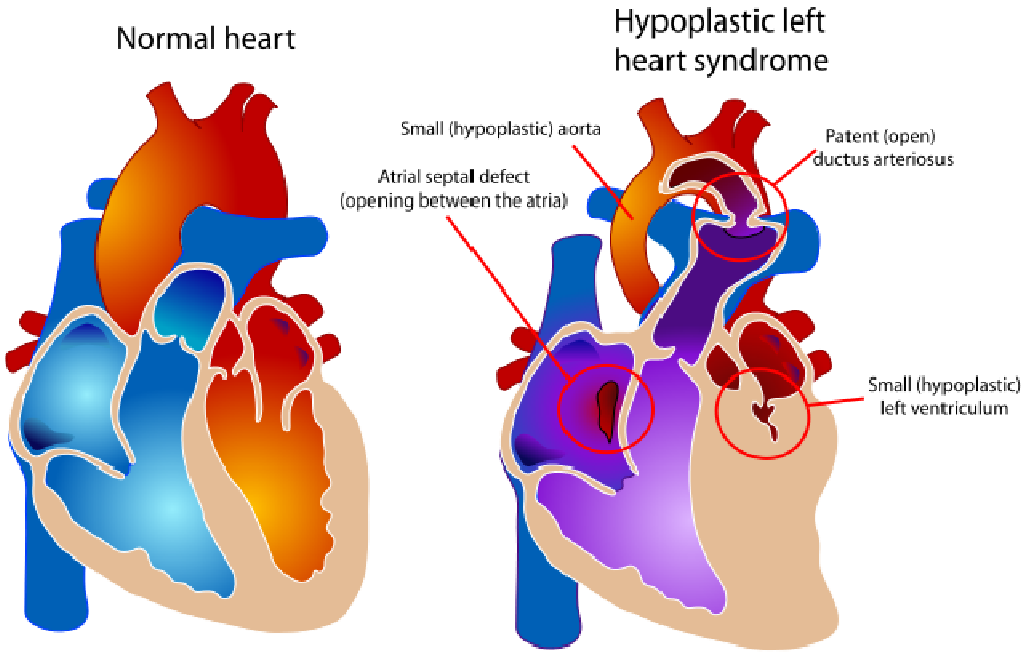


Table 4.1: List of genes causing human phenotypes overlapping LVOTO. HLHS = hypoplastic left heart syndrome ; AS = aortic stenosis; CoA = Coarctation of the aorta. Shone complex includes supravalvular mitral membrane, parachute mitral value, subaortic stenosis and CoA.

Gene	HLHS	AS	CoA	IAA	Shone complex
ACVR2B	✓	.	.	.	.
BRAF	.	.	✓	.	.
CBL	.	.	✓	.	.
CHD7	.	.	.	✓	.
CREBBP	✓	.	✓	.	.
DHCR7	✓	.	✓	.	.
DTNA	✓	.	.	.	.
EVC,EVC2	✓	.	.	.	.
FGD1	.	✓	✓	.	.
FOXC1	.	✓	.	.	.
FOXF1-FOXC2-FOXL1 deletion	✓	.	.	.	.
JAG1	.	.	✓	.	.
KCNJ2	.	.	✓	.	.
KRAS	.	.	✓	.	.
LEFTY2	✓	.		.	.
MKS1	.	.	✓	.	.
MLL2	.	.	✓	.	.
NF1	.	.	✓	.	.
NOTCH2	.	.	✓	.	.
NRAS	.	.	✓	.	
NSDHL	.	.	✓	.	✓
PEX genes	✓	.		.	.
PTPN11	.	.	✓	.	.
RAF1	.	.	✓	.	.
SHOC2	.	.	✓	.	.
SOS1	.	.	✓	.	.
TBX1	.	.	.	✓	.
TBX5	✓	.	.	.	.
ZIC3	✓	.	✓	.	.

resources (1000 genomes, ESP, ExAC) and 5,492 individuals from the ARIC study without severe cardiac malformation sequenced on the same platform. We first constructed a list of *a priori* candidate cardiac malformation genes which includes those implicated in similar human disorders, overlapping phenotypes in model organism knockouts, and expression in the developing heart tissue. The intersection of this candidate gene filter with rare loss-of-function (LOF) variation within cases identified implicating mutations in 9% of LVOTO cases, including nine *de novo* point mutations and three genes with recessive or hemizygous inheritance.

These results highlight the complex genetics contributing to LVOTO, and the utility of exome sequencing in a large informative sample set to identify novel genes or gene mutations for a rare disease. The general analysis framework we present may also apply to similar sequence-based analyses of rare disease cohorts of unrelated individuals (**Figure 4.2**).

Methods

Sample Selection

This study included 342 unrelated LVOTO cases without known extracardiac malformations or unexplained developmental delay ascertained from the Texas Medical Center in Houston, TX (**Table 4.2**). Parents and affected family members (if any) of LVOTO cases were also recruited as study participants. Our analyses also included 5,492 European American (EA) individuals from the population-based Atherosclerosis Risk in Communities (ARIC) study[40] as a comparison group for variant filtering and statistical analyses. ARIC samples with any of the following criteria were excluded from these analyses: prevalent heart failure, major Q-wave, or LVH by the Cornell definition. In addition, as a validation/replication set we examined 4,750 independent cases referred for clinical exome sequencing at the Baylor Miraca

Figure 4.2. Discovery strategy for LVOTO cohort. Imposing a candidate list, constructed independently, of disease gene candidates over exome-wide analyses facilitates genes discovery in rare disease cohorts.

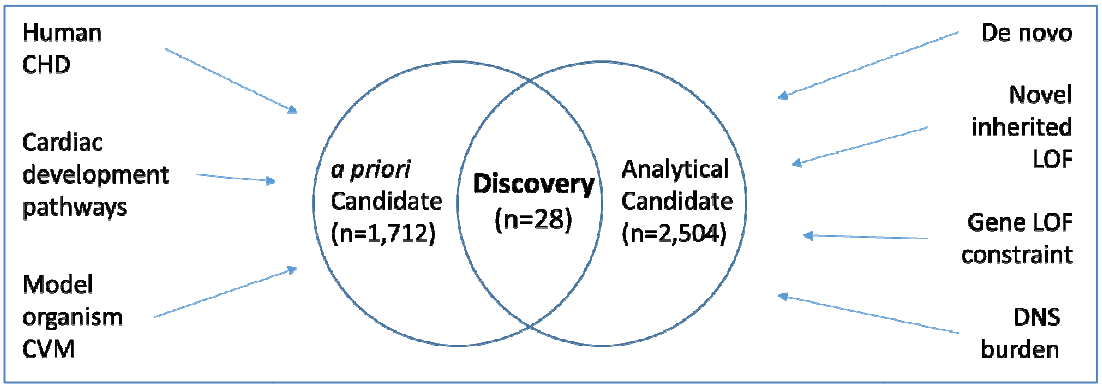


Table 4.2. Overview of LVOTO cases. This table summarizes basic demographical and clinical information of 342 LVOTO probands.

	Sex		
Ethnicity	**Female**	**Male**	**Total**
African	0	1	1
Caucasian	83	163	246
Hispanic	25	70	95
Total	108	233	342

Genetics Laboratory (http://www.bmgl.com) for rare LOF variants in a subset of genes given priority after analysis of the initial research discovery set of 342 cases.

Whole-exome sequencing

Whole exome sequencing (WES) was performed on cases and comparison samples with the Illumina HiSeq platform using the Mercury pipeline[41]. ARIC samples were captured using VCRome 2.1 (42Mb) reagents with an average coverage of 88x, LVOTO cases were captured using HGSC core (52Mb), and all analyses were restricted to exonic regions shared between these two reagents. Read mapping to Genome Reference Consortium Human Build 37 (GRCh37) was performed with Burrows-Wheeler alignment[42], and allele calling was performed with the Atlas2 suite (Atlas-SNP, Atlas-Indel)[43]. The Variant Call File (VCF) contained flags for low-quality variants which were excluded from all analyses, including SNPs with poster probability lower than 0.95, total depth of coverage less than 10x, fewer than 3 variant reads, allelic fraction less than 10%, 99% reads in a single directions, and homozygous reference alleles with < 6x coverage. In addition, we removed low-quality indels with a total depth less than 30x and allelic fraction below 30%. Individuals presenting extremely high or low numbers of heterozygous variant sites (6 standard deviations) were flagged and excluded from the burden analyses.

***A priori* gene prioritization**

To facilitate novel gene and variant discovery, we compiled *a priori* evidence from public resources to identify potential novel LVOTO genes. We compiled a list of 1,712 human genes with a putative role in the development of cardiovascular malformation from a variety of

public resources**.** Genes related to overlapping human disorders including CVM were ascertained from NCBI and literature searches. Relevancy to biological pathways and interactions (Hedgehog, *NOTCH*, *TGFB*, *PITX2*) was determined using the Kyoto Encyclopedia of Genes and Genomes (KEGG) database[8081]. Two model organism databases (ZFIN[82], MGI[83]) were also used to ascertain additional genes with a potential role in cardiovascular development. ZFIN was queried for genes expressed in the zebrafish heart, and MGI was queried for genes causing abnormal cardiac morphology in mouse models (MP:0000266). Additional quantitative measures used to prioritize genes included (a) measures of observed LOF prevalence (OP ratio[25]), (b) tolerance to functional variation (RVIS[48]) and (c) the probability of de novo mutation[84]. Gene expression in the heart was gathered from the literature[79] and Tissue-specific Gene Expression and Regulation (TiGER) database (http://bioinfo.wilmer.jhu.edu/tiger/).

Variant annotation

Variants were annotated to Refseq gene definitions using ANNOVAR[85]. Conservative loss-of-function (LOF) annotation was performed by selecting premature stopgains in the non-terminal exon, variants disrupting essential splice sites used by all gene isoforms, and frameshift indels similarly mapping to all isoforms. Damaging nonsynonymous (DNS) variation was defined as protein-altering substitutions predicted to be damaging by a consensus of at least 3 out of 6 prediction scores downloaded via dbNSFP[29] (SIFT, Polyphen2 HDIV, LRT, Mutation Taster, Mutation Assessor, FATHMM). A PHRED-like scaled C-score (CADD[86]) was also used to assess pathogenicity of variants (LOF and DNS), but was not used to exclude candidate sites.

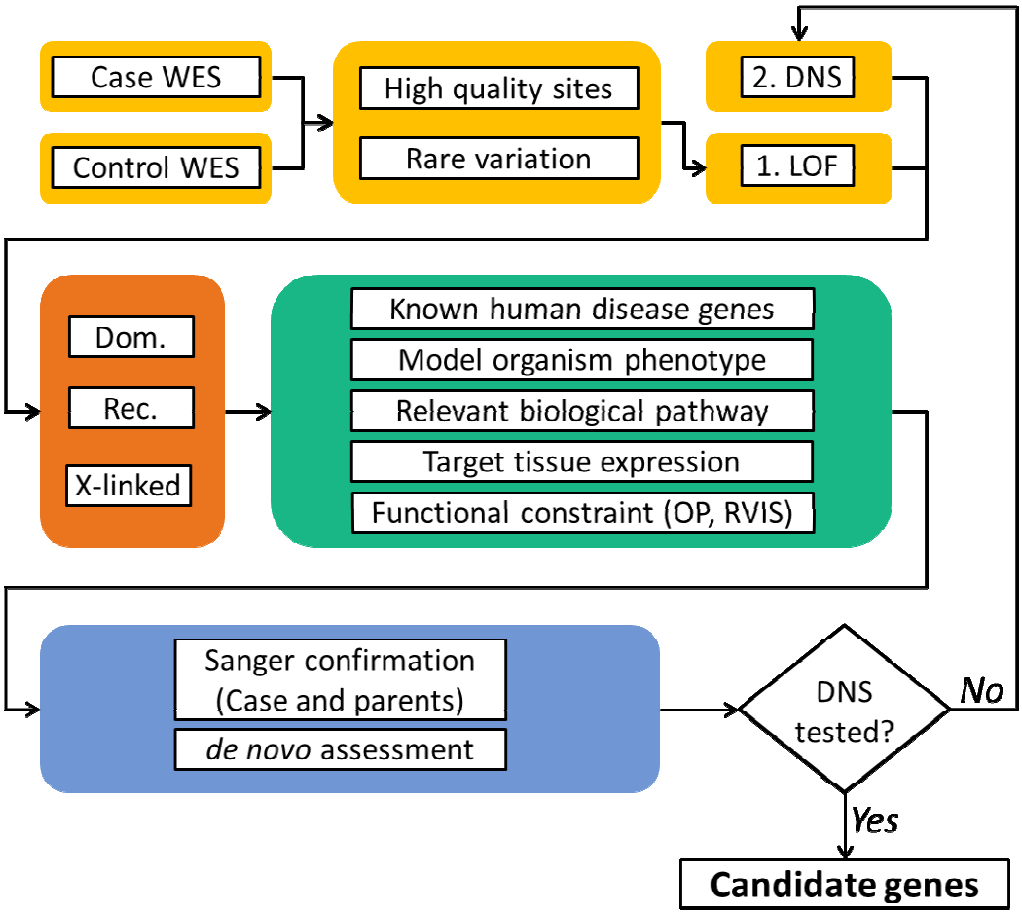
Analytical methods

Initially, we focused on the most damaging class of variation (extremely rare LOF) and considered multiple modes of Mendelian inheritance. We first identified all LOF sites exclusive to cases which included heterozygous (dominant), homozygous (recessive), hemizygous (X-linked males) or multiple compound-heterozygous genotypes in a given gene (recessive). Fisher's exact test was used to compare allele frequencies in LVOTO cases to ARIC study participants. Within the set of genes presenting case-exclusive LOF variation, we next performed "functional expansion" to include similarly-segregating DNS sites, which were analyzed by the same methods (**Figure 4.3**).

Firth logistic regression[87] was used to assess more complex genetic models by grouping rare DNS variation (observed MAF $\leq$1%). Using case-control status as an outcome, the total number of heterozygous sites per individual was included in the model as a covariate to address platform differences between sequencing batches. We defined a p-value of 2.86×10^{-6} as statistically significant for burden analyses by performing a Bonferroni correction for the number of genes harboring DNS (n=17,487).

After these exome-wide analyses, we further enriched for genes likely to contribute to LVOTO in two ways. First, for each gene the sum of all observed LOF alleles in ARIC was calculated (gene-wise observed LOF) and compared to all potential simulated LOF sites to calculate the ratio of observed to potential LOF alleles (OP ratio[25]). Only genes with a very low OP ratio (zero, or lowest 30th percentile) were considered strong candidates for disease. In addition, we filtered for the set of cardiac genes compiled *a priori* (**Figure 4.1**) to identify those with supporting evidence for a role in LVOTO.

Figure 4.3: Detailed analytical framework to assess in rare disease cohorts. This figure depicts a 2-stage analysis strategy used to assess whole-exome sequence (WES), with emphasis on LOF variation and follow-up using DNS variation. In the variant selection stage (yellow), high quality sites are identified, annotated for functional effect and population frequency. Rare LOF variation is analyzed by three inheritance models (orange) to identify case-exclusive LOF sites. Next, gene prioritization (green) compares an *a priori* list of gene candidates to identify the best gene candidate with both genetic and biological support. Variants are then validated (blue) in probands and parents when available to detect de novo inheritance. Finally, functional expansion into DNS alleles is then performed within the set of genes identified by LOF.



All variants identified to be potentially related to the development of LVOTO were validated via Sanger sequencing. Only those variants sites that did validate are reported here.

Results

Exome sequence variation

WES of LVOTO cases initially revealed 243,609 variants within the VCRome capture regions (239,726 single nucleotide substitutions, and 3,883 small indels ranging from -51 to +26 nucleotides in length). On average, each case presented with 14,669 heterozygous and 8,321 homozygous non-reference genotypes (**Table 4.3**). Thirty samples presented extremely high or low heterozygosity (4 cases, 26 controls; beyond 6 standard deviations from sample mean) and were excluded from burden analyses.

Annotation of exonic sites to multiple population data resources (ESP, 1000 Genomes, ExAC) revealed that 132,182 of the total LVOTO variants sites (129,329 SNV, 2,853 indel) were either novel or extremely rare (MAF 0.5%) in the comparison groups. Functional annotation revealed 4,161 rare predicted LOF variants (1,469 premature stop, 602 splice, and 2,090 frameshift) in 1,660 genes, and 34,100 DNS variants in 11,822 genes. The mean number of these rare variants per LVOTO patient was 54.4 LOF and 118.8 DNS (range LOF = 35-74; range DNS = 88-158, **Figure 4.4**).

LOF variation in LVOTO cases reveals known and novel cardiac genes

Twenty nine genes from our *a priori* cardiac gene candidate list harbored case-exclusive LOF alleles (**Table 4.4**). Sanger sequencing of these alleles in parents revealed nine to have arisen *de novo*, all in different genes (ACVR1, *JARID2*, *KMT2D*, *NF1*, *NR2F2*, *PLRG1, SMURF1, TBX20*, and *ZEB2*). Mutations in *NF1* (MIM 162200), *NR2F2* (MIM 615779), *TBX20*

Table 4.3: Summary of non-reference genotypes in exome sequence samples. This table describes the average value per individual, with standard deviation in parenthesis.

	ARIC	**LVOTO**
Het	14,493.97 (587.73)	14,669.61 (986.81)
Homo	8,204.09 (301.99)	8,321.78 (534.60)
Ti	22,895.69 (744.71)	23,084.29 (1,442.65)
Tv	7,582.58 (264.27)	7,669.32 (485.89)
TiTv	3.02 (0.03)	3.01 (0.031)

Figure 4.4: Distribution of rare sites within LVOTO cases. These histograms depict the number of rare (a) DNS and (b) LOF sites per sample.

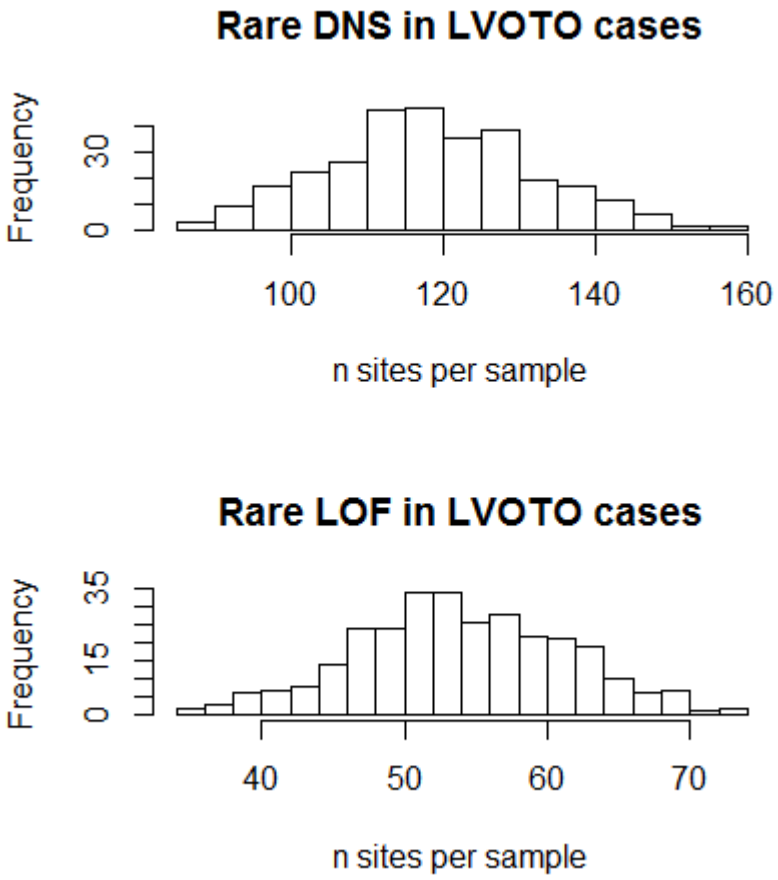


Table 4.4: Discovery genes presenting case-exclusive LOF sites and evidence for a role in LVOTO. This table highlights 29 genes with rare LOF sites in LVOTO cases and supporting evidence for a role in cardiac malformation. The number of samples in the Baylor Miraca lab presenting CVM and LOF alleles, distinct from LVOTO cases, is provided. One LVOTO case presenting mutation in multiple candidates is denoted (□). Unknown inheritance indicates only parent was available for validation that did not carry the mutation. Gene support symbols are defined as follows: CVM = known role in human cardiovascular malformation, MGI = overlapping phenotype in mouse, HE = human heart expressed ; PITX2 = related to PITX2 transcription; ZFIN = overlapping phenotype in zebrafish; TGFB = TGFB pathway.

LVOTO mode	Gene	Chrom.	LVOTO cases	BCM Miraca CVM	Gene support	LOF OP (%ile)
de novo	*ACVR1*	2	1	0	MGI,HE,TGFB	0.201
	JARID2	6	1	1	MGI,PITX2	0
	KMT2D	12	1	7	CVM	0
	NF1	17	1	1	CVM,MGI,HE	0.166
	NR2F2	15	1	0	MGI	0
	PLRG1	4	1	0	MGI	0
	SMURF1	7	1	1	TGFB	0
	TBX20	7	1	0	CVM,MGI,HE	0
	ZEB2	2	1	6	ZFIN, CVM	0
Inherited	*ARHGEF11*	1	1	2	ZFIN	0.156
	CCDC91	12	1	1	PITX2	0
	CDH2	18	1	0	MGI,HE,PITX2	0.077
	E2F6	2	1	0	PITX2	0
	FGF19	11	1	0	MGI	0
	GJC1	17	1	0	MGI,ZFIN	0
	GLRX3	10	1	0	MGI,HE	0
	LATS2	13	1	2	MGI	0.119
	LTBP1	2	1	1	MGI,TGFB	0
	PCDHGA2	5	1	0	HE	0
	PCSK6	15	1	3	MGI	0
	RAC1	7	1	1	MGI,HE	0
X-linked	*OFD1*	X	1	1	CVM,MGI,ZFIN	0
Recessive	*DNAH5*	5	1	0	CVM,MGI,HE,PCD	0.299
	MNDA	1	1	0	CHD candidate	0.265
Unknown	*JMJD6*	17	1□	0	MGI	0
	BMP1	8	1	0	MGI	0
	KMT2D	12	1	7	CVM	0
	ROCK1	18	1□	0	MGI,TGFB,PITX2	0

(MIM 611363) and *ZEB2* (MIM 235730) are known to cause human genetic disorders, and cardiac malformations in these syndromes occur in 3% to 50% of patients. *SMURF1* is involved in the *TGFB* pathway, *JARID2* is regulated by the *PITX2* transcription factor that has been associated with cardiac malformation, and mutant alleles in *PLRG1* cause malformation of the left ventricle in mouse model systems[88]. We expanded our evaluation of these genes to include DNS variants that were absent in controls or in public data resources. In this analysis, we identified three additional *de novo* variants (*KMT2D, TBX20, ZEB2*), providing further evidence for their role as primary LVOTO genes (**Table 4.5**).

Transmitted LOF variants were present in the heterozygous state, with the exception of *DNAH5, MNDA, and OFD1*. One LVOTO case presented a homozygous LOF variant in DNAH5, which is reported to cause ciliary dyskinesia, primary, 3, with or without situs inversus (MIM 608644), a known autosomal recessive condition. Two frameshift mutations were observed within in another case located in *MNDA* (Sanger validated, trans-inherited), a gene which has been previously suggested as an LVO candidate[79]. Mutations in *OFD1* (X-linked) cause Joubert syndrome, which can include congenital heart malformation within its phenotypic spectrum[89]. These observations support recessive and sex-linked forms of LVOTO, highlighting the complex genetics underlying this disease.

Aggregation of DNS variation

Expanding the scope of inheritance models, we next grouped rare variation by genes and pathways to detect enrichment of deleterious alleles in the sample of LVOTO cases. Aggregating DNS variation by gene revealed significant enrichment of rare alleles in 19 genes (MAF<1%, $p < 2.86 \times 10^{-6}$), including one gene (*DDX11*) present in the list of *a priori*

Table 4.5 . List of all Sanger-validated sites in LVOTO cases. Variant fields (":" delimited) are chromosome, hg19 position, reference allele, alternative allele; CADD v1.2 PHRED-like scales scores obtained online (http://cadd.gs.washington.edu/score); RVIS downloaded from Petrovski et al[48]; OP ARIC ratio calculated from ARIC EA samples; OP ExAC from v0.3, total AC field. "NA" in OP ratio field indicates zero LOF alleles observed in this gene by our custom stringent annotation methods.

Inheritance	Gene	Variant	Variant Type	n	CADD	RVIS (%ile)	OP ARIC (%ile)	OP ExAC (%ile)
inherited	*NOTCH2*	1:120458122:A:T	DNS	1	19.38	2.15	0.011	0.978
inherited	*ARHGEF11*	1:156905835:A:AG	frameshift	1	35	2.5	0.156	0.131
inherited	*MNDA*	1:158812053:CA:C	frameshift	1	19.18	67.3	0.265	0.869
inherited	*MNDA*	1:158812100:GAAAA:G	frameshift	1	23.3	67.3	0.265	0.869
unknown	*ROR1*	1:64644454:C:G	stopgain	1	37	3.05	NA	0.107
inherited	*GLRX3*	10:131959078:C:T	stopgain	1	19.57	48.35	NA	0.176
inherited	*LGR4*	11:27390487:TTAGGATGCCAG:T	frameshift	1	37	34.88	0.128	0.51
inherited	*FGF19*	11:69514190:TGA:T	frameshift	1	35	NA	NA	0.12
inherited	*CCDC91*	12:28459810:C:T	stopgain	1	38	91.26	NA	0.337
de novo	*KMT2D*	12:49416115:G:A	stopgain	1	22.7	NA	NA	NA
de novo	*KMT2D*	12:49420607:G:A	DNS	1	24.7	NA	NA	NA
unknown	*KMT2D*	12:49445202:C:CG	frameshift	1	23.1	NA	NA	NA
inherited	*LATS2*	13:21557671:TTGTC:T	frameshift	1	37	74.71	0.119	0.155
inherited	*SGCG*	13:23808789:C:T	DNS	1	34	43.77	NA	0.493
inherited	*PCSK6*	15:101866685:C:T	splicing	1	19.11	NA	NA	0.161
de novo	*NR2F2*	15:96875777:G:T	splicing	1	22.2	26.23	NA	NA
de novo	*NF1*	17:29562641:C:T	stopgain	1	41	0.47	0.166	0.916
inherited	*NF1*	17:29670054:C:T	DNS	1	22.3	0.47	0.166	0.916
inherited	*NF1*	17:29677234:G:A	DNS	1	22.3	0.47	0.166	0.916
inherited	*GJC1*	17:42882475:AC:A	frameshift	1	35	44.89	NA	NA

inherited	*JMJD6*	17:74714943:C:A	frameshift	1	27.2	18.44	NA	0.269
unknown	*ROCK1*	18:18629813:ATC:A	frameshift	1	35	18.9	NA	0.222
inherited	*CDH2*	18:25564963:C:T	splicing	1	33	7.52	0.077	0.257
de novo	*SMAD7*	18:46474796:G:T	DNS	1	18.05	NA	NA	0.907
unknown	*GATAD2A*	19:19576173:C:T	stopgain	1	36	9.27	NA	0.109
inherited	*E2F6*	2:11587812:CTT:C	frameshift	1	23.2	56.64	NA	0.234
de novo	*ZEB2*	2:145153979:C:A	stopgain	1	48	10.03	NA	0.011
de novo	*ZEB2*	2:145187539:T:C	DNS	1	22.3	10.03	NA	0.011
de novo	*ACVR1*	2:158637081:G:T	stopgain	1	35	21.41	0.201	0.205
inherited	*LTBP1*	2:33526713:T:G	splicing	1	25.5	24.47	NA	0.12
de novo	*PLRG1*	4:155460345:G:A	stopgain	1	22.3	33.2	NA	0.856
inherited	*DNAH5*	5:13868103:TA:T	frameshift	1	0.781	40.16	0.299	0.969
inherited	*PCDHGA2*	5:140720879:AGCCAG:A	frameshift	1	35	70.78	NA	NA
inherited	*JARID2*	6:15374390:C:T	DNS	1	21.6	2.42	NA	0.448
inherited	*JARID2*	6:15487603:G:C	DNS	1	22.5	2.42	NA	0.448
inherited	*JARID2*	6:15496604:A:G	DNS	1	24.5	2.42	NA	0.448
unknown	*JARID2*	6:15496768:C:T	DNS	1	35	2.42	NA	0.448
de novo	*JARID2*	6:15508596:TG:T	frameshift	1	38	2.42	NA	0.448
de novo	*TBX20*	7:35244085:G:A	stopgain	1	39	12.77	NA	0.177
inherited	*TBX20*	7:35244154:G:A	DNS	1	34	12.77	NA	0.177
de novo	*TBX20*	7:35289584:A:G	DNS	1	19.24	12.77	NA	0.177
inherited	*RAC1*	7:6438350:G:A	splicing	1	27.8	41.25	NA	NA
de novo	*SMURF1*	7:98649018:CTG:C	frameshift	1	36	4.39	NA	0.292
inherited	*SMURF1*	7:98649062:G:C	DNS	1	22	4.39	NA	0.292
unknown	*BMP1*	8:22058684:C:T	stopgain	1	19.09	1.84	NA	0.126
inherited	*OFD1*	X:13785314:C:T	stopgain	1	a	40.56	NA	0.884
inherited	*TAZ*	X:153649060:G:A	DNS	1	22.6	59.76	NA	NA

Table 4.6: Overview of candidate LVOTO genes detected by gene-based aggregation. This table highlights 19 genes where cases present excess rare (MAF<1%) DNS alleles compared to ARIC. Of note, *DDX11* is implicated in human cardiovascular malformation and similar mouse phenotypes, while the potential biological role for the remaining genes remains unclear. Sites = number of DNS sites in this gene; p = T05 Burden test p-value; MAC = minor allele count; CVM = cardiovascular malformation; MGI = mouse genome informatics.

Gene	**p**	**beta**	**se**	**DNS sites**	**MAC** case	**MAC** control	**Evidence**
ZNF845	6.70E-60	4.76	0.29	3	63	16	unknown
ABCD1	4.86E-57	3.14	0.2	18	61	79	unknown
MAGEC1	1.09E-55	4.27	0.27	8	56	21	unknown
KMT2C	8.78E-49	1.96	0.13	149	93	346	unknown
OR4B1	5.00E-33	3.48	0.29	11	32	23	unknown
GPX1	2.11E-30	5.02	0.44	5	39	6	unknown
DHRS4L1	2.90E-27	2.44	0.23	9	33	70	unknown
FMN2	3.20E-26	2.22	0.21	26	36	97	unknown
CHIT1	9.55E-24	2.83	0.28	18	24	33	unknown
UMODL1	7.01E-19	2.23	0.25	26	26	53	unknown
AQP7	5.95E-18	2.34	0.27	21	21	48	unknown
ABCB1	2.87E-12	1.8	0.26	39	20	66	unknown
OR2T4	9.74E-12	3.76	0.55	6	15	5	unknown
DDX11	**1.36E-10**	**1.32**	**0.21**	**50**	**29**	**188**	**CVM,MGI**
NDUFV2	2.68E-08	3.05	0.55	9	6	8	unknown
ZAN	8.20E-08	1.28	0.24	43	19	110	unknown
FRG1B	4.03E-07	1.1	0.22	19	23	157	unknown
KIAA1377	8.15E-07	2.85	0.58	7	5	8	unknown
IL25	4.97E-06	2.48	0.54	4	5	11	unknown

congenital heart defect candidate genes (**Table 4.6**). *DDX11* encodes a DNA helicase which is a cause of Warsaw breakage syndrome (MIM #613398), a human disorder which includes ventral septal defects within its phenotypic spectrum. Additionally, mutation of this gene in mouse models leads to developmental defects including abnormal heart looping[90].

Clinical database supports novel LVOTO candidate genes

To identify additional CVM cases with similar underlying etiology, we surveyed an additional 4,750 patients referred for clinical diagnostic testing for LOF variation in the LVOTO discovery genes. Thirty individuals with LOF variation in LVOTO discovery genes also present cardiovascular malformation, including severe left ventricular dysfunction (*GRIP1*) hypoplastic left heart (*KMT2D*), heterotaxy syndrome (*ARHGEF11*) and more (**Table 4.7**). The overlapping gene set also includes four genes ascertained by de novo mutations (*JARID2*, *KMT2D*, *NF1*, *SMURF1*, *ZEB2*), bolstering support for these genes as contributors to congenital heart defects.

Discussion

Performing whole exome sequencing and analysis of a cohort of 342 unrelated LVOTO cases, we detected 29 LOF variants among 28 genes with strong potential for involvement in cardiac development. Our discovery gene set included six Mendelian genes previously implicated in congenital heart malformation (*DNAH5*, *KMT2D*, OFD1, *NF1*, *TBX20*, *ZEB2*) and 16 genes associated with overlapping cardiac phenotypes in mammalian models (*ACVR1*, *BMP1*, *CDH2*, *FGF19*, *GJC1*, *GLRX3*, *JARID2*, *JMJD6*, *LATS2*, *LTBP1*, *NR2F2*, *OFD*, *PCSK6*, *PLRG1*, *RAC1*, and *ROCK1*) (**Table 4.3**). Among these are nine *de novo* mutations which were verified through targeted parental sequencing.

Table 4.7: Samples referred for clinical exome sequencing with cardiovascular malformation. 30 patients suspected of having a rare disorder (not necessarily including cardiovascular symptoms) referred to the BCM/Miraca lab for clinical exome sequencing also presented congenital heart disease and LOF mutations in our set of 28 discovery genes. TOF = Tetralogy of Fallot; PDA = patent ductus arteriosis; ASD = atrial septal defects; VSD = ventral septal defects; PFO = Patent foramen ovale.

Gene	Posn	Nucleotide	AminoAcid	Zyg.	Sex	Age	CHD
ARHGEF11	1:156918388	c.29delA	p.E10fs	Het	F	4y	Nonsyndromic cardiomyopathy leading to heart transplantation
ARHGEF11	1:156909357	c.3958delT	p.Y1320fs	Het	F	0m	Heterotaxy syndrome, right-sided and dilated ascending aorta, TOF
CCDC91	12:28605463	c.978delA	p.R326fs	Het	F	2y	PDA
ERRFI1	1:8075667	NM_018948		Het	F	2y	Moderate secundum ASD, VSD, moderate tricuspid valve insufficiency, hypoplastic tricuspid valve, bicuspid aortic valve, moderately hypoplastic right ventricle, peripheral pulmonary artery stenosis
GRIP1	12:66747201	c.2999T>A	p.L1000X	Het	M	10m	Severe left ventricular dysfunction
JARID2	6:15520368	c.3628_3629del	p.1210_1210del	Het	M	9m	Aortic & pulmonic stenosis
KMT2D	12:49438290	c.4964-1_4978del	N/A	Het	F	1y	Mildly hypoplastic aortic arch, VSD, pulmonary hypertension
KMT2D	12:49427675	c.10813C>T	p.Q3605X	Het	F	14y	VSD, resolved
KMT2D	12:49426760	c.11728C>T	p.Q3910X	Het	M	0m	Hypoplastic left heart
KMT2D	12:49425821	c.12667C>T	p.Q4223X	Het	M	1m	Coarctation of the aorta
KMT2D	12:49425644	c.12844C>T	p.R4282X	Het	F	0m	Shone's complex with small left sided structures, mild coarctation, large VSD
KMT2D	12:49420607	c.15142C>T	p.R5048C	Het	M	9y	Shone syndrome, VSD, mitral valve stenosis, aortic arch hypoplasia
KMT2D	12:49416416	c.16295G>A	p.R5432Q	Het	M	2y	Bicuspid aortic valve
LATS2	13:21565531	c.355C>T	p.R119X	Het	M	54y	Myocardial cyst
LATS2	13:21563131	c.786_787del	p.262_263del	Het	M	3y	Aortic root dilation
LGR4	11:27493693	c.156delC	p.P52fs	Het	F	7m	Distal abdominal aorta narrowing
LTBP1	2:33172449	c.59delC	p.S20fs	Het	F	7m	Distal abdominal aorta narrowing
NF1	17:29496923	c.499_502del	p.C167fs	Het	M	10y	TOF
OFD1	X:13785314	c.2668C>T	p.R890X	Hem	M	1m	Situs inversus, possible VSD
PCSK6	15:101905203	c.1903_1904insAG	p.V635fs	Het	F	12y	PDA, PFO, cardiomegaly

PCSK6	15:101853668	c.2569_2570del	p.857_857del	Het	M	39y	Cardiac arrest, ventricular fibrillation; maternal history of sudden cardiac death
RAC1	7:6438350	c.282+1G>A	N/A	Het	M	8y	Dilated aortic root & abnormal aortic valve
SMAD7	18:46468987	c.41delA	p.K14fs	Het	F	10m	Dysplastic pulmonary valve
SMURF1	7:98647229	c.988C>T	p.R330X	Het	M	2y	Dextrocardia
ZEB2	2:145161630	c.643_659del17	p.Y215fs	Het	F	3y	Dysplastic pulmonic valve, small ASD, small anterior muscular VSD, trivial mitral regurgitation, trivial distal right PTAS
ZEB2	2:145157824	c.928_931delins8	p.Y310fs	Het	M	9m	ASD, VSD, narrow aortic arch, leaky tricuspid valve
ZEB2	2:145156682	c.2072G>A	p.W691X	Het	M	4y	VSD, pulmonic valve stenosis
ZEB2	2:145156671	c.2083C>T	p.R695X	Het	M	17y	Aortic stenosis
ZEB2	2:145154138	c.2908C>T	p.Q970X	Het	M	1y	Pulmonary valve stenosis
ZEB2	2:145154011	c.3034delA	p.S1012fs	Het	M	1y	Large flow PDA, secundum ASD, muscular VSD, pulmonic stenosis with dysplastic valve

By focusing on case-exclusive LOF variation which intersected an *a priori* candidate gene list, we identified contributing, if not causative, diagnosis for 7.9% (27/342) of our starting cohort. Refining candidate gene selection may increase the number of discoveries within large sample sets. Large-scale model organism knockout projects with deep phenotypes, such as the Knockout Mouse Phenotyping Program (KOMP2), have great potential to facilitate the identification of putative human disease genes. Similarly, functional expansion of variation to include DNS variation facilitates disease gene discovery. This value is demonstrated by the detection of excess DNS burden in *DDX11*, which identified risk alleles in an additional 26 (7.6%) LVOTO cases.

The results presented here offer insights into the complexity of the inheritance of abnormal cardiac development. *De novo* mutations have been previously reported in CVM cases[79], and the role of this non-Mendelian-inherited variation is also supported by 9 genes in the data reported here. In addition, it is well-established that up to 30% of parents of LVOTO probands will have a similar, often more subtle, left-sided lesion, of the type that is only apparent with imaging of the heart. Review of echocardiogram data for available parents of affected probands showed that some of the transmitted damaging variants were inherited from parents with milder LVOTO phenotypes (e.g. mild left sided lesions), whereas some were inherited from parents without any evidence of current cardiac involvement.

This study implicates rare LOF and DNS along a broad spectrum of known and postulated cardiac genes in the complex pathogenesis of LVOTO. These mutations may arise de novo or be inherited from parents with milder but overlapping forms of CVM as well as in apparently unaffected parents (perhaps reflecting incomplete penetrance)?. Our approach illustrates the value of integrating an appropriately matched control group with model organism

bioinformatics, which compliments a traditional family-based approach to assess inheritance with the goal of identifying genes implicated in rare human disorders.

Chapter 5: Synthesis and Discussion

Genes and LOF trends

The previous chapters have described the analytical approaches and results of our studies on the contribution on LOF variation to a broad spectrum of human traits, including common chronic disease biomarkers, levels of small molecular metabolites, and rare congenital cardiovascular malformation. Whole exome sequencing coupled with detailed annotation has proved useful for discovery of novel gene candidates contributing to a broad phenotype spectrum.

The function of certain genes during human development may be extremely sensitive to or intolerant of gene dosage variation during viable human development, and large-scale sequencing projects provide data to help identify these genes. Despite sequencing 8,554 ARIC study participants, we did not observe any LOF variation in 11,380 genes (**Figure 2.5**). Currently, the largest collection of publicly available exome sequence data (ExAC r0.3; n~65000 individuals) describes at least one LOF allele in ~83% of human genes (17,005 out of 20,319 protein-coding Ensembl genes), suggesting that most human genes may tolerate some level of LOF variation. As expected, many of the genes not tolerant to this variation represent essential pathways to cellular function, such as ribosome formation, ubiquitination, and splicing (**Table 5.1**). The prevalence of LOF variation will continue to be refined and characterized in ongoing sequencing projects, and provide valuable insights into the robustness of the genome to these mutations.

Metrics for gene tolerance to LOF variation will also be refined and informed by ongoing studies, especially those informed by empirical LOF frequency in populations. The OP ratio metric we developed in chapter 2 has already demonstrated potential for application towards prioritizing novel gene contributing to human health. In chapter 2 we described a lower OP

Table 5.1: Classification of genes intolerant to LOF variation. This table presents the top 8 KEGG pathways enriched for genes presenting no LOF alleles in ExAC. We identified 3,314 gene presenting no LOF alleles in the ExAC v0.3 database, and father selected 656 of these which (1) encode for a known protein product and (2) have at least two exons, making them eligible LOF candidates by our definition. This list of 656 genes was uploaded to the NIH DAVID (http://david.abcc.ncifcrf.gov/) online bioinformatic resource for annotation and pathway enrichment analysis.

KEGG pathway	Genes	
	Count	%
Ribosome	32	5.9
Ubiquitin mediated proteolysis	13	2.4
Proteasome	7	1.3
TGF-beta signaling pathway	8	1.5
Chemokine signaling pathway	13	2.4
Spliceosome	10	1.8
Neurotrophin signaling pathway	9	1.7
Thyroid cancer	4	0.7

ratio in human paralogs of mouse embryonic lethal genes, and in chapter 4 we used this metric to prioritize novel candidate genes for rare monogenic disease. Moving forward, the utility of such a metric can be bolstered with refined LOF annotation methods.

Refining and expanding LOF annotation

Our annotation strategy was designed to enrich for exonic variation predicted to induce EJC-mediated NMD degradation of mRNA transcripts. However, mRNA regulation is complex and other mechanisms that influence gene dosage were not included in this study. For example, the degradation of aberrant mRNA transcripts may be suppressed by miRNA silencing of essential EJC-mediated NMD factors[91]. In addition, EJC-independent NMD of transcripts has been reported, likely related to length of 3' UTR[92], which we did not consider in our analysis.

Expanding the genetic scope beyond protein-coding regions may capture additional classes of LOF variation. For example, whole genome sequencing can be used to detect large structural variants including large deletions that can span a significant portion of a reading frame (especially the first exon) and may prevent protein formation[24]. In addition, certain intergenic motifs are known to influence transcriptional efficiency. Mutations within gene enhancer or silencer regions may influence gene expression levels[93]. Many of these loci are described in the ENCODE project and these data are available to genome studies[94]. These additional categories of variation affecting gene function may bolster the study of rare LOF variation and human phenotypes, just as the inclusion of small frameshift indels in our analyses complimented single nucleotide substitutions with a predicted LOF effect.

In addition to adding new classes, there are also potential to refine the selection and annotation of more familiar categories of LOF variation. Recent mRNA sequencing studies

provide opportunities to study the effect of genetic variation on mRNA transcripts. The Gene-Tissue Expression pilot project has performed RNAseq data on 43 tissue types from 175 individuals[95], giving valuable insights into which transcript isoforms are expected in a given tissue. The approach in these chapters was conservative, selecting predicted LOF variation mapping to genomic regions used by all known isoforms for a given gene. However, an alternative approach would be to restrict to variants mapping to isoforms expressed in tissues known to influence the studied phenotype (ie, pancreas and insulin levels). It is important to note that while a target-tissue approach may prevent the over-filtering of LOF sites from analyses, it may limit the potential for discovery and the relevant tissue for a given phenotype may not be known *a priori*.

In addition, recent studies suggest that LOF sites may be capable of allele-specific expression (ASE) at a higher rate than non-LOF sites in coding regions[96]. At these loci, the observed ratio of diploid transcript levels is not equal, with expression from one chromosome dominating the other. These loci may confound the study of gene dosage, especially heterozygous LOF variation, as these loci may effectively "escape" degradation of the prematurely truncated transcript by upregulating the more functional transcript. Conservatively, LOF variants mapping to these loci could be omitted from analyses, or a more quantitative weight could be assigned based on the extent of allelic imbalance for a particular locus.

Applications & Future directions

Genes that influence human heath via dosage mutations have great potential to improve healthcare. Specific applications may vary for each gene, especially whether LOF variation is

associated with a protective or deleterious health effect, although some applications may apply in either case.

Family history is already an important clinical component of assessing the risk for an individual to develop certain diseases, and incorporating genetic variation is simply a refinement of this predictive approach. Common population polymorphisms conveying modest risk are already available to consumers through commercial genetic services, such as 23andMe (www.23andme.com). In the clinic, rare mutations causing Mendelian disorders may be used to assess risk within families with a history of disease, and can distinguish the individuals most likely to develop disease[97], but these sites may not be observed outside a few families. LOF variation bridges the gap between these two extremes, in that the variation may be ascertained more frequently than Mendelian mutations and the effects are larger than commonly applied GWAS SNPs.

The potential for drug discovery with large deeply-phenotyped cohorts is enormous. LOF studies identify genes in which the total transcript levels, rather than gene composition (amino acid sequence) contribute to health. As such, the drugs which are developed as a result of association studies may seek to influence overall gene levels, rather than mimic a specific amino acid substitution as many synthetic forms of insulin act[98]. This presents opportunities for multiple avenues (monoclonal antibodies, miRNA) of therapy. These drugs represent a transition into a genomic era of medicine, where treatment options move beyond symptom alleviation towards preventing the causes of disease before their onset.

References

1. Ritchie, M. D. The success of pharmacogenomics in moving genetic association studies from bench to bedside: Study design and implementation of precision medicine in the post-GWAS era. Hum. Genet. 131, 1615–1626 (2012).

2. Corder, E. & Saunders, A. Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. Science (80-.). (1993). at <http://www.sciencemag.org/content/261/5123/921.short>

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

Montgomery G.W., Meitinger T., McPherson R., McCarthy M.I., McArdle W., Masson D., Martin N.G., Marroni F., Mangino M., Magnusson P.K., Lucas G., Luben R., Loos R.J., Lokki M.L., Lettre G., Langenberg C., Launer L.J., Lakatta E.G., Laaksonen R., Kyvik K.O., Kronenberg F., König I.R., Khaw K.T., Kaprio J., Kaplan L.M., Johansson A., Jarvelin M.R., Janssens A.C., Ingelsson E., Igl W., Kees Hovingh G., Hottenga J.J., Hofman A., Hicks A.A., Hengstenberg C., Heid I.M., Hayward C., Havulinna A.S., Hastie N.D., Harris T.B., Haritunians T., Hall A.S., Gyllensten U., Guiducci C., Groop L.C., Gonzalez E., Gieger C., Freimer N.B., Ferrucci L., Erdmann J., Elliott P., Ejebe K.G., Döring A., Dominiczak A.F., Demissie S., Deloukas P., de Geus E.J., de Faire U., Crawford G., Collins F.S., Chen Y.D., Caulfield M.J., Campbell H., Burtt N.P., Bonnycastle L.L., Boomsma D.I., Boekholdt S.M., Bergman R.N., Barroso I., Bandinelli S., Ballantyne C.M., Assimes T.L., Quertermous T., Altshuler D., Seielstad M., Wong T.Y., Tai E.S., Feranil A.B., Kuzawa C.W., Adair L.S., Taylor H.A Jr., Borecki I.B., Gabriel S.B., Wilson J.G., Holm H., Thorsteinsdottir U., Gudnason V., Krauss R.M., Mohlke K.L., Ordovas J.M., Munroe P.B., Kooner J.S., Tall A.R., Hegele R.A., Kastelein J.J., Schadt E.E., Rotter J.I., Boerwinkle E., Strachan D.P., Mooser V., Stefansson K., Reilly M.P., Samani N.J., Schunkert H., Cupples L.A., Sandhu M.S., Ridker P.M., Rader D.J., van Duijn C.M., Peltonen L., Abecasis G.R., Boehnke M., Kathiresan S. Biological, clinical and population relevance of 95 loci for blood lipids. Nature 466, 707–713 (2010).

4. Wadelius, M., Chen, L. Y., Downes, K., Ghori, J., Hunt, S., Eriksson, N., Wallerman, O., Melhus, H., Wadelius, C., Bentley, D. & Deloukas, P. Common VKORC1 and GGCX polymorphisms associated with warfarin dose. Pharmacogenomics J. 5, 262–70 (2005).

5. Phillips, I. R. & Shephard, E. a. Flavin-containing monooxygenases: mutations, disease and drug response. Trends Pharmacol. Sci. 29, 294–301 (2008).

6. Boileau, C., Guo, D.-C., Hanna, N., Regalado, E. S., Detaint, D., Gong, L., Varret, M., Prakash, S. K., Li, A. H., d'Indy, H., Braverman, A. C., Grandchamp, B., Kwartler, C. S., Gouya, L., Santos-Cortez, R. L. P., Abifadel, M., Leal, S. M., Muti, C., Shendure, J., Gross M.S., Rieder M.J., Vahanian A., Nickerson D.A., Michel J.B., National Heart, Lung, and Blood Institute (NHLBI) Go Exome Sequencing Project, Jondeau G., Milewicz D.M. TGFB2 mutations cause familial thoracic aortic aneurysms and dissections associated with mild systemic features of Marfan syndrome. Nat. Genet. 1–8 (2012). doi:10.1038/ng.2348

7. Sheridan, C. Phase 3 data for PCSK9 inhibitor wows. Nat. Biotechnol. 31, 1057–8 (2013).

8. Robinson, J. G., Farnier, M., Krempf, M., Bergeron, J., Luc, G., Averna, M., Stroes, E. S., Langslet, G., Raal, F. J., Shahawy, M. El, Koren, M. J., Lepor, N. E., Lorenzato, C., Pordy, R., Chaudhari, U. & Kastelein, J. J. P. Efficacy and Safety of Alirocumab in Reducing Lipids and Cardiovascular Events. N. Engl. J. Med. 2015, 150315080052000 (2015).

9. Sabatine, M. S., Giugliano, R. P., Wiviott, S. D., Raal, F. J., Blom, D. J., Robinson, J., Ballantyne, C. M., Somaratne, R., Legg, J., Wasserman, S. M., Scott, R., Koren, M. J. & Stein, E. a. Efficacy and Safety of Evolocumab in Reducing Lipids and Cardiovascular Events. N. Engl. J. Med. 150315080057008 (2015). doi:10.1056/NEJMoa1500858

10. Crosby, J., Peloso, G. M., Auer, P. L., Crosslin, D. R., Stitziel, N. O., Lange, L. a, Lu, Y., Tang, Z., Zhang, H., Hindy, G., Masca, N., Stirrups, K., Kanoni, S., Do, R., Jun, G., Hu, Y., Kang, H. M., Xue, C., Goel, A., Farrall M., Duga S., Merlini P.A., Asselta R., Girelli D., Olivieri O., Martinelli N., Yin W., Reilly D., Speliotes E., Fox C.S., Hveem K., Holmen O.L., Nikpay M., Farlow D.N., Assimes T.L., Franceschini N., Robinson J., North K.E., Martin L.W., DePristo M., Gupta N., Escher S.A., Jansson J.H., Van Zuydam N., Palmer C.N., Wareham N., Koch W., Meitinger T., Peters A., Lieb W., Erbel R., Konig I.R., Kruppa J., Degenhardt F., Gottesman O., Bottinger E.P., O'Donnell C.J., Psaty B.M., Ballantyne C.M., Abecasis G., Ordovas J.M., Melander O., Watkins H., Orho-Melander M., Ardissino D., Loos R.J., McPherson R., Willer C.J., Erdmann J., Hall A.S., Samani N.J., Deloukas P., Schunkert H., Wilson J.G., Kooperberg C., Rich S.S., Tracy R.P., Lin D.Y., Altshuler D., Gabriel S., Nickerson D.A., Jarvik G.P., Cupples L.A., Reiner A.P., Boerwinkle E., Kathiresan S. Loss-of-function mutations in APOC3, triglycerides, and coronary disease. N. Engl. J. Med. 371, 22–31 (2014).

11. Adzhubei, I., Schmidt, S. & Peshkin, L. A method and server for predicting damaging missense mutations. Nat. … 7, 248–249 (2010).

12. Sim, N.-L., Kumar, P., Hu, J., Henikoff, S., Schneider, G. & Ng, P. C. SIFT web server: predicting effects of amino acid substitutions on proteins. Nucleic Acids Res. 40, W452–7 (2012).

13. Thusberg, J., Olatubosun, A. & Vihinen, M. Performance of mutation pathogenicity prediction methods on missense variants. Hum. Mutat. 32, 358–368 (2011).

14. Chang, Y.-F., Imam, J. S. & Wilkinson, M. F. The nonsense-mediated decay RNA surveillance pathway. Annu. Rev. Biochem. 76, 51–74 (2007).

15. Le Hir, H., Moore, M. J. & Maquat, L. E. Pre-mRNA splicing alters mRNP composition: Evidence for stable association of proteins at exon-exon junctions. Genes Dev. 14, 1098–1108 (2000).

16. Le Hir, H., Gatfield, D., Izaurralde, E. & Moore, M. J. The exon-exon junction complex provides a binding platform for factors involved in mRNA export and nonsense-mediated mRNA decay. EMBO J. 20, 4987–4997 (2001).

17. Amrani, N., Dong, S., He, F., Ganesan, R., Ghosh, S., Kervestin, S., Li, C., Mangus, D. a, Spatrick, P. & Jacobson, a. Aberrant termination triggers nonsense-mediated mRNA decay. Biochem. Soc. Trans. 34, 39–42 (2006).

18. Nagy, E. & Maquat, L. E. A rule for termination-codon position within intron-containing genes: When nonsense affects RNA abundance. Trends in Biochemical Sciences 23, 198–199 (1998).

19. Gossage, D. L., Norby-slycord, C. J., Hershflekj, M. S. & Marker, M. L. A homozygous 5 base-pair deletion in exon 10 of the adenosine deaminase (ADA) gene in a child with severe combined immunodeficiency and very low levels of ADA mRNA and protein. 2, 1493–1494 (1993).

20. Lejeune, F. & Maquat, L. E. Mechanistic links between nonsense-mediated mRNA decay and pre-mRNA splicing in mammalian cells. Curr. Opin. Cell Biol. 17, 309–315 (2005).

21. Brocke, K. S., Neu-Yilik, G., Gehring, N. H., Hentze, M. W. & Kulozik, A. E. The human intronless melanocortin 4-receptor gene is NMD insensitive. Hum. Mol. Genet. 11, 331–335 (2002).

22. Maquat, L. E. & Li, X. Mammalian heat shock p70 and histone H4 transcripts, which derive from naturally intronless genes, are immune to nonsense-mediated decay. RNA 7, 445–456 (2001).

23. Neu-Yilik, G., Amthor, B., Gehring, N. H., Bahri, S., Paidassi, H., Hentze, M. W. & Kulozik, A. E. Mechanism of escape from nonsense-mediated mRNA decay of human beta-globin transcripts with nonsense mutations in the first exon. RNA 17, 843–854 (2011).

24. MacArthur, D. G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., Jostins, L., Habegger, L., Pickrell, J. K., Montgomery, S. B., Albers, C. a, Zhang, Z. D., Conrad, D. F., Lunter, G., Zheng, H., Ayub, Q., DePristo, M. a, Banks, E., Hu, M., Handsaker R.E., Rosenfeld J.A., Fromer M., Jin M., Mu X.J., Khurana E., Ye K., Kay M., Saunders G.I., Suner M.M., Hunt T., Barnes I.H., Amid C., Carvalho-Silva D.R., Bignell A.H., Snow C., Yngvadottir B., Bumpstead S., Cooper D.N., Xue Y., Romero I.G., 1000 Genomes Project Consortium, Wang J., Li Y., Gibbs R.A., McCarroll S.A., Dermitzakis E.T., Pritchard J.K., Barrett J.C., Harrow J., Hurles M.E., Gerstein M.B., Tyler-Smith C. A systematic survey of loss-of-function variants in human protein-coding genes. Science 335, 823–8 (2012).

25. Li, A. H., Morrison, A. C., Kovar, C., Cupples, L. A., Brody, J. a, Polfus, L. M., Yu, B., Metcalf, G., Muzny, D., Veeraraghavan, N., Liu, X., Lumley, T., Mosley, T. H., Gibbs, R. a & Boerwinkle, E. Analysis of loss-of-function variants and 20 risk factor phenotypes in 8,554 individuals identifies loci influencing chronic disease. Nat. Genet. 47, 640–642 (2015).

26. Cohen, J. & Boerwinkle, E. Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. New Engl. J. … 1264–1272 (2006). at <http://www.nejm.org/doi/full/10.1056/NEJMoa054013>

27. Margaritte, P., Bonaiti-Pellie, C., King, M. C. & Clerget-Darpoux, F. Linkage of familial breast cancer to chromosome 17q21 may not be restricted to early-onset disease. Am. J. Hum. Genet. 50, 1231–4 (1992).

28. Liu, X., Jian, X. & Boerwinkle, E. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. Hum. Mutat. 32, 894–9 (2011).

29. Liu, X., Jian, X. & Boerwinkle, E. dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. Hum. Mutat. 34, E2393–402 (2013).

30. Dang, V. T., Kassahn, K. S., Marcos, A. E. & Ragan, M. a. Identification of human haploinsufficient genes and their genomic proximity to segmental duplications. Eur. J. Hum. Genet. 16, 1350–7 (2008).

31. Georgi, B., Voight, B. F. & Bućan, M. From mouse to human: evolutionary genomics analysis of human orthologs of essential genes. PLoS Genet. 9, e1003484 (2013).

32. Li, B. & Leal, S. M. Methods for Detecting Associations with Rare Variants for Common Diseases : Application to Analysis of Sequence Data. 311–321 (2008). doi:10.1016/j.ajhg.2008.06.024.

33. Cohen, J., Pertsemlidis, A., Kotowski, I. K., Graham, R., Garcia, C. K. & Hobbs, H. H. Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9. Nat. Genet. 37, 161–5 (2005).

34. Jin, Y., Sharma, A., Bai, S., Davis, C., Liu, H., Hopkins, D., Barriga, K., Rewers, M. & She, J.-X. Risk of type 1 diabetes progression in islet autoantibody-positive children can be further stratified using expression patterns of multiple genes implicated in peripheral blood lymphocyte activation and function. Diabetes 63, 2506–15 (2014).

35. Gizer, I. R., Ehlers, C. L., Vieten, C., Seaton-Smith, K. L., Feiler, H. S., Lee, J. V, Segall, S. K., Gilder, D. a & Wilhelmsen, K. C. Linkage scan of nicotine dependence in the University of California, San Francisco (UCSF) Family Alcoholism Study. Psychol. Med. 41, 799–808 (2011).

36. Barbaric, I., Miller, G. & Dear, T. N. Appearances can be deceiving: phenotypes of knockout mice. Brief. Funct. Genomic. Proteomic. 6, 91–103 (2007).

37. Schäffler, A. & Buechler, C. CTRP family: linking immunity to metabolism. Trends Endocrinol. Metab. 23, 194–204 (2012).

38. Sheridan, C. Phase 3 data for PCSK9 inhibitor wows. Nat. Biotechnol. 31, 1057–8 (2013).

39. Morrison, A. C., Voorman, A., Johnson, A. D., Liu, X., Yu, J., Li, A., Muzny, D., Yu, F., Rice, K., Zhu, C., Bis, J., Heiss, G., O'Donnell, C. J., Psaty, B. M., Cupples, L. A., Gibbs, R. & Boerwinkle, E. Whole-genome sequence-based analysis of high-density lipoprotein cholesterol. Nat. Genet. 45, 899–901 (2013).

40. The ARIC Investigators. THE ATHEROSCLEROSIS RISK IN COMMUNITIES (ARIC) STUDY□: DESIGN AND OBJECTIVES. Am. J. Epidemiol. 129, (1989).

41. Reid, J. G., Carroll, A., Veeraraghavan, N., Dahdouli, M., Sundquist, A., English, A., Bainbridge, M., White, S., Salerno, W., Buhay, C., Yu, F., Muzny, D., Daly, R., Duyk, G., Gibbs, R. a & Boerwinkle, E. Launching genomics into the cloud: deployment of Mercury, a next generation sequence analysis pipeline. BMC Bioinformatics 15, 30 (2014).

42. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26, 589–95 (2010).

43. Challis, D., Yu, J., Evani, U. S., Jackson, A. R., Paithankar, S., Coarfa, C., Milosavljevic, A., Gibbs, R. a & Yu, F. An integrative variant analysis suite for whole exome next-generation sequencing data. BMC Bioinformatics 13, 8 (2012).

44. Grove, M. L., Yu, B., Cochran, B. J., Haritunians, T., Bis, J. C., Taylor, K. D., Hansen, M., Borecki, I. B., Cupples, L. A., Fornage, M., Gudnason, V., Harris, T. B., Kathiresan,

S., Kraaij, R., Launer, L. J., Levy, D., Liu, Y., Mosley, T., Peloso, G. M., Psaty B.M., Rich S.S., Rivadeneira F., Siscovick D.S., Smith A.V., Uitterlinden A., van Duijn C.M., Wilson J.G., O'Donnell C.J., Rotter J.I., Boerwinkle E. Best practices and joint calling of the HumanExome BeadChip: the CHARGE Consortium. PLoS One 8, e68095 (2013).

45. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 38, e164 (2010).

46. Kelso, J., Visagie, J., Theiler, G., Christoffels, A., Bardien, S., Smedley, D., Otgaar, D., Greyling, G., Jongeneel, C. V., Mccarthy, M. I., Hide, T., Hide, W., Africa, S. & Technology, I. eVOC□: A Controlled Vocabulary for Unifying Gene Expression Data. 1222–1230 (2003). doi:10.1101/gr.985203.bioinformatics

47. Kamburov, A., Stelzl, U., Lehrach, H. & Herwig, R. The ConsensusPathDB interaction database: 2013 update. Nucleic Acids Res. 41, D793–800 (2013).

48. Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S. & Goldstein, D. B. Genic intolerance to functional variation and the interpretation of personal genomes. PLoS Genet. 9, e1003709 (2013).

49. Suhre, K. & Gieger, C. Genetic variation in metabolic phenotypes: study designs and applications. Nat. Rev. Genet. 13, 759–69 (2012).

50. Shah, S. H., Bain, J. R., Muehlbauer, M. J., Stevens, R. D., Crosslin, D. R., Haynes, C., Dungan, J., Newby, L. K., Hauser, E. R., Ginsburg, G. S., Newgard, C. B. & Kraus, W.

E. Association of a peripheral blood metabolic profile with coronary artery disease and risk of subsequent cardiovascular events. Circ. Cardiovasc. Genet. 3, 207–214 (2010).

51. Dennis, M. K., Field, A. S., Burai, R., Ramesh, C., Whitney, K., Bologa, C. G., Oprea, T. I., Yamaguchi, Y., Hayashi, S., Sklar, L. a, Hathaway, H. J., Arterburn, J. B. & Prossnitz, E. R. Metabolic Profiles Predict Adverse Events Following Coronary Artery Bypass Grafting. Am. Assoc. Thorac. Surg. 127, 358–366 (2012).

52. Shah, S. H., Kraus, W. E. & Newgard, C. B. Metabolomic profiling for the identification of novel biomarkers and mechanisms related to common cardiovascular diseases form and function. Circulation 126, 1110–1120 (2012).

53. Shin, S.-Y., Fauman, E. B., Petersen, A.-K., Krumsiek, J., Santos, R., Huang, J., Arnold, M., Erte, I., Forgetta, V., Yang, T.-P., Walter, K., Menni, C., Chen, L., Vasquez, L., Valdes, A. M., Hyde, C. L., Wang, V., Ziemek, D., Roberts, P., Xi L., Grundberg E., Multiple Tissue Human Expression Resource (MuTHER) Consortium, Waldenberger M., Richards J.B., Mohney R.P., Milburn M.V., John S.L., Trimmer J., Theis F.J., Overington J.P., Suhre K., Brosnan M.J., Gieger C., Kastenmüller G., Spector T.D., Soranzo N. An atlas of genetic influences on human blood metabolites. Nat. Genet. 46, 543–50 (2014).

54. Kettunen, J., Tukiainen, T., Sarin, A.-P., Ortega-Alonso, A., Tikkanen, E., Lyytikäinen, L.-P., Kangas, A. J., Soininen, P., Würtz, P., Silander, K., Dick, D. M., Rose, R. J., Savolainen, M. J., Viikari, J., Kähönen, M., Lehtimäki, T., Pietiläinen, K. H., Inouye, M., McCarthy, M. I., Jula A., Eriksson J., Raitakari O.T., Salomaa V., Kaprio J., Järvelin M.R., Peltonen L., Perola M., Freimer N.B., Ala-Korpela M., Palotie A., Ripatti S.

Genome-wide association study identifies multiple loci influencing human serum metabolite levels. Nat. Genet. 44, 269–276 (2012).

55. Ohta, T., Masutomi, N., Tsutsui, N., Sakairi, T., Mitchell, M., Milburn, M. V, Ryals, J. a, Beebe, K. D. & Guo, L. Untargeted metabolomic profiling as an evaluative tool of fenofibrate-induced toxicology in Fischer 344 male rats. Toxicol. Pathol. 37, 521–535 (2009).

56. Evans, A. M., DeHaven, C. D., Barrett, T., Mitchell, M. & Milgram, E. Integrated, nontargeted ultrahigh performance liquid chromatography/electrospray ionization tandem mass spectrometry platform for the identification and relative quantification of the small-molecule complement of biological systems. Anal. Chem. 81, 6656–6667 (2009).

57. Levey, A. S., Stevens, L. a, Schmid, C. H., Zhang, Y. L., Castro, A. F., Feldman, H. I., Kusek, J. W., Eggers, P., Van Lente, F., Greene, T. & Coresh, J. A new equation to estimate glomerular filtration rate. Ann. Intern. Med. 150, 604–612 (2009).

58. Yu, B., Li, A. H., Muzny, D., Veeraraghavan, N., de Vries, P. S., Bis, J. C., Musani, S. K., Alexander, D., Morrison, A. C., Franco, O. H., Uitterlinden, A., Hofman, A., Dehghan, A., Wilson, J. G., Psaty, B. M., Gibbs, R., Wei, P. & Boerwinkle, E. Association of Rare Loss-Of-Function Alleles in HAL, Serum Histidine: Levels and Incident Coronary Heart Disease. Circ. Cardiovasc. Genet. 8, 351–5 (2015).

59. Mahagita, C., Grassl, S. M., Piyachaturawat, P. & Ballatori, N. Human organic anion transporter 1B1 and 1B3 function as bidirectional carriers and do not mediate GSH-bile acid cotransport. Am. J. Physiol. Gastrointest. Liver Physiol. 293, G271–G278 (2007).

60. Mootha, V. K. & Hirschhorn, J. N. Inborn variation in metabolism. Nat. Genet. 42, 97–98 (2010).

61. Yu, B., Zheng, Y., Alexander, D., Morrison, A. C., Coresh, J. & Boerwinkle, E. Genetic Determinants Influencing Human Serum Metabolome among African Americans. PLoS Genet. 10, (2014).

62. Peterson, J. W., Boldogh, I., Popov, V. L., Saini, S. S. & Chopra, A. K. Anti-inflammatory and antisecretory potential of histidine in Salmonella-challenged mouse small intestine. Lab. Invest. 78, 523–534 (1998).

63. Watanabe, M., Suliman, M. E., Qureshi, A. R., Garcia-lopez, E. & Ba, P. Consequences of low plasma histidine in chronic kidney disease patients□: associations with inflammation , oxidative stress , and. 1860–1866 (2008).

64. Gerber, D. A. Low free serum histidine concentration in rheumatoid arthritis. A measure of disease activity. J. Clin. Invest. 55, 1164–1173 (1975).

65. Tousoulis, D., Antoniades, C. & Stefanadis, C. Assessing inflammatory status in cardiovascular disease. Heart 93, 1001–7 (2007).

66. Feng, R. N., Niu, Y. C., Sun, X. W., Li, Q., Zhao, C., Wang, C., Guo, F. C., Sun, C. H. & Li, Y. Histidine supplementation improves insulin resistance through suppressed

inflammation in obese women with the metabolic syndrome: a randomised controlled trial. Diabetologia 56, 985–94 (2013).

67. Drife, J. O. SLCO1B1 Variants and Statin-Induced Myopathy — A Genomewide Study. BMJ Br. Med. J. 318, 1565 (1999).

68. Zhao, Y., Peng, J., Lu, C., Hsin, M., Mura, M., Wu, L., Chu, L., Zamel, R., Machuca, T., Waddell, T., Liu, M., Keshavjee, S., Granton, J. & De Perrot, M. Metabolomic heterogeneity of pulmonary arterial hypertension. PLoS One 9, (2014).

69. Hoffman, J. I. & Kaplan, S. The incidence of congenital heart disease. J. Am. Coll. Cardiol. 39, 1890–1900 (2002).

70. Van Der Linde, D., Konings, E. E. M., Slager, M. a., Witsenburg, M., Helbing, W. a., Takkenberg, J. J. M. & Roos-Hesselink, J. W. Birth prevalence of congenital heart disease worldwide: A systematic review and meta-analysis. J. Am. Coll. Cardiol. 58, 2241–2247 (2011).

71. Pradat, P., Francannet, C., Harris, J. a. & Robert, E. The epidemiology of cardiovascular defects, Part I: A study based on data from three large registries of congenital malformations. Pediatr. Cardiol. 24, 195–221 (2003).

72. West, E. R., Xu, M., Woodruff, T. K. & Shea, L. D. Epidemiology of noncomplex left ventricular outflow tract obstruction malformations (aortic valve stenosis, coarctation of the aorta, hypoplastic left heart syndrome) in Texas, 1999-2001. October 28, 4439–4448 (2008).

73. Clark, E. B. Pathogenetic mechanisms of congenital cardiovascular malformations revisited. Semin. Perinatol. 20, 465–472 (1996).

74. Jefferies, J. L., Pignatelli, R. H., Martinez, H. R., Robbins-Furman, P. J., Liu, P., Gu, W., Lupski, J. R. & Potocki, L. Cardiovascular findings in duplication 17p11.2 syndrome. Genet. Med. 14, 90–94 (2012).

75. Ware, S. M. & Jefferies, J. L. New Genetic Insights into Congenital Heart Disease. (2012). doi:10.4172/2155-9880.S8-003.New

76. Garg, V., Kathiriya, I. S., Barnes, R., Schluterman, M. K., King, I. N., Butler, C. a, Rothrock, C. R., Eapen, R. S., Hirayama-Yamada, K., Joo, K., Matsuoka, R., Cohen, J. C. & Srivastava, D. GATA4 mutations cause human congenital heart defects and reveal an interaction with TBX5. Nature 424, 443–447 (2003).

77. McBride, K. L., Pignatelli, R., Lewin, M., Ho, T., Fernbach, S., Menesses, A., Lam, W., Leal, S. M., Kaplan, N., Schliekelman, P., Towbin, J. a & Belmont, J. W. Inheritance analysis of congenital left ventricular outflow tract obstruction malformations: Segregation, multiplex relative risk, and heritability. Am. J. Med. Genet. A 134A, 180–6 (2005).

78. Kerstjens-Frederikse, W. S., Du Marchie Sarvaas, G. J., Ruiter, J. S., Van Den Akker, P. C., Temmerman, A. M., Van Melle, J. P., Hofstra, R. M. W. & Berger, R. M. F. Left ventricular outflow tract obstruction: should cardiac screening be offered to first-degree relatives? Heart 97, 1228–1232 (2011).

79. Zaidi, S., Choi, M., Wakimoto, H., Ma, L., Jiang, J., Overton, J. D., Romano-Adesman, A., Bjornson, R. D., Breitbart, R. E., Brown, K. K., Carriero, N. J., Cheung, Y. H., Deanfield, J., DePalma, S., Fakhro, K. a, Glessner, J., Hakonarson, H., Italia, M. J., Kaltman, J. R., Kaski J., Kim R, Kline J.K., Lee T., Leipzig J., Lopez A., Mane S.M., Mitchell L.E., Newburger J.W., Parfenov M., Pe'er I., Porter G., Roberts A.E., Sachidanandam R., Sanders S.J., Seiden H.S., State M.W., Subramanian S., Tikhonova I.R., Wang W., Warburton D., White P.S., Williams I.A., Zhao H., Seidman J.G., Brueckner M., Chung W.K., Gelb B.D., Goldmuntz E., Seidman C.E., Lifton R.P. De novo mutations in histone-modifying genes in congenital heart disease. Nature 498, 220–3 (2013).

80. Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. Data, information, knowledge and principle: Back to metabolism in KEGG. Nucleic Acids Res. 42, 199–205 (2014).

81. Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H. & Kanehisa, M. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 27, 29–34 (1999).

82. Sprague, J., Bayraktaroglu, L., Clements, D., Conlin, T., Fashena, D., Frazer, K., Haendel, M., Howe, D. G., Mani, P., Ramachandran, S., Schaper, K., Segerdell, E., Song, P., Sprunger, B., Taylor, S., Van Slyke, C. E. & Westerfield, M. The Zebrafish Information Network: the zebrafish model organism database. Nucleic Acids Res. 34, D581–D585 (2006).

83. Blake, J. a., Bult, C. J., Eppig, J. T., Kadin, J. a. & Richardson, J. E. The Mouse Genome Database: Integration of and access to knowledge about the laboratory mouse. Nucleic Acids Res. 42, 810–817 (2014).

84. Samocha, K. E., Robinson, E. B., Sanders, S. J., Stevens, C., Sabo, A., McGrath, L. M., Kosmicki, J. a, Rehnström, K., Mallick, S., Kirby, A., Wall, D. P., MacArthur, D. G., Gabriel, S. B., DePristo, M., Purcell, S. M., Palotie, A., Boerwinkle, E., Buxbaum, J. D., Cook, E. H., Gibbs R.A., Schellenberg G.D., Sutcliffe J.S., Devlin B., Roeder K., Neale B.M., Daly M.J. A framework for the interpretation of de novo mutation in human disease. Nat. Genet. 46, 944–950 (2014).

85. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 38, e164 (2010).

86. Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M. & Shendure, J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet. 46, 310–5 (2014).

87. Firth, D. Bias Reduction of Maximum Likelihood Estimates. Biometrika Trust Stable. 80, 27–38 (1993).

88. Kleinridders, A., Pogoda, H.-M., Irlenbusch, S., Smyth, N., Koncz, C., Hammerschmidt, M. & Brüning, J. C. PLRG1 is an essential regulator of cell proliferation and apoptosis during vertebrate development and tissue homeostasis. Mol. Cell. Biol. 29, 3173–3185 (2009).

89. Elmali, M., Ozmen, Z., Ceyhun, M., Tokatlioğlu, O., Incesu, L. & Diren, B. Joubert syndrome with atrial septal defect and persistent left superior vena cava. Diagn. Interv. Radiol. 13, 94–96 (2007).

90. Cota, C. D. & García-García, M. J. The ENU-induced cetus mutation reveals an essential role of the DNA helicase DDX11 for mesoderm development during early mouse embryogenesis. Dev. Dyn. 241, 1249–59 (2012).

91. Wang, G., Jiang, B., Jia, C., Chai, B. & Liang, A. MicroRNA 125 represses nonsense-mediated mRNA decay by regulating SMG1 expression. Biochem. Biophys. Res. Commun. 435, 16–20 (2013).

92. Silva, A. L. & Romão, L. The mammalian nonsense-mediated mRNA decay pathway: to decay or not to decay! Which players make the decision? FEBS Lett. 583, 499–505 (2009).

93. Corradin, O., Saiakhova, A., Akhtar-Zaidi, B., Myeroff, L., Willis, J., Cowper-Sallari, R., Lupien, M., Markowitz, S. & Scacheri, P. C. Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. Genome Res. 24, 1–13 (2014).

94. The ENCODE Project Consortium. A user's guide to the Encyclopedia of DNA elements (ENCODE). PLoS Biol. 9, (2011).

95. GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. (2015).

96. Rivas, M. A., Pirinen, M., Conrad, D. F., Lek, M., Tsang, E. K., Karczewski, K. J., Maller, J. B., Kukurba, K. R., Deluca, D. S., Fromer, M., Ferreira, P. G., Smith, K. S., Zhang, R., Zhao, F., Banks, E., Poplin, R., Ruderfer, D. M., Purcell, S. M., Tukiainen, T., Minikel E.V., Stenson P.D., Cooper D.N., Huang K.H., Sullivan T.J., Nedzel J., GTEx Consortium; Geuvadis Consortium, Bustamante C.D., Li J.B., Daly M.J., Guigo R., Donnelly P., Ardlie K., Sammeth M., Dermitzakis E.T., McCarthy M.I., Montgomery S.B., Lappalainen T., MacArthur D.G. Effect of predicted protein-truncating genetic variants on the human transcriptome. Science (80-.). 348, (2015).

97. Cruchaga, C., Chakraverty, S., Mayo, K., Vallania, F. L. M., Mitra, R. D., Faber, K., Williamson, J., Bird, T., Diaz-Arrastia, R., Foroud, T. M., Boeve, B. F., Graff-Radford, N. R., St. Jean, P., Lawson, M., Ehm, M. G., Mayeux, R., Goate, A. M. & for the NIA-LOAD - NCRAD Family Study Consortium. Rare variants in APP, PSEN1 and PSEN2 increase risk for AD in late-onset Alzheimer's disease families. PLoS One 7, (2012).

98. Egan, A. G., Blind, E., Dunder, K., de Graeff, P. A., Hummer, B. T., Bourcier, T. & Rosebraugh, C. Pancreatic safety of incretin-based drugs--FDA and EMA assessment. N. Engl. J. Med. 370, 794–7 (2014).

Vita

Alexander Hung Li was born in Houston, Texas on February 6, 1983, the son of Joe Li and Beverly Li. After completing his work at Clear Lake High School, Houston, Texas in 2001, he entered the University of Rochester in Rochester, New York. He received the degree of Bachelor of Science with a major in Ecology and Evolutionary Biology in May 2005. For the next three years, he worked as a research technician in the Department of Cell Biology at the Baylor College of Medicine and the University of Texas Medical School. In May of 2008 he entered The University of Texas Graduate School of Biomedical Sciences at Houston. He received the degree of Master of Science in Biomedical Sciences in August 2010. In September of 2008 he entered the PhD program at The University of Texas Graduate School of Biomedical Sciences at Houston.

Permanent address:

1402 Richmond #107

Houston, Texas 77006