Metabolic Adaptations in Modern Human Populations: Evidence, Theory, and Investigation

Metabolic Adaptations in Modern Human Populations:

Evidence, Theory, and Investigation

A dissertation presented

by

Elizabeth Anne Brown

to

The Department of Human Evolutionary Biology

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Human Evolutionary Biology

Harvard University

Cambridge, Massachusetts

April 2015

Dissertation Advisors:
Professor Maryellen Ruvolo
Professor Pardis C. Sabeti

Author:
Elizabeth Anne Brown

Metabolic Adaptations in Modern Human Populations:

Evidence, Theory, and Investigation

ABSTRACT

Diverse climates, infectious agents, and subsistence patterns drove humans to adapt metabolically to different environments since the migration out of Africa 100,000 years ago. In this dissertation, I review current literature on the genetic underpinnings, and the molecular and physiological manifestations of these metabolic adaptations in diverse human populations. Then, I develop a theory regarding pregnancy as a critical period in life history that mediated recent selection on human metabolism. Finally, I investigate the function and evidence for selection of derived genetic variants at increased frequency in East Asian populations. I find multiple standing variants that increase expression of the gene *IVD* and increase the efficiency of leucine catabolism, which lie on positively selected haplotypes in East Asians. I use this research process as a model for how to develop and study novel hypotheses of human metabolic adaptation. Such adaptations often impact health in the modern environment, so more evolutionary research will provide useful guidance to the medical community in how to treat people from diverse ethnicities.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

I would like to thank my committee, Maryellen Ruvolo, Terence Capellini, and Pardis Sabeti, for their generosity and support during this endeavor. In addition, I would like to thank the many faculty members, postdocs, researchers, teachers, and graduate colleagues who have taught, inspired, and mentored me both prior to and during graduate school. Lastly, I would like to thank my supporting network of family members and friends for their kindness and love.

CHAPTER 1

INTRODUCTION

RECENT SELECTION ON HUMAN METABOLISM

Within the past 100,000 years, human populations migrated out of Africa encountering diverse climates, infectious agents, and food sources within and outside of Africa. Positive selection acting within this time frame drove changes in allele frequency that correlate with environmental variables and subsistence patterns (Hancock et al., 2011a; Angela M. Hancock et al., 2010), and produced long regions of homozygosity that surround derived alleles at high frequency (Grossman et al., 2010). Such regions are enriched for genes involved in immune response, sensory perception, and metabolism (Grossman et al., 2013). For example, Table 1 lists genes in regions under selection in the enriched Gene Ontology category for catalytic activity. This indicates that exposure to new infectious diseases, climates, and foods, along with cultural developments like increased population density, agriculture, and pastoralism, drove important selective change in humans.

Despite these patterns of selection in human genomes, few discrete genes and phenotypes of adaptive significance have been successfully pulled from these regions of selection. This is partially because such regions are large, containing many genes and variants for which function may be unknown or pleiotropic. In addition, limited information on environmental, cultural, and phenotypic differences across human populations add additional layers of uncertainty. However, well-studied examples of

Table 1. CMS genes in regions under selection in the GO category "catalytic activity"

AASDHPPT	CBL	FKBP2	MOGS	PSMA1	STT3A
ABCB9	CCS	FOLH1	MON2	PSMC3	SUGCT
ABCC8	CD44	FOXRED1	MSRB3	PSMD9	SUOX
ABCC9	CDC42BPA	FRS2	MST1	PTGS1	SUV420H1
ABCG4	CDC42BPG	FUT4	MST1R	PTP4A3	SYTL2
ACACB	CDC42EP2	GABARAPL2	MTMR2	PTPMT1	TAAR2
ACAD10	CDC45	GAL3ST3	MUS81	PTPN22	TAOK3
ACADS	CDK17	GALNT18	MVK	PTPN5	TBC1D15
ACAT1	CDK2	GALNT6	MYH9	PTPN9	TBC1D30
ACCS	CDK4	GANAB	MYLK	PTPRB	TBK1
ACCSL	CDKN1B	GATC	MYO1A	PTPRJ	TCIRG1
ACER3	CELA1	GDPD4	MYO1H	PTPRM	TDG
ACP2	CEMIP	GDPD5	MYO5C	PTPRO	TESC
ACSS3	CHD2	GIT2	MYO7A	PTPRQ	TGM4
ACVR1B	CHEK1	GLS2	MYO9A	PTPRR	THY1
ACVRL1	CHKA	GLT8D2	NAA40	PTS	TM7SF2
ACY3	CHPT1	GLYAT	NAALAD2	PUS3	TMBIM6
ADAMTS15	CHRM1	GLYATL1	NAALADL1	PUS7L	TMED2
ADAMTS20	CHST11	GLYATL2	NADSYN1	PYROXD1	TMEM225
ADAMTS6	CIT	GNPTAB	NARS2	QARS	TMPRSS12
ADAMTS8	COQ5	GNS	NAT10	RAB1B	TMPRSS13
ADAP2	CORO1C	GOLT1B	NCKAP1L	RAB21	TMPRSS4
ADAT1	CPM	GPX1	NDST2	RAB28	TMPRSS5
ADCY3	CPT1A	GUCY2C	NDUFA12	RAB30	TPH1
ADCY6	CRY2	GYLTL1B	NDUFC2	RAB35	TRAF2
ADH7	CS	GYS2	NDUFC2-KCTD14	RAB38	TRAF6
ADRBK1	CSAD	HAL	NDUFS8	RAB39A	TRAPPC4
AGAP2	CSNK2A3	HECTD4	NDUFV1	RAB3IL1	TRHDE
AGAP5	CSRNP2	HECW2	NEDD4	RAB5B	TRIAP1
AGBL2	CST6	HELB	NEK10	RAB6A	TRMT112
AK7	CTDSP2	HEPHL1	NEU3	RABGEF1	TRPT1
ALDH1L2	CTSC	HERC1	NF1	RAD51B	TRPV1
ALDH2	CTSW	HIPK3	NNMT	RAD9B	TTC9C
ALDH3B1	CUL5	HMBS	NOS1	RAG1	TUBA1A
ALDH3B2	CWF19L2	HPD	NOX4	RAPGEF3	TUBA1B
ALG10	CYB561A3	HRASLS2	NRG3	RARRES3	TUBA1C
ALG10B	CYB5R4	HSD17B12	NT5DC3	RASAL1	TUT1
ALG8	CYP27B1	HSD17B6	NTAN1	RASGRP2	TWF1
ALG9	CYP2R1	HSPB2	NUAK1	RCE1	TXNRD1
ALKBH3	DAGLA	HSPB8	NUDT4	RDH5	UBC
ALKBH8	DAK	IGHMBP2	NUDT8	RECQL	UBE2L6

Table 1 (Continued).

AMBRA1	DALRD3	IMMP1L	OAS1	RELN	UBE2N
AMDHD1	DAO	IMPDH2	OAS2	RERG	UBE3B
AMHR2	DARS	IRAK3	OGFOD2	REXO2	UBE4A
AMPD3	DCLRE1B	IRAK4	OTOG	RFC5	UNG
AMT	DDX10	ISCU	OTOGL	RIC8B	USP19
ANKK1	DDX11	ITPR2	OTUB1	RNF10	USP2
APEH	DDX25	IVD	OVCH1	RNF115	USP28
APOA1	DDX47	KAT5	P4HA3	RNF121	USP30
APOA5	DDX54	KCTD10	P4HTM	RNF141	USP35
ARAP1	DDX55	KCTD14	PA2G4	RNF34	USP4
ARFGAP2	DDX6	KDM4E	PAFAH1B2	RNF41	USP44
ARHGAP26	DENND5B	KIF21A	PAH	RPS3	USP47
ARHGAP42	DERA	KIF5A	PAK1	RPS6KA4	USP54
ARHGDIB	DGAT2	KLC2	PAMR1	RPS6KB2	UVRAG
ARHGEF12	DGKA	KLHL42	PAN2	RPUSD4	VNN1
ARHGEF17	DGKZ	KMT2D	PARN	RRAS2	XRCC6BP1
ARHGEF3	DHCR7	KSR2	PAX6	RRNAD1	XYLT1
ARHGEF38	DHX37	LALBA	PC	SBF2	ZDHHC13
ARIH2	DIABLO	LAMTOR1	PCMTD1	SC5D	ZDHHC17
ARL1	DIP2C	LARGE	PCSK7	SCYL1	ZDHHC24
ART4	DLAT	LCT	PDE11A	SCYL2	ZDHHC3
ASAP2	DLG2	LDHA	PDE1B	SDR9C7	ZDHHC5
ASRGL1	DNAH10	LDHAL6A	PDE2A	SDS	ZFP91
ATG3	DNAJC24	LDHB	PDE6H	SDSL	ZNF738
ATL3	DNM1L	LDHC	PDHX	SENP8	
ATP2A2	DPP3	LIPT2	PDZD3	SERGEF	
ATP2B1	DPY19L2	LMF1	PDZRN4	SERPING1	
ATP5B	DUSP16	LRR1	PFKM	SERPINH1	
ATP6V0A2	DUSP18	LRRK2	PGA3	SETD1B	
B3GAT3	DUT	LRTOMT	PGA4	SETD8	
B3GNT6	DYNC2H1	LTA4H	PGA5	SETX	
B4GALNT1	DYRK2	LYZ	PGM2L1	SHMT2	
BACE1	EIF2B1	MACROD1	PGM5	SHPK	
BCAS3	ELMOD1	MADD	PICALM	SIAE	
BCAT1	ENDOG	MAGI3	PIK3C2A	SIK2	
BCDIN3D	ENDOU	MAP3K11	PIK3C2G	SIK3	
BIRC2	EPHB1	MAP3K12	PIP4K2C	SIRT4	
BIRC3	ERBB3	MAP4K2	PLA2G1B	SLC27A6	
BLK	ERP27	MAPK8IP1	PLCZ1	SLC37A4	
BRAP	ERP29	MAPKAPK5	PLEKHG7	SLC3A2	
BRMS1	ESPL1	MARK2	POC1B-GALNT4	SLN	

Table 1 (Continued).

C11orf54	ETNK1	ME3	POLR2G	SNX15
C7orf60	EVI5	MED21	POLR3B	SOAT2
C9orf156	F2	METAP2	PPM1H	SOCS2
CABP1	FADD	METTL1	PPME1	SORL1
CAMK2G	FADS1	METTL12	PPP1CA	SPCS2
CAMKK2	FADS2P1	METTL15	PPP1CC	SPINT2
CAMKV	FADS3	METTL20	PPP1R12A	SPPL3
CAND1	FAM76B	METTL25	PPP1R14B	SPTSSB
CAPN1	FAM86C2P	METTL7A	PPP1R1A	SRGAP1
CAPN13	FBXO21	METTL7B	PPP2R5B	SRPK2
CAPN5	FBXO3	MGAT4C	PPP6R3	SSH1
CARD16	FBXW8	MICAL2	PPTC7	SSH3
CARD17	FDXACB1	MMAB	PPWD1	SSU72
CARNS1	FEN1	MMP10	PRCP	ST14
CASP1	FGD4	MMP13	PRDM11	ST3GAL4
CASP12	FGD6	MMP19	PRIM1	ST8SIA1
CASP4	FICD	MMP3	PRKAB1	STK32B
CASP5	FKBP11	MMP8	PRMT3	STK38L

adaptations in recent human evolution, along with new genomic research, show a way forward.

The few examples of adaptation that have been described so far, such as malaria-resistance (Currat et al., 2002), lactase-persistence (Tishkoff et al., 2007b), skin melanation (Izagirre, García, Junquera, de la Rúa, & Alonso, 2006), hypoxia-response (Cynthia M Beall et al., 2010a), heat production (Hancock, Clark, Qian, & Di Rienzo, 2010a), and sweat-gland activity (Kamberov et al., 2013), motivate future research. They reflect human migrations and culture, and they impact human health. In particular, they highlight metabolism as sensitively tuned by environmental and dietary selective factors, and pregnancy and infancy as life stages that often mediate the actions of selection. Such insights can help point researchers to more likely candidate genes and phenotypes for further study.

In addition, new genomic research allows researchers to generate *de novo* hypotheses of adaptation based upon particular functional elements under selection in individual human populations. Researchers can do this by combining the data from more sequenced genomes from diverse populations (Abecasis et al., 2012) with improved power in statistical tests for selection (Grossman et al., 2010) and genomic functional annotation from high-throughput assays (Consortium et al., 2015; Dunham et al., 2012). These recent developments in genomics, along with insights garnered from in depth studies of genes under selection, render the barriers to working out new examples of adaptation more tractable.

New Genomic Tools for Studying Selection

One major new genomic dataset that aids in studying selection in diverse human populations is the 1000 genomes project resequencing data for more than 2,500 individuals from 26 populations around the world in its phase 3 release. Previous analysis relied on far fewer populations (*e.g.*, the phase 1 release of 1000 genomes consisted of just Yoruba, Han Chinese, Japanese, and Northern Europeans), genotyping data rather than full sequence data (*e.g.*, HapMap), or far fewer numbers of individuals in each population (*e.g.*, the Human Genome Diversity Panel). The advantages of this dataset are the possibility of detecting adaptations in more diverse human populations, and the more complete and accurate picture of genetic variation in these populations.

Genomic functional annotation from high-throughput assays also aids in the detection of selection in the genome because it provides a map of functional regions and variants upon which selection may act. Functional variation comes in two major varieties: non-synonymous coding variants that impact protein form and function and regulatory variants that impact gene expression levels. A recent scan for selection in human populations found only 35 non-synonymous coding variants with high-probability for being under selection in 412 regions identified as being under strong, positive selection (Grossman et al., 2013). In addition, another recent analysis of genetic variation showing shifts in frequency correlating to climate found a ten-fold enrichment of regulatory variants among these candidate adaptive loci compared to non-synonymous variants (Fraser, 2013). These analyses indicate that regulatory variation may be behind the majority of human adaptations.

Two broad categories of genomic analysis aid detection of the regulatory variants that could have been the substrates for selection. The first type of analysis correlates gene expression data across individuals from RNA-seq studies with genetic variants in those individuals. The genetic variants that most strongly associate with the expression of nearby genes are termed eQTLs, or expression quantitative trait loci, and are candidate loci for regulating expression of those genes. The second type of analysis detects chemical signatures (*e.g.*, histone methylation and acetylation, DNase I, chromatin-immunoprecipitation [ChIP]-seq for transcription factor binding) in regions of the genome that typically correspond to regulatory function. Finding eQTLs within regions identified as regulatory by these genomic functional assays lends stronger evidence for genuine regulatory variation. Larger RNA-seq studies on denser maps of genetic variation from the 1000 genomes project (Lappalainen et al., 2013) combined with large datasets of genomic functional annotation from the ENCODE project (Dunham et al., 2012) and the Epigenomics Roadmaps project (Consortium et al., 2015) improve the power to detect regulatory variation in each of these arenas.

Another tool that helps pinpoint new adaptations in the human genome is a refined statistical technique for detecting selection in the genome called the Composite of Multiple Signals test, or CMS (Grossman et al., 2010). Tests of selection are designed around the fact that positive selection creates disturbances in the expected allele frequency spectra of selected loci and surrounding genetic variation. These disturbances, or signatures of selection, can be measured by comparing allele frequency spectra of particular loci to a genome-wide null distribution. Different test statistics specialize in detecting different signatures of selection. For example, allele frequency differentiation of a locus selected only

in a subset human populations is predicted to be high, and can be measured using the fixation index (Fst) (Wright, 1950a). In addition, haplotype homozygosity surrounding positively selected newly derived variation is expected to be extensive when recombination has not yet broken the linkage disequilibrium between the selected locus and surrounding genetic variation. Tests such as the integrated haplotype score (*iHS*) and cross-population extended haplotype homozygosity (*XP-EHH*) measure length of unbroken haplotypes (Sabeti et al., 2007; Voight, Kudaravalli, Wen, & Pritchard, 2006). Lastly, positive selected will increase the frequency not only of the selected allele, but also of nearby linked variation, causing an excess of high-frequency derived alleles in the region, which can be measured using the change in derived allele frequency (ΔDAF) (Grossman et al., 2010). CMS leverages the signatures of selection inherent in each of these tests to pinpoint individual regions and variants within those regions with highest probability for being under selection (Figure 1). CMS greatly enhances the power to localize selection in the genome to a tractable number of variants for functional analysis, enabling researchers to generate novel adaptive hypotheses.

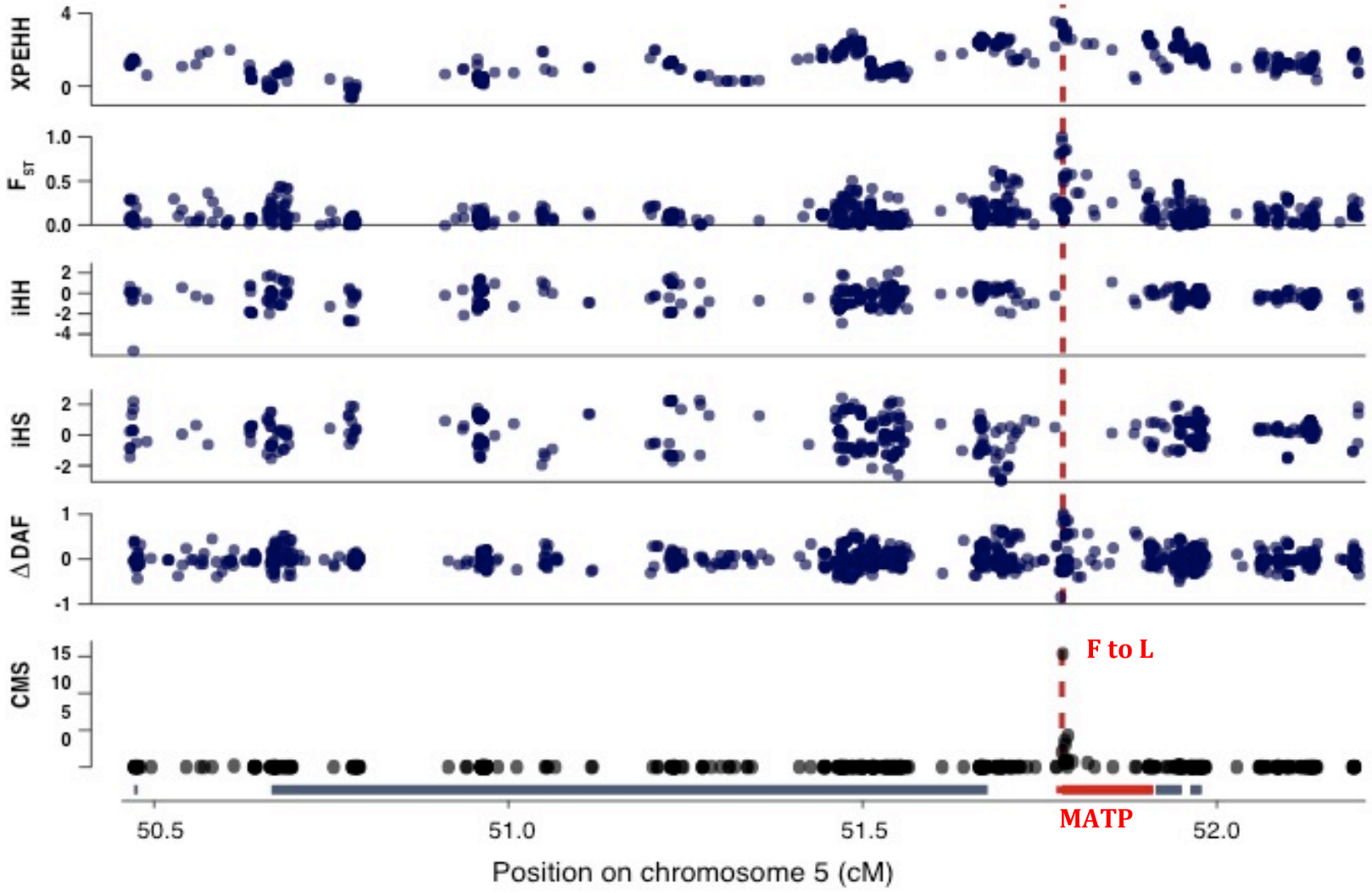
Outline of Dissertation

This dissertation reviews human metabolic adaptations in diverse populations (chapter 2), considers how recent human adaptations fit into human life history and physiology (chapter 3), and describes novel research into an adaptation in Asian populations (chapter 4). The second chapter begins by putting recent human metabolic adaptations into the broader context of human metabolic evolution since the lineage split between *Homo* and *Pan*. Then, it outlines the complex genetic underpinnings of metabolic

Figure 1. CMS identifies an adaptive variant in MATP for skin color
The Composite of Multiple Signals (CMS) test for selection combines five statistical tests for selection (*XP-EHH*, F_{ST}, ΔiHH, *iHS*, and ΔDAF) described in the text to localize the signal of selection to individuals variants with high probability for being under selection. This example shows the localization of a signal of selection on chromosome 5 in European haplotypes to a non-synonymous coding variant in the *MATP* gene, the membrane-associated transporter protein, which is involved in skin color adaptation in Europe.

Figure 1 (Continued)

CMS identifies an adaptive variant in MATP for skin color



traits alongside the genetic patterns indicative of selection in human populations. Finally, it reviews the genetic and phenotypic evidence for adaptation for heat production and high-altitude hypoxia response in human populations, as well as a few candidate genes and phenotypes for adaptation, to expose the challenges and limitations researchers face in studying metabolic adaptations.

The third chapter approaches the problem of recent selection in humans from the perspective of the life history stages and physiological traits in humans that are most sensitive to the actions of selection. The chapter begins by laying out the importance of reproduction—pregnancy and infancy—to Darwinian fitness and the specific constraints that make that make these stages challenging. This provides grounding for why the immune system and metabolism experience so much selection in diverse human populations. Furthermore, the chapter examines variable human metabolic phenotypes, such as blood glucose levels and salt sensitivity, in the context of pregnancy to generate novel hypotheses of why they vary across people. The aim of this chapter is to provide researchers with a stronger theoretical framework for generating and evaluating new hypotheses of adaptation.

The fourth chapter describes primary research investigating a metabolic gene under selection in East Asians from a functional, computational, and physiological perspective. New genomic tools, described above, are used to hone in on candidate genetic variants for study. Then, functional assays in cell culture are used to test the regulatory impact of variation. Next, surrounding patterns of genetic variation are explored in terms of population differentiation and selection. Finally, knowledge of gene function and human diet are used to generate a novel hypothesis of adaptation in the East Asian population.

This chapter provides a model for how selection can be explored functionally and computationally in a hypothesis-generating fashion.

CHAPTER 2

GENETIC EXPLORATIONS OF RECENT HUMAN METABOLIC ADAPTATIONS:

HYPOTHESES AND EVIDENCE

ABSTRACT

Since humans and chimpanzees split from a common ancestor over 6 million years ago, human metabolism has changed dramatically. This change includes adaptations to a high quality diet, the evolution of an energetically expensive brain, dramatic increases in endurance abilities and capacity for energy storage in white adipose tissue. Human metabolism continues to evolve in modern human populations in response to local environmental and cultural selective forces. Understanding the nature of these selective forces and the physiological response during human evolution is a compelling challenge for evolutionary biologists. The complex genetic architecture surrounding metabolic phenotypes indicates that selection probably altered allelic frequencies across many loci in populations experiencing adaptive metabolic change to fit environment (Hernandez et al., 2011; J. K. Pritchard, Pickrell, & Coop, 2010). A recent analysis supports this hypothesis, finding that classic selective sweeps at single loci were rare in the past 250,000 years of human evolution (Hernandez et al., 2011). Detection of selective signatures at multiple loci, as well as exploration of physiological adaptation to environment in humans, will require cross-disciplinary collaboration, including the incorporation of biological pathway analysis to inform this evolutionary research. This paper explores the Thrifty Genotype Hypothesis (Neel, 1962), high altitude adaptation, and cold resistance adaptation and genetic evidence

surrounding these proposed metabolic adaptations in hopes of clarifying current challenges and avenues for future advancement of understanding.

INTRODUCTION

Changes in Metabolism during Human Evolution

Since humans and chimpanzees diverged approximately 6 million years ago (Goodman, 1999), humans have become adapted to a high quality diet. Humans exhibit much smaller digestive tracts than chimpanzees and other apes and much reduced prognathism and tooth size than fossil hominins living before 1.9 million years ago (Rachel N. Carmody & Wrangham, 2009; Lieberman, Pilbeam, & Wrangham, 2009). The nature and timing of the shift to a high quality diet remain contentious, though some components are known. Modern humans eat meat and animal products in much higher quantities than wild chimpanzees and other modern apes (Cordain et al., 2000). Meat exhibits vastly different protein and fat profiles than do plant foods. Furthermore, cooking and extensive manual and chemical processing of food are unique human behaviors that increase the caloric availability of food (Rachel N Carmody, Weintraub, & Wrangham, 2011; Rachel N. Carmody & Wrangham, 2009; Wrangham & Conklin-Brittain, 2003). Tubers and meat have both been hypothesized to be important substrates for cooking in the hominin diet, though the timing of when hominins became habituated to cooking is highly debated (Rachel N Carmody et al., 2011; Rachel N. Carmody & Wrangham, 2009; Wrangham & Conklin-Brittain, 2003).

Human energetic demands and metabolism also changed in non-dietary ways. Chimpanzees and humans are largely dichotomous in their athletic phenotypes. Chimpanzees are stronger and more muscular, while humans have made unique gains in

endurance ability compared to other primates (Lieberman et al., 2009). Humans have also evolved an energetically expensive brain that demands protection from the interruption of energy inflow during the infant years as it continues to grow (Kuzawa, 1998). To protect the brain from interruption of energy inflow, and possibly other reasons, humans have unusually large fat stores by percent body weight in the form of white adipose tissue during infancy and continuing into adulthood (Kuzawa, 1998; Wells, 2006).

Other human metabolic adaptations arose more recently in response to the unique environments that individual populations encountered when they moved outside of Africa during the past 100,000 years (Forster & Matsumura, 2005; Macaulay et al., 2005). Cold temperatures and high altitude hypoxic conditions are two environmental factors to which human populations may have metabolically adapted. Variation in natural food sources, such as marine versus terrestrial environments, might have further altered human metabolism in specific populations. Finally, cultural and behavioral shifts have acted as selective agents altering human metabolism in the populations that adopted them. Agriculture and pastoralism have altered human energy metabolism by changing the macromolecular content of the human diet, increasing the seasonality of energetic input and output, and possibly changing the pattern of energetic exertion. Throughout this paper, the term "environment" is used in a comprehensive sense to refer to all external stimuli that interact with the organism including climatic, dietary, cultural, and behavioral components.

Broad shifts in energy metabolism during the span of human evolution, as well as more recent variation in energy metabolism exhibited by different populations, have

exerted fitness effects on humans, selecting adaptive genetic variants in novel climatic, dietary, and behavioral environments.

Hypotheses

Some variable phenotypes among human populations are hypothesized to be adaptations to particular environmental conditions. For example, humans conform to Bergmann's and Allen's Rules: people from northern latitudes have broader frames with shorter distal to proximal limb ratios than equatorial people (Katzmarzyk & Leonard, 1998; D. F. Roberts, 1953). This has been proposed to be an adaptation to ambient climatic temperature as more spherical body forms retain heat and more elongated body forms dissipate heat with higher efficiencies. Furthermore, human adaptation to variable UV exposure at different latitudes via variable melanin production has been extensively described (Nina G. Jablonski & Chaplin, 2000).

Several other hypotheses regarding adaptation to local environments in humans have been proposed and have accrued varying degrees of support from genetic and physiological investigation. Adaptations to high altitude, cold resistance, and the digestion of milk by pastoral adults have the strongest supporting evidence, while adaptations to water loss or salt-sensitivity, folate deficiency, iodine deficiency, detoxification of xenobiotic compounds found in plants, and heightened starch consumption have also been proposed for various human populations (X. Chen et al., 2009; Hancock, Clark, Qian, & Di Rienzo, 2010b; López Herráez et al., 2009; Luca et al., 2008; George H. Perry et al., 2007; Storz, 2010; Thompson et al., 2004; Tishkoff et al., 2007a). Evidence has also accumulated for selection upon genetic variants linked to particular metabolic phenotypes due to change in energy metabolism among human groups. One famous attempt was made to resolve the

agent behind potential selective events upon energy metabolism. The proposal, known as the Thrifty Genotype Hypothesis, attempted to explain variable prevalence of diabetes using a "quick insulin trigger" in the context of energy availability, the transition to agriculture, and energy storage (Neel, 1962). Neel later expanded this hypothesis to encompass more specifically type II diabetes, obesity, and hypertension (aspects of the metabolic syndrome) (Neel, 1999). Essentially, the Thrifty Genotype Hypothesis proposes that certain human populations have been selected for an enhanced ability to store carbohydrates as fat in order to maximally exploit resources to survive long periods of famine in highly seasonal environments. For example, as agriculture has led to heightened seasonality of environments, agricultural populations are expected to sequester calories as fat more efficiently, resulting in the modern epidemic of obesity and type II diabetes in the current environment of ubiquitously available highly caloric food sources. Originally proposed in 1962, this hypothesis has been modified since, and evidence for it is mixed (see Pollard (2008) for a review of recent research).

However, few researchers since have attempted to tackle the thorny problem of deciphering what precise selective agent may have caused changes in the metabolic phenotypes of certain populations. Prevalence of metabolic phenotypes such as type II diabetes, obesity, resting metabolic rate, hypertension, and triglyceride levels are reported to vary in populations. For example, African Americans are reported to have higher prevalence of hypertension and obesity; certain Native Americans and Oceanic groups are reported to have higher prevalence of obesity and type II diabetes; and Asians and African Americans at lower BMIs tend to have higher prevalence of type II diabetes than European Americans (Chiu, Austin, Manuel, Shah, & Tu, 2011; Diamond, 2003; McKeigue, Shah, &

Marmot, 1991; Miller, 2010; Seidell, 2000; Youfa Wang & Beydoun, 2007). This variation may be due to environmental, epigenetic, or genetic variation (Cheng et al., 2010; Fernández et al., 2003; H. Tang et al., 2006; R. C. Williams, Long, Hanson, Sievers, & Knowler, 2000). However, though environmental mismatch and epigenetic Thrifty Phenotype hypotheses have been proposed as alternative mechanisms behind variable metabolic patterns in diverse populations, few if any modern alternatives to the Thrifty Genotype Hypothesis exist that purport to explain natural selection upon metabolic phenotypes in response to shifts in energy metabolism (though see (Corbett, McMichael, & Prentice, 2008)).

Genetic Architecture of Variable Phenotypes and Metabolic Disorders

The genetic architecture underlying phenotypes that vary among humans has been challenging to uncover. Genome-wide association studies (GWAS) have revealed that many variants of small effect size may underlie human variation in metabolic phenotypes. For example, the latest GWAS on lipid profiles revealed 95 genetic variants that affect lipid profiles in Europeans (Teslovich et al., 2010). In total these 95 genetic variants explain roughly 12% of the variance for each lipid profile, or 25-30% of the heritable variance of the traits (Teslovich et al., 2010). As sample sizes and marker densities increase, the number of identified variants also is expected to increase. Even so, most heritable genetic components of metabolic phenotypes still go unexplained by the identified loci. One possible explanation is that the identified loci are not the causal alleles and are just in linkage disequilibrium with the causal alleles, losing some of the recognizable heritable effect size of the variant (Manolio et al., 2009). However, this uncertainty allows for the possibility that still many more variants of even smaller effect size may exist. The genetic

architecture underlying metabolic phenotypes in humans seems to be complex with a few variants of large effect size and many variants of small effect size (De Silva & Frayling, 2010; Speliotes et al., 2010). Additionally, diverse under-sampled populations contain variants that may be population-specific risk factors (De La Vega, Bustamante, & Leal, 2011). These variants may be common in the under-sampled population, but rare and thus undetected in other world samples. As the vast majority of GWAS are conducted on Europeans, the genetic assessment of phenotypic risk may be subject to population bias (De La Vega et al., 2011). Furthermore, sex-gene interactions, gene-gene interactions, and gene-environment interactions also complicate the architecture of human metabolic phenotypes. This complexity affects both the way in which phenotypic selection plays out in the genome and, also, the way in which selection on genetic variants affecting metabolic phenotypes may be detected.

Statistical Tests for Signatures of Natural Selection

In order to detect evidence of natural selection having acted on particular genetic variants in a population, several statistical tests are employed that identify signatures of these selective events. Hard sweeps occur when a newly derived allele confers a beneficial phenotype in a population. The fitness-enhancing allele rises quickly in frequency in the population, skewing the pattern of genetic variation surrounding the allele (Smith & Haigh, 1974). Patterns left by hard sweeps include high allelic frequency differentiation between populations (high Fst or SNP frequency change) (Wright, 1950b), extended regions of haplotype homozygosity surrounding the selected allele due to the rapid increase in allele frequency outpacing the local recombination rate (high *iHS* or *XP-EHH*) (Sabeti et al., 2007; Voight, Kudaravalli, Wen, & Pritchard, 2006), and an excess of low frequency derived

alleles in the population at the site of selection, as many older alleles will have been driven out by the positively selected allele (Tajima, 1989). This excess of low frequency derived alleles can alternatively be an indication of purifying selection to conserve a region against deleterious mutations (Kirk E. Lohmueller et al., 2011). The Composite of Multiple Signals (CMS) test exploits all of the patterns of genetic variation left by hard sweeps (Grossman et al., 2010).

Soft sweeps occur when multiple standing variants confer phenotypic effects in a population, and these phenotypes are positively selected. Under this model, the fitness enhancing alleles rise subtly in frequency in the population. Unlike traditional hard sweeps in which one fitness enhancing allele is strongly selected, soft sweeps do not drive fast fixation of a given allele and have very little effect on the allele frequency spectrum and haplotype structure surrounding selected alleles (Hermisson & Pennings, 2005; Przeworski, Coop, & Wall, 2005). Soft sweeps lead to polygenic adaptation in the population. (J. K. Pritchard et al., 2010)

The Bayesian Likelihood Test (BLT) is a newly developed method that should have the power to detect soft sweeps occurring in response to a given environmental selective agent. This method relies upon using large numbers of populations spread worldwide to test whether allele frequencies correlate with an environmental variable, such as seasonal temperature, independent of demographic relatedness of populations. Dense sampling of world populations and precise measures of allele frequencies in those populations are essential to achieve statistical power to detect alleles involved in soft sweeps using the Bayesian Likelihood Test. Previous studies have included 600,000 genome-wide SNPs from

52 world populations in the HGDP. (Coop, Witonsky, Di Rienzo, & Pritchard, 2010; Hancock, Alkorta-Aranburu, Witonsky, & Di Rienzo, 2010)

Demographic processes (population expansion, subdivision, bottlenecks, and shared histories) may create genetic patterns that mimic those left by selection in both the hard and soft sweep models. For example, population expansion generates an excess of low frequency alleles on haplotypes, mimicking hard sweeps. The key distinguishing factor is that demographic processes affect the entire genome similarly as most of it is evolving neutrally, while selection targets particular loci (Kelley, Madeoy, Calhoun, Swanson, & Akey, 2006). Therefore, researchers wishing to control for the effects of demography must use null distributions of test statistics generated from genomic datasets to compare to the signal of selection for the region of interest (see Akey et al. (2004) for an example). Now many studies, such as Bigham et al. (2010a), generate their own null distributions of test statistics, which gives a more precise comparison since the data has been collected from the same populations using the same methods. Interestingly, widespread negative selection affects the allele frequency spectrum of linked neutral variation, indicating that even neutral variants may not be perfect markers of the demographic processes that researchers need to control for to detect positive selection (Kirk E. Lohmueller et al., 2011). This presents a novel challenge for creating null distributions of test statistics as controls for signatures of selection.

Unfortunately, another confounding factor known as ascertainment bias can also mimic patterns of allelic diversity left by selective sweeps when genotyping rather than full-sequencing data is used (Kelley, Turkheimer, Haney, & Swanson, 2009). Using multiple test statistics that focus on different aspects of the data to detect selection can help remove

false positives from individual tests. Controlling for confounding demographic processes and detecting selection on multiple standing variants across the genome are the major challenges for detecting recent selection in human populations.

Evidence from Genomics

Hard Sweeps

The emerging field of human genomics brings some evidence to bear on recent human metabolic adaptations. Many genomics studies assessing recent positive natural selection across the human genome have scanned dense genotype maps in diverse human populations, typically specializing in detecting hard sweeps. The lists of regions produced by these scans may be assessed for the types of genes contained within them using standard gene classifications, such as Gene Ontology (GO) or Protein ANalysis THrough Evolutionary Relationships (PANTHER) categories. Several of the lists of regions under selection reveal enrichment for genes involved in metabolic processes.

Genes of metabolic function show some of the strongest signatures of positive natural selection in the human genome (Akey, 2009; Nielsen, Hellmann, Hubisz, Bustamante, & Clark, 2007; Nielsen et al., 2005; Voight et al., 2006). Notably, genomic regions of selection within human populations are enriched for genes involved in protein modification and metabolism, carbohydrate metabolism, and phosphate metabolism (Akey, 2009). An analysis by Voight et al. (2006) of genetic data from the HapMap database, including Yorubans, Europeans, and East Asians, found enrichment for regions containing genes involved in the metabolism of carbohydrates, lipids, and phosphates among East Asians, and in vitamin transport among East Asians and Europeans. In addition, an analysis

by Wang et al. (2006) using Perlegen and HapMap data revealed enrichment for genes involved in protein metabolism. López Herráez et al. (2009) using 51 populations from the Human Genome Diversity Panel (HGDP) found enrichment for genes in GO categories related to "how humans interact with their environments, especially in terms of pathogens and diet…," specifically in the category "metabolism of xenobiotic compounds." Using the CMS test for recent positive selection on the HapMap populations, Grossman et al. (2010) detected several interesting metabolic and diet-related PANTHER categories enriched for genes in regions under selection. For example, calcium mediated signaling in Europeans, protein metabolism and modification in Europeans and Yorubans, and homeostasis in East Asians all show enrichment for being under selection.

In contrast, a recent analysis by Hernandez et al. (2011) of resequencing data from 179 human genomes indicates that classic hard sweeps occurring over the past 250,000 years are rare in human evolution. In addition, another recent analysis of resequencing data in human genomes by Lohmueller et al. (2011) concluded that negative or purifying selection against deleterious variants may be responsible for a large portion of the regions exhibiting reduced haplotype diversity and skewed allele frequency spectrum detected by the genomic scans, rather than positive selection on a beneficial allele. An important limitation of the analyses described above is that scans for hard sweeps in the genome implicitly assume that strong selection has occurred on single newly derived variants with strong phenotypic effect sizes. This constitutes a built-in ascertainment bias to genomic scan results. When selection happens on long-standing rather than newly derived variants, the allele frequency spectrum and haplotype structure in the population will lack the obvious signature of a hard sweep. Furthermore, much selection may be missed because

metabolic phenotypes have complex genetic architectures that consist of many variants of small effect sizes. The examples of pathway or category enrichment above require that signatures of selection be very strong surrounding each variant. Many variants underlying metabolic phenotypes have only small effect sizes and may be individually selected more weakly, with only the pathway as a whole being strongly selected, resulting in the signature of a soft sweep rather than a hard sweep.

Soft Sweeps and Polygenic Adaptation

In assessing small changes in allele frequency that correlate with climate or subsistence patterns, Hancock et al. (2010) detected putative signatures of soft sweeps in response to certain environmental pressures significantly above the neutral signal of demographic relatedness. These small increases in allele frequency across many loci result from positive selection of variants that confer increased fitness in the novel environment (Angela M. Hancock et al., 2010; J. K. Pritchard et al., 2010). Under this model, small changes in allele frequencies render members of the population better fit to their environment via polygenic adaptation (Angela M. Hancock et al., 2010; J. K. Pritchard et al., 2010).

Studying genes involved in the same pathway is one tool that has been used to detect polymorphisms that have been subject to soft sweeps and may constitute polygenic adaptation. For example, SNPs in genes involved in starch and sucrose metabolism and folate biosynthesis are enriched for frequency changes correlated with the consumption of roots and tubers as the primary dietary component (Angela M. Hancock et al., 2010). This is significant as roots and tubers are primarily composed of starch. Adaptation to starch via *AMY1* copy number variation has also been proposed for populations known to consume

more starch (George H. Perry et al., 2007). Two genes involved in the catabolism and anabolism of glycogen (*GAA* and *GBE1*) are among those with SNPs with the strongest correlation to roots and tubers dietary component in the Hancock et al. (2010) study(Hancock et al., 2010)(Hancock et al., 2010)(Hancock et al., 2010)(Hancock et al., 2010)(Hancock et al., 2010). Furthermore, roots and tubers are low in folates. Since folic acid is essential for neural development in infants and neural maintenance throughout life, enrichment for genes involved in the biosynthesis of folates seems adaptive for populations consuming folate-poor roots and tubers as the staple of their diets (Angela M. Hancock et al., 2010). Previously, selection of variants of *NAT2* conferring the slow-acetylator phenotype was proposed as an adaptation to the folate-poor diets of agricultural populations, as slow-acetylation by *NAT2* may aid in folate retention (Luca et al., 2008). Also, SNPs in genes involved in pyruvate metabolism and glycolysis and gluconeogenesis, pathways intimately connected with energy metabolism, are enriched for having changes in frequency correlated with polar ecoregions (Angela M. Hancock et al., 2010).

Finally, some genes with variants known to impact metabolic phenotypes such as lipid levels, fasting plasma glucose, and QT interval (the interval between the Q wave and the T wave in the heart's electrical cycle—long QT interval is commonly a result of obesity) were identified as having SNPs strongly correlated with certain environmental and subsistence variables such as humid tropical ecoregion, foraging, pastoralism, and roots and tubers or cereals as primary dietary components (Angela M. Hancock et al., 2010). One suggestive correlation was the increase in frequency of a SNP in *KCNQ1* (encoding a cardiac potassium channel), which is associated with increased risk of Type II diabetes, in populations who primarily consume cereals (Angela M. Hancock et al., 2010). More

examples of SNPs with associations to metabolic disorders with strong signals of correlation to subsistence, ecoregion, and climate were later published (Angela M Hancock, Gorka Alkorta-Aranburu, et al., 2010), and more analyses using denser sampling of world populations continue to be performed (Hancock et al., 2011b).

Case Studies of Selection on Metabolic Genes

A few instances of selection on genes underlying metabolic phenotypes in humans from different populations have been described. In addition to the genomic scans described above, directed investigations into candidate genes and phenotypes have the potential to reveal specific metabolic adaptations in which signatures of selected genetic regions can be linked to fitness-enhancing phenotypes in a population. Examining specific case studies can also provide insight into the importance of different models of selection. The degree and nature of evidence surrounding these examples varies on a case-by-case basis.

(*a*) *Cold Resistance Adaptation*

As described above, humans from different latitudes conform to Bergmann's and Allen's Rules, which serve either to conserve or dissipate body heat more effectively by altering the volume to surface area ratios of the body. Non-shivering thermogenesis (the generation of body heat by means other than mechanical shivering) (Seale, Kajimura, & Spiegelman, 2009) is another physiological mechanism that may have been subject to selection in response to cold stress in humans living in low temperature regions. The main site of non-shivering thermogenesis is brown adipose tissue, which burns fat to produce heat via the uncoupling of mitochondrial oxidative metabolism from ATP production (Seale et al., 2009). Brown-adipose-tissue has long been known to exist in rodents and mammalian infants (Hughes, Jastroch, Stoneking, & Klingenspor, 2009; Seale et al., 2009).

More recently, brown adipose tissue was detected in distinct deposits in adult humans, as well as interspersed in deposits of white adipose tissue (Seale et al., 2009). Because resting respiration rates and metabolic rates are known to increase from Africans to Europeans to populations living in extreme northern latitudes (Snodgrass, Sorensen, Tarskaia, & Leonard, 2007; Wong et al., 1999), mitochondrial oxidation, dependent upon respiratory and metabolic rates, in the non-shivering thermogenesis pathway in brown adipose tissue may have been subject to recent selection in response to cold.

A recent study investigating small changes in allele frequency among 52 worldwide populations detected a correlation between allele frequencies of SNPs in many metabolic genes and the polar ecoregion, using the Bayesian Likelihood Test (Angela M. Hancock et al., 2010). In particular, strong signals were found for mitochondrial malic enzymes 2 and 3 (*ME2* and *ME3*), involved in oxidative metabolism, suggesting a link between these genetic variants and cold adaptation (Angela M. Hancock et al., 2010).

These findings parallel earlier findings from a similar study, which found a strong correlation between the (rs1137100) 109R allele of *LEPR*, encoding the leptin receptor, and the winter *PC1* variable (i.e., the first principle component related to winter conditions—a combination of several measures of winter severity for the home locales of the populations studied) (Hancock et al., 2008). The 109R allele of *LEPR* is associated with an increased respiratory rate and decreased BMI, which fits the role of mitochondrial oxidative metabolism of fatty acids for thermogenesis in brown adipose tissue (Hancock et al., 2008). Signals of high *iHS* surround the *LEPR* 109R allele in Asian and European populations in the HapMap dataset, indicative of recent positive selection (Hancock et al., 2008). Interestingly, leptin serves as a satiety signal: leptin-deficient rodents and humans

have decreased respiratory rate, and increased food intake and body weight. (See (Rosenbaum & Leibel, 1999) and (Friedman & Halaas, 1998) for reviews of leptin function.) Given this information, the 109R allele of the leptin receptor could serve as a gain-of-function for leptin signaling to help mobilize energy stores in a cold environment.

Correlation was also found between the 192Q allele of *PON1* (rs662) and winter *PC1* (Hancock et al., 2008). The *PON1* protein product is known to protect lipids from oxidation, and the 192Q variant lowers protein activity *in vitro* which could increase levels of lipid oxidation (Hancock et al., 2008). Again, this is consistent with an enhanced ability by 192Q-carriers to mobilize energy stores in a cold environment. These examples are suggestive of genetic adaptation for cold tolerance among people living in colder northern climates via selection on the pathway surrounding non-shivering thermogenesis. (See Table 2 for a summary)

In addition, transcription factors and genes responsible for generation of brown adipose tissue from skeletal muscle progenitor cells have begun to be described. Key among these is *UCP1*, known to be critical for the uncoupling of mitochondrial oxidation from ATP production in brown adipose tissue (Hughes et al., 2009; Kajimura et al., 2009). A team of researchers recently tested variants in *UCP* genes for evidence that they have evolved adaptively for cold resistance (Angela M Hancock, Vanessa J Clark, et al., 2010b). Variants that increase expression of *UCP1* (such as the A allele of the -3826 A/G polymorphic site—rs1800592) increase metabolic rate and thermogenic capacity and decrease rate of weight gain (Angela M Hancock, Vanessa J Clark, et al., 2010b). The functions of the homologues *UCP2* and *UCP3* are less well defined. However, based upon rodent models and expression studies, a role for either of them in uncoupling and

Table 2. Genetic cold adaptations

Genetic Regions	Functional Connection	Evidence	Source
ME2 and *ME3*	These mitochondrial malic enzymes 2 and 3 are involved in oxidative metabolism.	SNP frequencies in *ME2* and *ME3* correlate with polar ecoregion across 52 world populations, controlling for neutral demographic processes.	Hancock et al. (2010c)
The 109R allele of *LEPR* (rs1137100)	The 109R allele of *LEPR* is associated with an increased respiratory rate and decreased BMI, consistent with a role in non-shivering thermogenesis adaptation.	109R of *LEPR* correlates with the winter *PC1* variable across 54 world populations. 109R was identified by network analysis of 873 tag SNPs in 82 candidate genes for common metabolic disorders. High *iHS* surrounds the 109R allele in East Asian and European HapMap populations.	Hancock et al. (2008)
The 192Q allele of *PON1* (rs662)	The PON1 protein product is known to protect lipids from oxidation. Since the 192Q variant lowers protein activity *in vitro*, 192Q could increase levels of lipid oxidation by mitochondria in brown adipose tissue.	192Q of *PON1* correlates with winter *PC1* across 54 world populations. 192Q was identified using 873 tag SNPs in 82 candidate genes for common metabolic disorders.	Hancock et al. (2008)
The -3826 A allele of *UCP1* (rs1800592)	*UCP1* expressed in brown adipose tissue is critical for the uncoupling of mitochondrial fatty acid oxidation from ATP production during thermogenesis. The -3826 A allele is associated with increased expression of *UCP1*, decreased rate of weight gain, and increased thermogenic capacity, consistent with a role in non-shivering thermogenesis adaptation.	The -3826 A allele of *UCP1* correlates with latitude across 52 world populations, controlling for neutral demographic processes. High *iHS* surrounds -3826A on some Italian haplotypes, indicating this allele may have been subject to recent positive selection as a standing variant in the population.	Hancock et al. (2010b)
The -55 T allele of UCP3 (rs1800849)	*UCP3* shares homology with *UCP1* and may share function in the decoupling of mitochondrial oxidation from ATP production in thermogenesis. The -55 T allele is associated with increased expressed of UCP3, decreased risk of obesity, and increased metabolic rate, consistent with a role in non-shivering thermogenesis adaptation.	The -55 T allele of *UCP3* correlates with minimum winter temperature across 52 world populations, controlling for neutral demographic processes. The Han Chinese population exhibits an excess of alleles at intermediate frequency in the *UCP3* gene region, indicative of either balancing or positive selection on multiple standing variants.	Hancock et al. (2010b)

thermogenesis in brown adipose tissue seems possible (Angela M Hancock, Vanessa J Clark, et al., 2010b).

Though analyses of allelic spectrum frequency and haplotype length failed to detect conclusive evidence of positive selection on haplotypes containing alleles known to increase expression of *UCP* proteins, the -3826 A allele of *UCP1* (rs1800592) demonstrated frequency correlation with latitude and the -55 T allele of *UCP3* (rs1800849) demonstrated frequency correlation with minimum winter temperature across 52 global populations, along with several other variants in the *UCP3* gene region. Both of these variants are associated with increased protein expression, decreased risk of obesity, and increased resting metabolic rate. Furthermore, some haplotypes containing the *UCP1* -3826 A allele in Italian populations do exhibit decreased allelic diversity, suggesting that selection may have acted on this allele as a standing variant in Northern populations, causing some allele-containing-haplotypes to exhibit signatures of selection, while others do not. The Han Chinese population also exhibits an excess of alleles at intermediate frequency in the *UCP3* gene region, which may be a signature of balancing selection (Tajima, 1989). Although some modern Italians and Han Chinese live in environments considered warm relative to more extreme cold regions, such populations may still have endured temperatures during their evolutionary history sufficient to select for cold resistance as humans are a tropical species with very limited fluctuation in body temperature tolerated for optimal function. These genetic patterns are consistent with positive selection acting on multiple alleles in the *UCP1* and *UCP3* gene regions as adaptations to cold resistance. (Angela M Hancock, Vanessa J Clark, et al., 2010b)

The evidence gathered from this study of *UCP* genes in relation to cold resistance, in addition to the evidence gathered from the other genes described (see Table 2), is consistent with the complex genetic architecture of metabolic phenotypes and lend insight into the way selection plays out genetically across multiple variants.

(*b*) *High Altitude Adaptation*

Though diverging as little as 3,000 years ago from sea level populations, highland populations, such as Andeans and Tibetans, exhibit unique suites of physiological characteristics that enable them to live in hypobaric hypoxic conditions 4,000 meters above sea level (Cynthia M Beall, 2007). Such conditions generally lead to increased hemoglobin levels and low birth weight for people from lowland populations. However, rather than having raised hemoglobin levels, Tibetans living at 3,500 to 4,500 meters above sea level increase oxygen availability to cells in spite of low pulmonary oxygen content by increasing resting ventilation and vasodilation and blood flow (Cynthia M Beall, 2007; C M Beall et al., 1997; Q. H. Chen et al., 1997; Ge et al., 1994; Groves et al., 1993). Increased ventilation and vasodilation are both results of increased circulating NO (nitric oxide) levels among Tibetans . In contrast, Andeans living above 2,500 meters do not have the same physiological adaptation to high altitude: they retain the high hemoglobin levels of lowland people living at high altitudes, and as a result have elevated risk of mountain sickness with age and low birth weight (Winslow et al., 1989). However, Andean physiology may have adapted to hypobaric hypoxic conditions in other unrecognized ways (Cynthia M Beall, 2007). Recently, researchers have been studying the genetic underpinnings of hypobaric hypoxia response in Tibetans and Andeans in order to determine whether it evolved adaptively under positive selection in these populations.

Much related research suggests that genes in the hypoxia-inducible factor (HIF) oxygen signaling pathway have been subject to strong and recent positive selection in Tibetan and Andean highlanders. However, the specific genes within the HIF pathway under selection differ between these populations.

One study of both Andean and Tibetan populations identified regions under selection, using neighboring lowland populations as outgroups to strengthen statistical power (Bigham et al., 2010a). This analysis revealed that both populations have experienced selection upon HIF pathway genes, including *EGLN1* (Bigham et al., 2010a). *EGLN1* negatively regulates HIF1α and HIF2α in an oxygen sensitive reaction, thereby decreasing the production of hemoglobin induced by these transcription factors in lower, oxygen-plentiful altitudes (Bigham et al., 2010a; Simonson et al., 2010a). Furthermore, mutations in *EGLN1* that prevent targeted degradation of HIFα's have been associated with excessive production of hemoglobin—excessive hemoglobin production is the maladaptive response of non-adapted humans from lowland populations to hypobaric hypoxic conditions (Simonson et al., 2010a).

Tibetans also experienced positive selection upon variants in *EPAS1* (also called *HIF2α*), which regulates expression of the *EPO* gene responsible for hemoglobin production (Bigham et al., 2010a). One mutation in *EPAS1* is known to cause excessive hemoglobin production (Yi et al., 2010). Finally, Andeans experienced positive selection upon variants in *PRKAA1* (a necessary cofactor for transcriptional regulation by *HIF1*) and *NOS2A* (responsible for NO production, involved in vascular response to hypoxia described above) (Bigham et al., 2010a).

This description of HIF pathway genes under selection in highland populations has been bolstered and expanded upon by other recent studies (see Table 3 for a summary). Another study by Simonson et al. (2010a) found that 10 of 240 genes included in GO categories related to pathways involved in high altitude adaptation lie in regions of high *iHS* or *XP-EHH* distinguishing Tibetans from Han Chinese and Japanese populations. These are significantly more genes under selection in Tibetans than are found to be for randomly sampled sets of 240 genes. Therefore, functional candidates for high altitude adaptation are enriched for signatures of selection above the rest of the genome. These candidate genes in regions under selection in Tibetan populations include *EPAS1*, *EGLN1*, and *PPARA*. *PPARA* is regulated by the HIF complex—in hypoxic mice, expression of *PPARA* is inhibited by HIF1. In humans, administration of a *PPARA* agonist decreased hemoglobin levels. In this study, *EGLN1* and *PPARA* haplotypes subject to incomplete sweeps among Tibetans were correlated with lower hemoglobin concentrations within the Tibetan population. (Simonson et al., 2010a)

A third study by Yi et al. (2010) compared Tibetan and Han Chinese exomes (genic coding regions) to assess SNP frequency change along the Tibetan lineage, using a Danish population for triangulation. Several SNPs of high frequency change in the Tibetan population are involved in oxygen transport and regulation (Yi et al., 2010). Also, the 34 genes in the GO category "response to hypoxia" were enriched for SNPs of high frequency change among Tibetans (Yi et al., 2010). Along the same lines as the other studies, *EPAS1* has the strongest signal for SNP frequency change. Of the dataset, an intronic variant in *EPAS1* at 87% in Tibetans and 9% in Han Chinese has the largest frequency change of all typed SNPs (Yi et al., 2010). This allele is associated with lower hemoglobin levels in the

Table 3. Genetic high-altitude adaptations

Genetic Regions	Functional Connection	Population	Evidence	Source
EGLN1 (*PHD2*)	*EGLN1* negatively regulates *HIF1α* and *HIF2α* in an oxygen sensitive reaction, decreasing the production of hemoglobin in lower, oxygen-plentiful altitudes. Mutations in *EGLN1* that prevent targeted degradation of HIFα's are associated with excessive production of hemoglobin. *EGLN1* haplotypes subject to sweeps in Tibetans correlate with lower hemoglobin concentrations in Tibetans.	Tibetans and Andeans	SNPs in the *EGLN1* gene region are significant for test statistics indicative of recent positive selection, including a high degree of frequency differentiation (LSBL), a reduction in heterozygosity of highlanders compared to lowlanders (ln*RH*), and an excess of rare variants in highlanders compared to lowlanders (a modified Tajima's *D* statistic, called the standardized difference of *D*).	Bigham et al. (2010)
		Tibetans	*EGLN1* lies in a region of extended haplotype homozygosity (measured by high *iHS* and *XP-EHH*), indicative of recent positive selection.	Simonson et al. (2010)
		Tibetans	*EGLN1* is enriched for SNPs of high frequency differentiation between Tibetans and Han Chinese.	Yi et al. (2010)
EPAS1 (*HIF2α*)	*EPAS1* regulates expression of *EPO*, responsible for hemoglobin production. *EPAS1* is expressed in fetal and adult lung, placenta, and vascular endothelial cells. One mutation in *EPAS1* causes excessive hemoglobin production. *EPAS1* mutations at increased frequency among Tibetans correlate with lower hemoglobin concentrations in Tibetans.	Tibetans	SNPs in the *EPAS1* gene region are significant for test statistics indicative of recent positive selection, including LSBL, ln*RH*, and the standardized difference of *D*.	Bigham et al. (2010)
		Tibetans	*EPAS1* lies in a region of extended haplotype homozygosity (measured by high *XP-EHH*), indicative of recent positive selection.	Simonson et al. (2010)
		Tibetans	*EPAS1* is enriched for SNPs of high frequency differentiation between Tibetans and Han Chinese. An intronic SNP in *EPAS1* at 87% in Tibetans and 9% in Han Chinese has the largest frequency change of all typed SNPs.	Yi et al. (2010)
		Tibetans	8 SNPs in the *EPAS1* gene region achieve genome-wide significance for frequency differentiation between Tibetans and Han Chinese.	Beall et al. (2010)

Table 3 (Continued)

PPARA	*PPARA* is regulated by the HIF complex. In hypoxic mice, expression of *PPARA* is inhibited by HIF1. In humans, administration of a *PPARA* agonist decreases hemoglobin levels. *PPARA* haplotypes subject to sweeps among Tibetans correlate with lower hemoglobin concentrations in Tibetans.	Tibetans	*PPARA* lies in a region of extended haplotype homozygosity (measured by high *iHS*), indicative of recent positive selection.	Simonson et al. (2010)
PRKAA1	*PRKAA1* is a necessary cofactor for transcriptional regulation by HIF1.	Andeans	SNPs in the *PRKAA1* gene region are significant for two test statistics indicative of recent positive selection, LSBL and the standardized difference of *D*.	Bigham et al. (2010)
NOS2A	*NOS2A* is responsible for NO production, involved in vascular response to hypoxia.	Andeans	SNPs in the *NOS2A* gene region are significant for two test statistics indicative of recent positive selection, LSBL and the standardized difference of *D*.	Bigham et al. (2010)

Tibetan population (Yi et al., 2010). Other interesting genes identified in this study include *EGLN1*, the only gene also identified as under selection in Andeans by Bigham et al. (2010a).

A fourth study comparing SNP frequencies of Tibetan highlanders with Han Chinese found 8 SNPs of genome-wide significance all located in the region from *EPAS1* to 235kb downstream of it (Cynthia M Beall et al., 2010b). This study further tested noncoding SNPs in the *EPAS1* region for correlation with hemoglobin levels. Twenty-six SNPs were significantly correlated with hemoglobin levels in two replicated highland Tibetan populations; the major allele is associated with lower hemoglobin concentrations in every case (Cynthia M Beall et al., 2010b). (See Table 3 for a summary of genetic regions implicated in high altitude adaptation.)

The research into cold resistance and hypobaric hypoxia adaptation examine multiple varieties of selective signatures left by recent positive selection. Significantly, while some adaptive variants lie in regions of extended haplotype homozygosity due to strong selection on newly mutated variants found on a single haplotype in the populations, others can only be identified by changes in allele frequency. Combining information on known functional pathways and environmental selective agents has proven to be the key to identifying these more subtle candidates for recent adaptive metabolic evolution. The research into high altitude adaptation has further connected genotype to phenotype by testing for polymorphism association with the low hemoglobin phenotype, a phenotype with fitness consequences in mountain sickness and low fetal birth weight. Similarly, some variants proposed as adaptive for cold resistance are correlated with altered protein expression and activity, resting metabolic rate, resting respiratory rate, and altered BMI

and obesity risk. Associations with such phenotypes suggest the variants have an adaptive role in non-shivering thermogenesis, given current mechanistic understanding of uncoupling of mitochondrial oxidative metabolism in brown adipose tissue.

(*c*) *Selection on Genes involved in Metabolic Disorders*

Beyond the examples of cold resistance and hypobaric hypoxia adaptation, signatures of selection have been described for genes with involvement in metabolic disorders. Studies of such candidate genes differ from the previous case studies of metabolic adaptation because the selective agents are unknown.

1. PCSK9

PCSK9 encodes a serine protease known to degrade LDLR—the low density lipoprotein receptor, which uptakes LDL into cells from the plasma. A variety of nonsynonymous mutations in *PCSK9* are known to impact function of the gene, resulting in increased serum LDL levels (with gain-of-function mutations in *PCSK9*) or decreased serum LDL levels (with loss-of-function mutations in *PCSK9*). Natural selection on *PCSK9* was hypothesized due to its relationship to *LDLR*, for which longer scale selection during primate evolution (Q. Wang et al., 2006) and more recent balancing selection in the 3'-UTR region has been demonstrated (Fagundes, Salzano, Batzer, Deininger, & Bonatto, 2005). Also, variation in *PCSK9* function among human populations has been described—the Africans surveyed have a higher frequency of loss-of-function mutations than the Europeans (Ding & Kullo, 2008).

A study of African- and European-Americans found that several test statistics indicate positive selection on gain-of-function mutations in *PCSK9* (Ding & Kullo, 2008). However, the populations differ in the direction of selection on specific alleles. African-

Americans exhibited positive selection on the derived (gain-of-function/higher LDL associated) allele of SNP rs562556 and the ancestral (lower LDL associated) allele of SNP rs505151, while European-Americans exhibited positive selection on the ancestral allele of SNP rs562556 (Ding & Kullo, 2008). While this example demonstrates selection on functional variants impacting metabolism phenotype in different human populations, the implications for metabolic adaptation remain unclear. In part this is likely due to the fact that the study used African-Americans, an admixed population. According to a recent analysis, the admixture in the African-American population is unlikely to produce false-positive results for test statistics of allele frequency spectrum: these test statistics (with the exception of Fay and Wu's H test) are, in fact, underpowered to detect selection on alleles prior to the admixture, so signatures of positive selection detected in African-Americans are likely to be real (Kirk E Lohmueller, Bustamante, & Clark, 2010). However, this population exhibits signals of positive selection on alleles conferring opposing phenotypic effects, muddying the pattern of phenotypic change in the population over time. Since the African-American population is derived from both diverse West African and European ancestries, selection could have favored different phenotypes in these different populations. Another possibility is that this represents fine-tuning in selection upon LDL levels in African-American ancestral populations, in which certain variants were selected for conferring increased LDL levels, while others were selected for conferring decreased LDL levels. Alternatively, the phenotype conferring the alteration in fitness could be an unidentified pleiotropic effect of the alleles.

Based upon research in the Framingham Heart Study, elevated LDL was still associated with increased coronary heart disease, even in a population with lower baseline

nutritional standards (P. W. Wilson, Anderson, & Castelli, 1991). Negative health impacts of having low LDL levels are currently unclear, though having sufficient cholesterol levels for cellular replication in conditions of an elevated metabolic rate (see research regarding *APOE* functional variants, below) could be critical. The selective agent in the case of *PCSK9* remains completely unknown, as comparison of two populations (one of them an admixed population with multiple ancestry sources) is insufficient to speculate upon environmental or cultural sources of variation that might affect fitness of variants underlying metabolic phenotypes. This example of recent positive selection on *PCSK9* is afflicted by uncertainty as to the overall phenotype under selection, the selective agent(s), and the populations that experienced selection, making it impossible to understand adaptation of *PCSK9* during recent human evolution.

2. ANGPTL4

Another investigation into potential adaptive evolution of metabolic genes was conducted by Romeo et al. (2007) on *ANGPTL4*; a gene that contains variants known to affect triglyceride and HDL levels. *ANGPTL4* is expressed in adipose and liver tissue during fasting, and is among a class of genes known to regulate glucose and lipid metabolism based on research in mice models. *ANGPTL4* may inhibit lipoprotein lipase, preventing hydrolysis of triglycerides into fatty acids taken up by adjacent cells, which is consistent with the hypertriglyceridemia and reduction in fat mass in mice with heightened expression of the protein. This study of *ANGPTL4* largely focused on finding functional variants via resequencing of the gene in various ethnic populations including European- and African-Americans. (Romeo et al., 2007)

An excess of nonsynonymous variants, some predicted to have structural impact on the expressed protein, were found in individuals with the lowest quartile of triglyceride levels, while synonymous and noncoding variants were evenly distributed among individuals with all levels of triglycerides. Furthermore, European-Americans were enriched for the nonsynonymous loss-of-function variants over African-Americans, with nonsynonymous to synonymous mutation ratios of 4:1 and 1.3:1 in these populations, respectively. One nonsynonymous variant, E40K, significantly lowered plasma triglyceride and LDL cholesterol in heterozygous European carriers (MAF—minor allele frequency ~1.3%). This finding was replicated in two larger population studies for individuals of European ancestry, [the Atherosclerosis Risk in Communities (ARIC; n=15,792) and the Copenhagen City Heart Study (CCHS; n=10,135)]. (Romeo et al., 2007)

Finally, analysis of test statistics of selection in the resequencing population revealed an excess of rare derived alleles in all ethnic populations, consistent with purifying selection acting upon the locus (Romeo et al., 2007). The excess of nonsynonymous variants with effects on lipid profiles in the European-Americans may be attributed to recent relaxation of selective constraint in this population (Romeo et al., 2007). Given the role of *ANGPTL4* in response to fasting, consideration of changes in lifestyle habits that could have impacted the frequency, severity, and duration of famine in different populations could prove a fruitful avenue for speculation on the selective agent responsible for acting on genetic variants in this region. A change in selective pressure in one human population begs an explanation involving shifts in selective agents across populations. However, the nature of this change in selective pressure remains elusive, again, due to ambiguity of the study populations compared. African-Americans are admixed

from European as well as various West African ancestries. Distinction between these two populations in terms of evolutionary ancestry and selective pressures is unclear, making speculation on the relationship between *ANGPTL4* selection and phenotypic fitness consequences in particular environments impossible.

3. ALMS1

ALMS1, a gene involved in ciliogenesis, is a third example of a gene under selection in certain human populations, with variants underlying a particular metabolic phenotype. Certain variants in the *ALMS1* gene region cause a metabolic disorder called Alström Syndrome, resulting in early onset obesity and type II diabetes (Laura B. Scheinfeldt et al., 2009). A recent resequencing study detected positive selection upon standing variation in Eurasian populations (Laura B. Scheinfeldt et al., 2009). This selection within the past ~15,000 years has resulted in near fixation of seven nonsynonymous mutations in the East Asian population sampled (Laura B. Scheinfeldt et al., 2009). The phenotypic consequences of these putatively selected nonsynonymous variants remain unknown (Laura B. Scheinfeldt et al., 2009). The fact that some mutations in *ALMS1* cause a severe metabolic syndrome makes it likely that these seven nonsynonymous polymorphisms also influence metabolites related to energy usage (Laura B. Scheinfeldt et al., 2009). However, more studies could be conducted, involving the use of cell-culture and transgenic mouse models, to clarify phenotypic change caused by the seven nearly fixed polymorphisms in East Asians. The study does find modest support for association between nonsyndromic variation in *ALMS1* and five insulin and glucose related phenotypes by pooling unpublished results from four previous GWAS studies (Laura B. Scheinfeldt et al., 2009). A relationship to insulin or glucose profiles could implicate changes in carbohydrate content or patterns

of energetic expenditure as responsible for altering the fitness of genetic variants in *ALMS1*. However, the selective agent remains unknown in the case of *ALMS1* metabolic adaptation due to a lack of understanding of the phenotypic consequences of selected variants as well as environmental or cultural variability across selected versus unselected populations worldwide. Scheinfeldt et al. (2009) acknowledge that any influence of *ALMS1* variants under selection on insulin or glucose profiles may be merely "tangential to the primary selective force," meaning that a phenotypic consequence of a variant under selection offers no proof of the selective agent. However, such phenotypic information would still be helpful in forming competing hypotheses of the selective regimen under which *ALMS1* evolved in diverse human populations.

4. TCF7L2

TCF7L2 is a transcription factor with a well-known association with type II diabetes, which has been replicated in many studies in diverse populations from Europe, East Asia, and West Africa (Dupuis et al., 2010; Helgason et al., 2007; J. R. B. Perry & Frayling, 2008). The study by Helgason et al. (2007) detected a signature of selection surrounding a haplotype at high frequency in East Asians (0.95), moderate frequency in Europeans (0.58), and low frequency in Yorubans (0.1). Recent positive selection in the past 10,000 years drove this non-diabetes associated haplotype to near fixation in the East Asian population and may have driven it to increased frequency in the European and Yoruban populations as well (Helgason et al., 2007). The phenotypic effect conferring fitness advantage of the non-diabetes associated haplotype is unclear, though the team detected association between the haplotype and increase in BMI (Helgason et al., 2007), which has not been replicated by any subsequent GWAS. Interestingly, the team found that the haplotype associated with type II

diabetes is correlated with decreased BMI, opposite to the prediction made by the Thrifty Genotype Hypothesis (Helgason et al., 2007). Overall, though the timing of selection upon the non-diabetes conferring haplotype of *TCF7L2* correlates with the onset of agriculture in the East Asian and European populations (Helgason et al., 2007), the phenotypic effects of the selected haplotype are unclear so the mechanism by which the novel genotype conferred fitness advantage in selected populations cannot be verified.

5. APOE

APOE, involved in binding, uptake, and catabolism of lipoprotein cholesterol, has variants associated with cholesterol levels (Kathiresan, Willer, Peloso, Demissie, & K, 2008). In particular *APOE* ε4, a nonsynonymous ancestral polymorphism, has been associated with increased total cholesterol, low density lipoprotein cholesterol, cholesterol absorption, and responsiveness to dietary fats (Eisenberg, Kuzawa, & Hayes, 2010). Researchers hypothesized that due to the important role of cholesterol in cell proliferation and tissue growth, *APOE* ε4 experienced positive selection for increased cholesterol absorption in populations exhibiting increased basal metabolic rates due to either extreme high or low environmental temperatures (Eisenberg et al., 2010). Consistent with this, they discovered a significant curvilinear relationship with ε4 frequency increasing in populations at both high and low absolute latitudes, rough proxies for historical temperatures, across 268 world populations controlling for neutral population structure (Eisenberg et al., 2010). This study is relatively strong due to the logical connection between phenotypic association of the variant and fitness consequences in certain environments, as well as the large number of populations and control for shared population histories. However, studies of this nature would still benefit from confirmation

of the link between genotype and phenotype. While the pattern of world frequency data in relation to presumed fitness consequences of the variant is intriguing, conducting experiments using cell and mouse models to confirm the phenotypic consequences would elevate understanding of human metabolic adaptation to local environments.

The Challenge of Pleiotropy

One challenge in establishing phenotypic adaptation via selection for a particular genetic variant is that the selection may be acting on a pleiotropic effect of the selected variant. When phenotypic differences are observed between populations, several possible explanations exist. The phenotypic differences could be a result of different environments, epigenetic effects, or genetic differences. If the phenotypic difference is genetically encoded, three possibilities exist. The phenotypic differences could be the result of drift; the phenotypic differences could be adaptive; or the phenotypic differences could be a result of selection on a linked or related trait—pleiotropy. Pleiotropy may be difficult to rule out in many cases. However, some recent examples highlight the importance of considering it in studies of recent human adaptation.

Short stature among rainforest pygmies living in Africa, South East Asia, certain Pacific Islands, and the Amazon has long been a mystery of recent human evolution. What adaptive value does this striking physiological change have for rainforest peoples? According to a recent investigation into selective signatures in a broad sample of world populations by López Herráez et al. (2009), this may have been the wrong question to ask. Their study detected a region of selection surrounding *TRIP4* in Mbuti Pygmies and *IYD* in Biaka Pygmies of central Africa (López Herráez et al., 2009). Both of these genes are

involved in the iodide-dependent thyroid hormone pathway, and *IYD*, in particular, catalyzes deiodination of metabolites of the thyroid hormone pathway and salvages iodide in the thyroid gland (López Herráez et al., 2009). Rainforest environments are generally deficient in iodine. Interestingly, Efe Pygmies (a population of Mbuti Pygmies) living in an iodine-deficient rainforest were recently found to have much lower rates of goiter than the unrelated Bantu ethnicity living nearby (9.4% versus 42.9%) (López Herráez et al., 2009). This relationship between selection on the same iodine-dependent pathway in two independent populations of pygmies, knowledge of environmental variation in iodine between rainforest and other ecoregions, and evidence of adaptation to iodine deficiency by one of the pygmy populations led the researchers to hypothesize that pygmy populations are adapted genetically to low iodine levels in rainforest environments (López Herráez et al., 2009). Significantly, variation in the thyroid hormone pathway has been connected to short stature (López Herráez et al., 2009). So, stature may not have been the target of selection at all, but rather iodine levels.

This hypothesis needs to be more thoroughly tested. However, in spite of current uncertainty as to the validity of selection on the iodine-dependent thyroid hormone pathway in creating the pygmy phenotype of rainforest populations, this example presents a realistic possibility that the short stature of pygmies may be a pleiotropic result of selection for coping with iodine deficiency. The easily recognized short stature phenotype may not be adaptive at all and may be merely a byproduct of selection on genes that affect iodine metabolism. Keeping this example in mind, wariness of premature conclusions about adaptive phenotypes is necessary. However, improved understanding of biological pathways and gene function may assist in this challenge as well. In the above example, the

relationship between the genetic variants under selection and iodine metabolism, as well as stature, enabled the generation of a novel alternative hypothesis of phenotypic adaptation to be tested. As genetic variants under selection continue to be described in the future, thorough examination of all the biological pathways they impact may prove to be important to deciphering true phenotypic adaptations.

Assessing Hypotheses

Given the current evidence for selection on metabolic genes during recent human evolution, can any conclusions be drawn about pre-existing hypotheses regarding recent human adaptations? Adaptation to high altitude among Tibetans seems certain, though cell-culture and transgenic mouse models (and other tests relating genotype to phenotype) could further clarify the phenotypic effects of selected genetic variants. Adaptation to cold stress also seems likely among Northern populations, though more work should be done verifying the molecular mechanisms by which this has occurred using the methods mentioned above as well as more extensive exploration of genetic selection among different populations in relevant pathways of mitochondrial oxidation and lipid metabolism for brown adipose tissue non-shivering thermogenesis.

Little evidence has accumulated that directly supports the Thrifty Genotype Hypothesis, while it remains the sole hypothesis to explain how vast shifts in energy attainment and usage have impacted human metabolic phenotypes and genetic evolution. While signatures of selection surrounding genes that affect metabolic disorders seem common, the direction of selection on metabolic phenotypes is unclear with respect to agricultural populations, who according to the hypothesis should have increased risk of

type II diabetes and metabolic syndrome as a result of shifts in selection pressure related to the agricultural lifestyle and diet. For example, the haplotype in *TCF7L2* associated with type II diabetes is associated with decreased BMI (Helgason et al., 2007), while the Thrifty Genotype Hypothesis predicts that variants that increase risk of Type II diabetes should also increase other symptoms of metabolic syndrome like obesity.

Also, little evidence exists to support different variants conferring risk for the same metabolic disease being selected in the same direction in the same populations. A study recently conducted by Klimentidis et al. (2011), assessing signatures of selection surrounding loci linked by GWAS to obesity and type II diabetes in 53 populations of the Human Genome Diversity Panel, was the first research to directly test the predictions of the Thrifty Genotype Hypothesis at a genetic level on world populations. This study found that regions surrounding type II diabetes risk factors were highly differentiated between East Asians and the other broad geographic groups, and also between sub-Saharan Africans and other groups (Klimentidis et al., 2011). However, the study did not state whether risk alleles were found at a higher or lower prevalence in these groups. In addition, type II diabetes associated loci exhibited a slight excess of extended haplotype length (high *XP-EHH*) among East Asians (Klimentidis et al., 2011). Also, South Asians and Europeans exhibited limited evidence for having an excess of extended haplotype length (high *REHH*) surrounding obesity risk factors, though this failed to achieve statistical significance (Klimentidis et al., 2011). These limited conclusions have some suggestive value with regards to selection and phenotypic change: in particular, the East Asian population exhibits selection surrounding loci that alter type II diabetes risk, while South and East

Asians experience higher risk of type II diabetes at lower BMIs than do other populations (Chiu et al., 2011; Seidell, 2000).

However, the broadness of the defined ethnic populations limits ability to parse populations accurately between subsistence regimens to properly test the predictions of the Thrifty Genotype Hypothesis. Furthermore, abundance of risk alleles is not assessed across the ethnic groups. Additionally, while the Thrifty Genotype Hypothesis predicts a concordance in the selective regimens upon type II diabetes and obesity loci, this study presents evidence of different selective regimens happening on Type II diabetes and obesity loci across these world populations. Finally, lack of complete knowledge of the genetic risk factors for metabolic diseases in different populations limits the implications of this analysis. Most risk factors for Type II diabetes and obesity identified by GWAS are solely identified in European populations; the genetic underpinnings of these metabolic diseases may vary in different populations, especially if rare variants, likely specific to particular populations, play a prominent role in disease architecture. Since the loci identified by GWAS as risk factors for disease are not necessarily the causal alleles, and may just be in linkage disequilibrium with the causal alleles, this could pose serious discrepancies when these alleles are assessed for risk on different genetic backgrounds (i.e. on different haplotypes in diverse ethnic populations). Importantly, GWAS with large East and South Asian cohorts have recently been published that may help to alleviate this bias (Cui et al., 2011; Kooner et al., 2011; Shu et al., 2010; Sim et al., 2011; Takeuchi et al., 2009; Xu et al., 2010; Yamauchi et al., 2010). While this preliminary analysis by Klimentidis et al. (2011) is interesting, much phenotypic variation among populations for risk of type II diabetes and metabolic syndrome may stem from environmental and epigenetic rather

than genetic causes (Stöger, 2008). Acknowledging the difficulties inherent in collecting the type of information necessary to properly test the Thrifty Genotype Hypothesis, evidence supporting the hypothesis remains inconclusive. Therefore, taking enrichment of metabolic genes for signatures of selection as support for the hypothesis may be premature.

Generating New Hypotheses

If the evidence surrounding the Thrifty Genotype Hypothesis has not been conclusive, what alternative hypotheses exist? The prevalence of genes of metabolic function in scans of selection across the genome is indicative of some profound shift in energy metabolism among certain populations. Agricultural and pastoral changes in diet that reduced food diversity, decreased folate content, altered starch/carbohydrate content, altered proportions of fatty acids, and decreased meat consumption may have contributed to this shift in energy metabolism. Additionally, agriculture and pastoralism increases the frequency of famines and the seasonality of energy consumption and energy expenditure among humans, especially those living at northern latitudes. Such hypotheses of selective agents upon recent human metabolism are currently challenging to test genetically as the biological pathways affected by such changes in energy metabolism have not been comprehensively described.

In addition to these potential dietary selection factors upon human metabolism, other aspects of the human metabolic phenotype, such as the heightened propensity for accumulation of large deposits of white adipose tissue, bear explaining in terms of adaptive evolutionary change. Differences in energy usage between humans and chimpanzees may have evolved at many points during the evolution of hominins. However, even adaptations

that begin early during human evolution may continue into recent human evolution (Carroll, 2003). This means that more ancient adaptive metabolic shifts initiated in the hominin lineage to accommodate larger brain size, cooking, endurance exercise, thermal regulation, and a higher quality diet may have continued to manifest in selective genetic change in recent human evolution. The time frame necessary for complete evolution of adaptive traits is theoretically challenging to assess. Long waiting times for fitness enhancing mutations, given small mutational target sizes, makes it possible that new adaptive peaks could be explored and discovered long after the evolution of a novel adaptation has begun.

In order to generate novel testable hypotheses regarding variation in human environments relating to phenotypic variability across human populations, two major improvements must take place. First, variation in human environments must be more precisely described, and second, variation in human phenotypes across populations must be more precisely described. In the first case, environmental variables such as temperature, altitude, UV radiation are easily quantified for different locales. On the other hand, climate change and migration of human populations may confound the validity of modern measurements of these variables, as human populations may remain adapted to a previous climate or locale given the lag-time necessary for selection to alter gene frequencies in a population. In addition, variation across human diets is more challenging to quantify than climate variables. Currently researchers must rely on ethnographic studies of variable quality in order to explore dietary variation across populations (e.g., Murdock, 1967), which may be inaccurate or incomprehensive. Furthermore, diets may have changed very recently in some human populations, meaning that the diet the population is adapted to

may not be what they are still eating. Still, understanding of variation across environments that human populations spread to and evolved in within the past 50,000 years stands to be improved greatly. Collaboration with climate scientists could improve understanding of recent changes in climate across diverse geographies. Collaboration with ethnographers in order to fact check and gain historical insight into changes in human diets would also be helpful. Macro- and micromolecular content of diverse foods could be more precisely quantified. Finally, demographers and historians could verify recent human population movements over the past several centuries. Understanding of variation across human environments (including diet and culture) will not be easy to accomplish, but efforts to improve this understanding will yield vastly more diverse selective pressures to be explored than simple variation in temperature, altitude, and UV exposure.

In addition, variation in human phenotypes across populations bears more full exploration. Obvious phenotypic differences, such as skin color, have already been described (Nina G. Jablonski & Chaplin, 2000). Defining additional relevant and finite human phenotypes for study poses a challenge for researchers. Nevertheless, the medical community will be increasingly helpful in this endeavor as phenotyping of people from more diverse ethnicities is important for health treatment. Variation in disease risk, metabolic profiles, measures of adiposity, aerobic metabolism, and capacities for cold tolerance and fasting tolerance across populations will infinitely contribute to the generation of hypotheses connecting environmental variation to adaptive phenotypic variation. As phenotype is a product of both genes and the environment, environmental differences (climates, diets, behaviors) across populations, as well as epigenetic effects, will impact some of these metabolic profiles. Yet, once these environmental and phenotypic

variables have been more fully described and considered, signatures of selection in particular genes, pathways, and populations can be more meaningfully explored.

A final consideration for testing novel hypotheses of recent metabolic adaptations in humans concerns the availability of world population samples for genetic data. Often in current studies the populations analyzed do not afford insight into environmental variation that might be acting as the selective agent. Populations are generally chosen for convenience rather than chosen to span relevant environmental variables of interest. Understanding the genetic underpinnings of metabolic disease in admixed populations such as African-Americans and Latin-Americans has very important implications for health treatment, but translating such information into evolutionary understanding is challenging due to the admixture. Moreover, when particular populations are tested for selection across certain genetic regions without any *a priori* hypothesis of the agent of selection, a large number of independent populations must be used in order to have the statistical power to evaluate correlations with environmental selective pressures. Unfortunately, the availability of diverse population samples is currently limited for many researchers to those included in existing panels of diversity, which may be insufficient to generate or test many selective hypotheses.

CONCLUSIONS

1. Recent human metabolic adaptations come at the end of a long line of metabolic changes that arose during human evolution. Genes with evidence of selection earlier in human and primate evolution may continue experiencing selection in modern human evolution.

2. Known examples of human genetic adaptations to cope metabolically with novel environments represent "low hanging fruit". Finding more genetic adaptations may pose a difficult challenge for modern research and will require a better understanding of:
    - how the diets and climates of diverse populations have changed over time.
    - how human phenotypes across populations are shaped by environmental as well as genetic variation.

    Collaboration of researchers from many fields, including climate, ethnographic, and medical research, may be required to answer these questions.
3. Determining the biological pathways that comprise genetic interaction requires painstaking molecular research. Further research that clarifies these biological pathways will aid in connecting genotypic variation in human populations to phenotypic variation.
4. Proof of metabolic significance of genetic elements under selection requires functional follow-up studies involving cell culture work and transgenic mouse models in order to verify quantifiable adaptations related to candidate loci under selection (Sholtis & Noonan, 2010). However, these methods involve considerable expense and may offer limited insight in cases in which the effect size of a variant is small or the mouse background interferes with the phenotype or function being tested.
5. Future discoveries of recent metabolic adaptations due to hard selective sweeps may be rare (Hernandez et al., 2011; Kirk E. Lohmueller et al., 2011). Since selection may be acting on multiple standing variants, generating new maps of signatures of

selection will require developing statistical tools to explore the timing and strength of selection during soft sweeps. Also, these new statistical tools must grapple with full-genome resequencing rather than genotyping data (Nielsen, 2010). These efforts, combined with better information on population histories, climates, and diets; genetic pathways; and functional follow-up studies will illuminate the mechanisms by which humans have recently adapted metabolically to local environments.

Final Comments

This article was published in 2012 in the journal of *Biological Reviews* (E. A. Brown, 2012). I conducted and wrote this review independently with helpful feedback from Maryellen Ruvolo, Peter Ellison, David Reich, Alexander Banks, Amanda Lobell, and two anonymous reviewers. Since conducting this review, research has advanced for several of the topics covered.

High-Altitude Adaptations

High-altitude adaptation among Ethiopian highlander populations has now been studied genetically in several analyses, finding different suites of hypoxia and other genes under selection in these groups (Alkorta-Aranburu et al., 2012; Huerta-Sánchez et al., 2013; Udpa et al., 2014). In addition, researchers discovered that genetic adaptations in the *EPAS1* gene in Tibetans, which increase hemoglobin levels, introgressed from an ancient Denisovan haplotype (Huerta-Sánchez et al., 2014).

Thermoregulatory Adaptations

Signatures of selection of genomic data from multiple high-latitude ancient and modern human groups have now been analyzed to detect potential adaptations to novel cold environments (Cardona et al., 2014; Sazzini et al., 2014). In addition, a selected mutation in the *EDAR* gene in East Asians was shown using a mouse model to impact sweat gland morphology indicating metabolic adaptation to hot environments among some Asian populations (Kamberov et al., 2013).

Tropical Rainforest Adaptations

Researchers have more fully characterized genetic diversity in tropical rainforest populations, and established short stature as having genetic determinants in West African pygmies (Becker et al., 2011; Verdu et al., 2009). In addition, a few studies have examined rainforest populations in Africa, Southeast Asia, and Papua New Guinea for genomic signatures of selection that might surround loci involved in the short-height phenotype (Mendizabal, Marigorta, Lao, & Comas, 2012; Migliano et al., 2013; George H Perry et al., 2014).

CHAPTER 3

MANY WAYS TO DIE; ONE WAY TO ARRIVE:

HOW SELECTION ACTS THROUGH PREGNANCY

ABSTRACT

When considering selective forces shaping human evolution, the importance of pregnancy to fitness should not be underestimated. Although specific mortality factors may only impact a fraction of the population, birth is a funnel through which all individuals must pass. Human pregnancy places exceptional energetic, physical, and immunological demands on the mother to accommodate the needs of the fetus, making the woman more vulnerable during this time period. Here, we examine how metabolic imbalances, infectious diseases, oxygen deficiency, and nutrient levels in pregnancy can exert selective pressures on women and their unborn offspring. Numerous candidate genes under selection are being revealed by next-generation sequencing, providing the opportunity to further study the relationship between selection and pregnancy. This relationship is important to consider to gain insight into recent human adaptations to unique diets and environments worldwide.

SELECTION AND PREGNANCY

Some of the earliest records of mortality from London in John Graunt's "Bills of Mortality" for 1662 reveal a number of distinct causes of death in the population (Graunt, 1662). Any specific cause of death impacts only a fraction of the population, lessening each particular factor's importance to fitness (Figure 2) [*e.g*. (Heron, 2012)]. Managing to be

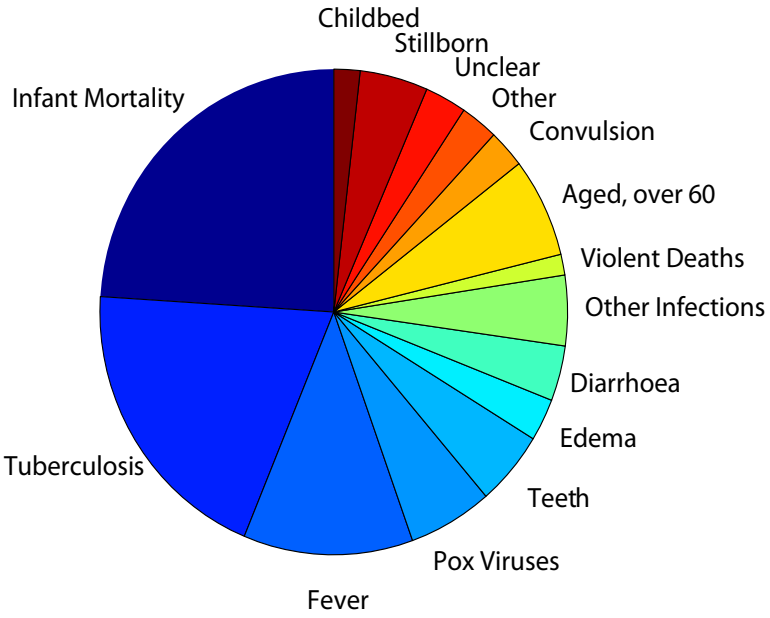
Figure 2. Multiple Varied Causes of Death in Modern Historic Populations

(A)Many different factors caused death for individuals who died in London in 1632. 'Childbed' referred to mothers who died during or after labor, often due to infections. Over a quarter of deaths occurred in infants and unborn fetuses. (B) By contrast, the leading causes of death in modern, developed countries, such as the USA in 2009, are very different, with heart disease and cancer accounting for fully half of the deaths.

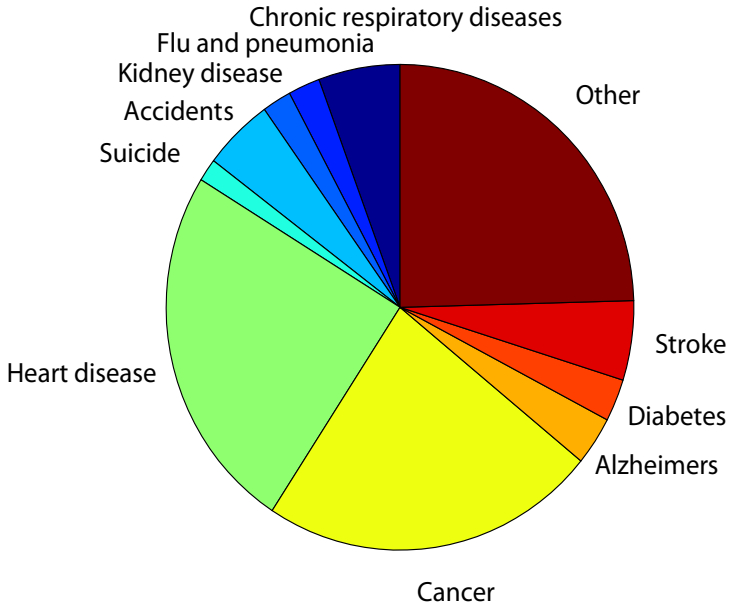
Figure 2 (Continued)

Multiple Varied Causes of Death in Modern Historic Populations

(a) **Reported Causes of Death in London, 1632**



(b) **Ten Leading Causes of Death in US, 2009**



born, however, is a universal requirement for fitness. Thus, factors that influence fecundity and pregnancy are likely to strongly shape human evolution.

The many physiological compromises of pregnancy make it a tremendous challenge for both mothers and infants and a potential selective force. In order to provide for the growing fetus, mothers increase blood sugar (Butte, 2000), blood volume, and hemoglobin count (J. Pritchard, 1965); remodel uterine arteries (Kaufmann, Mayhew, & Charnock-Jones, 2004); and decrease vascular resistance (Sladek, Magness, & Conrad, 1997). These changes put the mother at risk of diabetes, high blood pressure, strokes, hemorrhaging, and seizures (Butte, 2000; Hermida et al., 2000; James, Bushnell, Jamison, & Myers, 2005; Jolly, Sebire, Harris, Regan, & Robinson, 2003). Moreover, properties of the immune system are down-regulated to prevent immune response to the "foreign" fetus, potentially contributing to pregnant women's greater susceptibility to infectious disease (Robinson & Klein, 2012).

These difficulties for mothers translate into problems for infants as well: pre-industrial data show nearly a quarter of babies died during labor and infancy, while maternal mortality was nearly 1.5% per birth due to infectious diseases, diabetes, eclampsia, and jaundice (Woods, 2009). Similarly, modern foraging populations and sub-Saharan African nations in 1970 also had infant mortality rates of 20% to 25%, in contrast to Norway, for example, at only ~1.6% (Marlowe, 2005; Rajaratnam et al., 2010). Maternal mortality in sub-Saharan Africa was ~1.0% in the year 2000 (comparable to 16th and 17th century England) with hemorrhage, hypertension (preeclampsia/eclampsia), and infectious diseases as the major causes. By contrast, maternal mortality in Northern Europe was only 0.02% in the year 2000 (Ronsmans & Graham, 2006). These data from historic,

foraging, and developing country populations only serve as rough proxies for the conditions facing humans during recent evolution, but they give some indication of the difficulty of pregnancy experienced by pre-modern foraging and Neolithic populations.

In addition to the challenges of pregnancy, the number of babies a woman births, compounded across generations, can have huge evolutionary impact. For example, landless Finnish women living 1760-1849 had an average of 4.27 babies, whereas landowning women had an average of 4.55 babies: a change in absolute fitness of this magnitude would cause a geometric rise in the number of descendants in a few generations (Courtiol, Pettay, Jokela, Rotkirch, & Lummaa, 2012) (Figure 3a). The nutritional benefits of the Industrial Revolution (circa 1880) boosted average Finnish fertility to 5.3 babies (Liu, Rotkirch, & Lummaa, 2012). Any such increase in fertility from either environmental or genetic factors will dramatically increase women's fitness (Figure 3b). An earlier revolution, the development of agriculture and pastoralism, may have conferred similar fertility benefits, especially to women with genetic mutations allowing them to maximally exploit these new resources—lactase persistence, described below, may be an example of this (Laland, Odling-Smee, & Myles, 2010). Furthermore changes in female fertility could have played an important role during human population migrations. For example, a large study of Quebecois settlers indicated that women on the wave front of territory expansion had a fertility 15-20% advantage and a heritable component for fertility, suggesting that genes influencing fertility may be shaped by selection (Moreau et al., 2011).

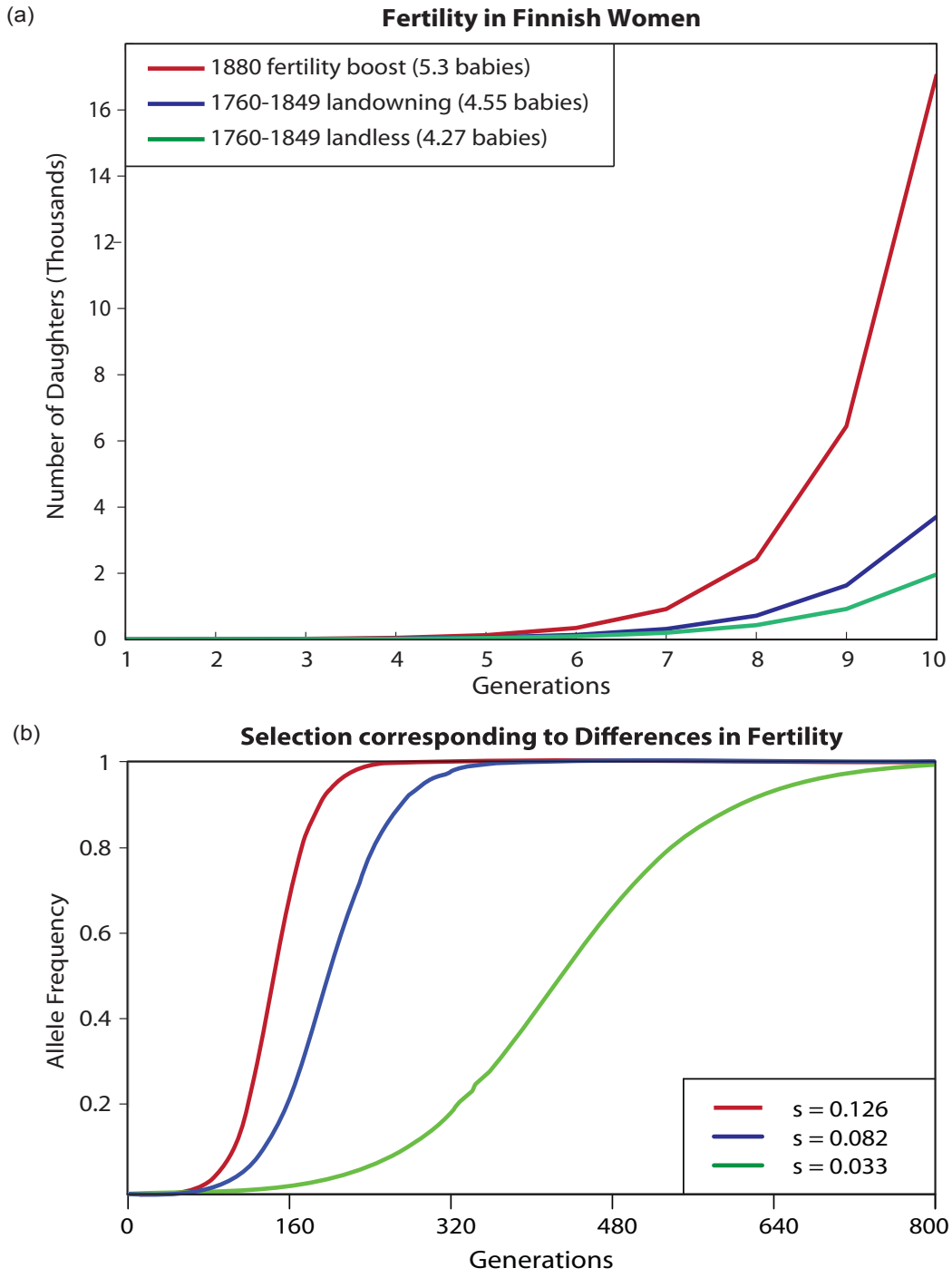
Considering the impact of female fertility alongside the challenges of pregnancy may be critical for understanding recent human adaptations. This chapter explores how selection may have acted through pressures on mothers and infants during pregnancy

Figure 3. Rapid Change in Prevalence of Fertility Enhancing Traits

(A) The increase in number of female descendants (y axis in thousands), compounded across generations, for maternal lineages with an average of 5.3, 4.55, or 4.27 babies over a lifetime, based on pre-industrial data on differences in female fertility in Finland [14,15]. (B) The increase in frequency of new mutations conferring fertility advantages that correspond to the differences in fertility for the three groups of Finnish women (selection coefficient s = 0.126 for 5.3 vs 4.27 babies; s = 0.82 for 5.3 vs 4.55 babies; s = 0.033 for 4.55 vs 4.27 babies). This demonstrates how readily any mutation with a positive impact on female reproduction will sweep through a population over a very short time due to the compounding effect across generations. Reprinted from "Many Ways to Die, One Way to Arrive," by E. A. Brown, M. Ruvolo, and P. C. Sabeti, 2013, *Trends in Genetics*, 29(10), 586.

Figure 3 (Continued)

Rapid Change in Prevalence of Fertility Enhancing Traits



given the changing environment, diet, and behavior of the last 10,000 years. These factors are critical to bear in mind as opportunities for evolutionary geneticists to generate new adaptive hypotheses proliferate, fueled by next-generation-sequencing data and new statistical tools for predicting adaptive variants in diverse populations.

Metabolic Disorders and Selection during Pregnancy

Theories of human adaptation surrounding metabolic disorders, such as hypertension and type 2 diabetes, are constrained by the fact that these diseases typically strike at post-reproductive ages. The related disorders of gestational diabetes mellitus (GDM) and preeclampsia (hypertension in pregnancy), however, occur precisely during the critical reproductive period of pregnancy. GDM occurs as a mother's blood glucose level rises to nourish the fetus, increasing risk of maternal diabetes (Barbour et al., 2007). Preeclampsia occurs as a mother increases blood volume and remodels vasculature for fetal ventilation, raising the risk of maternal hypertension (Hermida et al., 2000). Women predisposed for these conditions can be pushed into metabolic dysfunction.

GDM and preeclampsia are common diseases with grave consequences in pregnancy and thus may strongly impact reproductive fitness. GDM affects 4%-20% of pregnancies in different populations worldwide (Savitz, Janevic, Engel, Kaufman, & Herring, 2008). It can cause macrosomia, in which the fetus grows too large to fit through the maternal pelvis (Dunsworth, Warrener, Deacon, Ellison, & Pontzer, 2012; Langer, Yogev, Most, & Xenakis, 2005; Rosenberg & Trevathan, 2002; Sermer et al., 1998). Before the advent of Caesarian-sections, GDM could lead to fetal morbidity and mortality and maternal hemorrhage and tearing during delivery (Jolly et al., 2003; Langer et al., 2005). Preeclampsia is the leading

cause of maternal mortality worldwide accounting for 10%-19% of deaths (Duley, 1992; Moodley, 2008; WHO, 2005). It can cause fetal hypoxia and oxidative stress, low birth-weight, and maternal hemorrhage and seizures (eclampsia) if not treated by premature delivery (WHO, 2005). (See the section on oxygen and selection during pregnancy for a discussion of high-altitude adaptation and the risks of preeclampsia.)

The rates of GDM and preeclampsia vary significantly in different populations, even when controlling for environmental factors like obesity (Caughey, Cheng, Stotland, Washington, & Escobar, 2010; Hunsberger, Rosenberg, & Donatelle, 2010). This raises the possibility that selective pressures during pregnancy have fine-tuned metabolism to suit different environments and diets around the world, resulting in the current distribution of disease prevalence. On the other hand, alternative explanations, discussed in the section below on alternative hypotheses and avenues of research, may also account for these patterns—distinguishing between these competing hypotheses is an important avenue for future research.

Intriguingly, incidence of GDM among modern populations is inversely related to traditional consumption of dietary components known to increase risk for diabetes and GDM (Table 4). These include high glycemic carbohydrates, which produce large glucose responses in the blood, and dairy products, which produce large insulin responses due to the effect of leucine in whey proteins (Holt, Miller, & Petocz, 1997; Hoppe, Mølgaard, Vaag, Barkholt, & Michaelsen, 2005; Hoyt, Hickey, & Cordain, 2007; C. Zhang, Liu, Solomon, & Hu, 2006; C. Zhang & Ning, 2011). Europeans have the lowest prevalence of GDM in the world—3.6% in a study of over a million births in New York City (NYC) (Savitz et al., 2008)—yet have the longest history of high glycemic diets. In the past 10,000 years,

Table 4. Relationship between metabolic diseases of pregnancy and traditional diets

GDM Incidence, Glycemic Index, and Dairy Consumption						
Population	**GDM Incidence**	**Diet**	**Dairy**	**Agriculture**	**Glycemic Index**	**Reference**
European-Americans	3.6%[a]	70% Carbohydrate; Grain-based	Yes	Yes	High	15, 30, 31
Hunter-Gatherers	?	3-50% Carbohydrate; Game, Tubers, Vegetables, Fruits, Nuts	No	No	Moderate	30
Bangladeshis	7-9%[b] 21.2%[a]	Rice, Fish	No	Yes	Moderate	15, 32, 33
African-Americans	4.3%[a]	Agriculture, Pastoralism, or Hunter-Gatherer	Mixed	Mixed	Moderate	15

Preeclampsia Incidence and Traditional Salt Consumption				
Population	**Preeclampsia Incidence**	**Salt Consumption**	**Obesity**	**Reference**
European-Americans	2%[a]	?	High	40, 41
Sub-Saharans	3.3-3.9%[a]	Low, especially in rainforests	Low	40, 41
African-Americans	4.6%[a]	Low, mixed ancestry	High	40, 41
Iranians	0.6%[a]	High, due to soil salinity	Medium	40, 41, 42
Japanese	1.2%[a]	High, due to seafood	Medium	40, 41, 43

[a]Incidence for populations living in New York City
[b]Incidence for populations living in Bangladesh

European grain-based agriculture increased carbohydrate consumption to roughly 70% of diet, while hunter-gatherers consume only 3-50% (Ströhle & Hahn, 2011). In the past 8,000 years, Europeans also began consuming dairy products in large quantities (Myles et al., 2005). In comparison, South Central Asians had a much higher incidence of GDM in the NYC cohort (14.3%) with Bangladeshis the highest at 21.2% (Savitz et al., 2008). Traditionally, Bangladeshis have had high consumption of fish, a low glycemic food; rice, of moderate glycemic index due to little processing; and no dairy (Atkinson, Foster-Powell, & Brand-Miller, 2008; Itan, Jones, Ingram, Swallow, & Thomas, 2010). Finally, among African-Americans, incidence of GDM was intermediate at 4.3% (Savitz et al., 2008). This is consistent with their admixed ancestry and the mixed consumption of dairy across populations in West Africa, the origin of most U.S. African-Americans.

Given the inverse correlation between traditional consumption of dietary components increasing GDM risk and current incidence of GDM, high glycemic foods and dairy may have acted as selective agents on metabolism during pregnancy. Since GDM is very likely to have a genetic basis - 67% of the risk of type 2 diabetes for adults younger than 60 is heritable (Almgren et al., 2011), and women with GDM have 7-12x elevated risk for type 2 diabetes (Bellamy, Casas, Hingorani, & Williams, 2009; Metzger et al., 2007) - natural selection can act on its underlying risk factors. Therefore, any population environmentally at risk for GDM without access to C-sections should experience selection against genetic risk factors for GDM. Conversely, any population without access to high glycemic food items should experience selection to make blood sugars more available to the fetus, perhaps through increasing insulin resistance by increasing the frequency of genetic risk factors for GDM. Supporting these predictions, evidence suggests Europeans

may have a blunted glycemic response to food compared to other populations, which could be a result of this selection on maternal metabolism to suit diet (Dickinson, Colagiuri, Faramus, Petocz, & Brand-Miller, 2002; Henry et al., 2008).

Like GDM, preeclampsia has an incidence that varies across populations, and it appears to have an inverse relationship with the dietary risk factor of salt-intake (Table 4) (Reyes, Garcia, Ruiz, Dehghan, & López-Jaramillo, 2012). In a study of preeclampsia in NYC, preeclampsia rates were lower among immigrants from East Asia (1.4%), especially Japan (1.2%) and Taiwan (0.9%), and lowest in the world among Iranians (0.6%) (Gong, Savitz, Stein, & Engel, 2012) compared to an incidence of 3-5% of pregnancies in other developed countries (WHO, 2005). Although these populations are less obese than Americans, Japanese and Iranians have historically high salt-intakes due to consumption of coastal foods (Japan) and high soil salinity (Iran) (I. J. Brown, Tzoulaki, Candeias, & Elliott, 2009; FAO, 2012; INTERSALT, 1988).

High salt-consuming populations, such as Japanese and Iranians, may have experienced strong selection to protect them from the deadly threat of preeclampsia. Since the heritability of preeclampsia is 0.55 according to a study done in a Swedish cohort (Cnattingius, Reilly, Pawitan, & Lichtenstein, 2004), this provides variation for selection to act on. Populations consuming large amounts of salt should experience strong selection against genetic risk factors for preeclampsia in the absence of modern medical support for premature deliveries. Supporting this, insensitivity to salt in the diet is common in Japanese: women consuming the most salt (20.6g/day) have no more hypertension than those consuming the least 8g/day (Miura et al., 2010). For comparison, the WHO recommends less than 5g/day of salt consumption for adults (WHO, 2010).

Adaptation for consuming a high glycemic, high dairy diet may have been the result of selection in Europeans through the pressure of GDM, whereas adaptation for consuming a high salt diet may have evolved in Japanese and Iranians through the selective pressure of preeclampsia. On the other hand, alternative hypotheses may also explain the trends described (see alternative hypotheses section). In the past several thousand years, populations migrated to new environments and invented new methods of food extraction and processing, such as agriculture, pastoralism, and fishing. The hypotheses presented here focus on how selective pressures during pregnancy may cause strong selection in response to changing diets in recent human evolution.

Nutrients and Selection during Pregnancy

Access to nutrients has been critical in human evolution, contingent upon dietary resources and the physiological processes that determine the bioavailability of ingested nutrients. Two selective pressures in humans that changed the amount and bioavailability of nutrients in the diet were exposure to solar UV radiation and adult milk-drinking. The ways in which these impacted fecundity and pregnancy may explain why UV radiation and milk-drinking exerted such strong fitness effects.

Skin pigmentation closely correlates with UV radiation worldwide (Nina G. Jablonski & Chaplin, 2000), perhaps partly because UV radiation exerted strong selection across populations during pregnancy in addition to other stages of life. Lighter or darker pigmentation impacts absorption of UV radiation on folate and vitamin D3, critical micronutrients during pregnancy (Nina G. Jablonski & Chaplin, 2000; Nina G Jablonski & Chaplin, 2010). Folate—obtained from eating plants—is stored in cutaneous blood vessels

and can be destroyed by UV radiation (Steindal, Tam, Lu, Juzeniene, & Moan, 2008). Folate deficiency causes failure of neural tubes to close during fetal development, resulting in anencephalus and spina bifida, defects lethal to the fetus (Fleming & Copp, 1998). Neural tube defects rarely occur in darkly pigmented people as their melanin protects their folate stores in equatorial areas (Nina G. Jablonski & Chaplin, 2000). Therefore, increased melanin production among equatorial populations of Africa, as well as Asia, Australia, and the Pacific where populations migrated, was potentially selected to protect folate stores in the skin during pregnancy.

On the other hand, melanin in the skin also blocks synthesis of vitamin D3 at higher latitudes (M F Holick, 1987). Vitamin D3 enables absorption of calcium for skeletal formation in the fetus and maintenance in the mother (Brunvand, Quigstad, Urdal, & Haug, 1996). Deficiencies cause malformation of the maternal pelvis, maternal osteoporosis, and rickets in fetuses and growing children (Fogelman, Rakover, & Luboshitzky, 1995; Henderson et al., 1987). In addition, vitamin D3 may assist development of the fetal innate immune system and critical organs (Michael F Holick, 2004; Norman, 2008). Therefore, balancing the synthesis of vitamin D3 with protection of folate-stores for pregnancy probably played a role in the strong selection for graded melanation with UV-radiation clines worldwide (Nina G. Jablonski & Chaplin, 2000; Nina G Jablonski & Chaplin, 2010).

Signatures of strong selection have been found in diverse populations surrounding genes with variants associated with skin pigmentation—notably *SLC24A5*, *MATP*, and *TYR* in Europeans, *DCT*, *EGFR*, and *DRD2* in East Asians, and *TYRP1*, *KITLG*, *ASIP* and *OCA2* in both populations (Alonso et al., 2008; Lao, de Gruijter, van Duijn, Navarro, & Kayser, 2007; Norton et al., 2007; Quillen et al., 2012). In addition, ancestral alleles of these genes that

tend to be associated with darker pigmentation and occur at a high frequency in Africans also tend to be highly frequent in darkly pigmented Melanesian populations. This may indicate convergent selection on the same genetic variants in diverse populations (Lao et al., 2007), although many populations remain to be tested.

Alternatively, UV radiation may have selected for appropriate skin pigmentation at other life stages such as childhood. Some detrimental effects of UV radiation on skin, such as skin cancer, occur post-reproductively, mitigating their importance to fitness (Blum, 1961; Nina G Jablonski & Chaplin, 2010). However, sun-burn alone causes significant morbidity for lightly pigmented people living in high UV regions because it damages the skin, increasing infection and water loss and decreasing thermoregulatory control. Furthermore, while vitamin D3 is critical for pregnancy, it is also important for bone density, immune function, and other effects in childhood and throughout life. To address this, one piece of evidence indicating that pregnancy, specifically, may have been important to selection on skin pigmentation is that women exhibit slightly lower levels of skin pigmentation on low exposure patches of skin than men, across world populations, indicating that the need for vitamin D3 may have been more critical to women than men (Nina G. Jablonski & Chaplin, 2000). Research clarifying the importance of vitamin D3 status to human health at different life stages could shed more light on this hypothesis.

Likewise, the ability to drink milk among pastoralists who keep dairy animals may also have been driven by selection on reproductive fitness. These pastoralists experienced strong selection in the past 10,000 years to continue digesting the lactose found in milk into adulthood, rather than losing this ability shortly after birth, as in most mammals (Ingram, Mulcare, Itan, Thomas, & Swallow, 2009). Strong selection has been detected

surrounding a number of different polymorphisms in diverse pastoralist populations from Europe, Africa, the Middle East, and Central Asia, each associated with regulation of *LCT* expression, encoding the enzyme lactase, which is responsible for cleaving lactose, the disaccharide in milk (Enattah et al., 2008; Heyer et al., 2011; Myles et al., 2005; Peng et al., 2012; Tishkoff et al., 2007a). Researchers have been surprised by the strength of this selection and have struggled to develop plausible explanations for it. Milk from animals provides an extra source of sugar, protein, fat, calcium, and hydration, beneficial not only for survival but also for reproduction.

Several possible hypotheses could link milk to reproductive fitness. First, milk from animals provided a safe source of hydration, especially for those living in hot, arid climates like Africa and the Middle East (Ingram et al., 2009). Considering the sensitivity of pregnant women to contaminated food and drink (Pouillot, Hoelzer, Jackson, Henao, & Silk, 2012), pregnant women able to drink sterile fresh milk may have experienced special fitness benefits. Second, the extra calcium in milk could be beneficial due to its role in skeletal development and maintenance and to female reproductive maturation, as large pelvises are required for vaginal delivery (Ellison, 1990). Third, because fat is more calorically dense than proteins and carbohydrates, fat from milk could help the mother nourish her infant during pregnancy and lactation. Fat stores and energy balance have also been linked to age of menarche and length of anovulatory period post pregnancy (Frisch, 1984; Panter-Brick, Lotstein, & Ellison, 1993).

A final hypothesis involves the fact that milk and other animal fats contain cholesterols used to synthesize reproductive hormones, critical for fecundity and early fetal development and growth (Herrera, 2002). The grain-based diets of Neolithic farmers were

lower in cholesterol than the diets of hunter-gatherer ancestors who consumed more wild game (Ströhle & Hahn, 2011). Less cholesterol in the diet correlates with lower levels of reproductive steroids (Goldin et al., 1982), reducing ovarian function and fecundity, suggesting that milk drinking could have provided a much-needed cholesterol and fertility boost for Neolithic Europeans. Therefore, the increase in fat, cholesterol, and calcium from drinking milk may have accelerated female skeletal maturation, increased caloric resources, and increased fecundity among women who could consume dairy, creating strong fitness benefits.

Oxygen and Selection during Pregnancy

Another environmental pressure detrimental to pregnant women is high-altitude hypoxia. When brought to high-altitudes, people from sea-level populations increase hemoglobin levels to carry more oxygen to the tissues. With long-term exposure and old age, increased hemoglobin causes altitude sickness and even death. However, pregnant women experience a special danger: preeclampsia caused by oxygen-restriction for the fetus. As described in the main text, preeclampsia often results in premature labor, small birth-weight babies, and hemorrhaging, seizures, and death for the mother (WHO, 2005).

Tibetans, Andeans, and the Ethiopian Amhara have each adapted to hypoxic high-altitude conditions possibly due to its impact on pregnancy. In these populations, strong signatures of selection surround genetic loci related to hypoxia and hemoglobin concentration, including *EGLN1, EPAS1*, *PPARA, THRB*, and *ARNT2* (Cynthia M Beall et al., 2010a; Bigham et al., 2010b; Laura B Scheinfeldt et al., 2012; Simonson et al., 2010b). However, Andeans are still at risk for altitude sickness in old age because they exhibit the

same elevated hemoglobin levels of low-landers at high-altitudes, indicating that selection for post-reproductive survival was not the primary force in this population (Mejía, Prchal, León-Velarde, Hurtado, & Stockton, 2005). Yet, some studies find that Andeans and Tibetans giving birth at high-altitudes have fewer instances of low-fetal birth weights and preeclampsia than low-landers at high-altitudes, possibly due to increased uterine capillary density (Cynthia M Beall, 2007; Moore, Zamudio, Zhuang, Sun, & Droma, 2001; M. J. Wilson et al., 2007). Also, some genes under selection among the Amhara are involved in fetal hemoglobin levels (*BCL11A*) and angiogenesis (*AIMP1* and *VAV3*), important features of pregnancy (Laura B Scheinfeldt et al., 2012). These pieces of evidence indicate that pressures during pregnancy may have been significant in adapting to high altitude hypoxia for Tibetans, Andeans, and the Amhara.

Infectious Disease and Selection during Pregnancy

Infectious diseases have exerted some of the strongest forces of selection on humans, most notably since the increase in population densities following the transition to agriculture and pastoralism 10,000 years ago. For example, genetic variants conferring resistance to malaria, such as alleles in the region of *HBB*, *HBA*, *FY*, *CD36*, *G6PD*, etc., were strongly selected among African populations and others where malaria is endemic (Campino, Kwiatkowski, & Dessein, 2006). Though infectious diseases are threats to survival generally, their differential impact on infants and pregnant women makes them especially powerful selective agents.

During pregnancy, the maternal immune system is suppressed so that the mother does not launch an adaptive immune response to the fetus's foreign cellular antigens

(Robinson & Klein, 2012). Though details are still being clarified, this response may make pregnant women less able to clear infections requiring strong inflammatory responses (Robinson & Klein, 2012). The outcome is that pregnant women experience spontaneous abortion and have higher morbidity and mortality in response to many infections than the general population (Robinson & Klein, 2012).

Malaria, influenza, and cholera are three infectious diseases that pose severe risks for pregnancy. In particular, African Plasmodium falciparum can infect the placenta (Robinson & Klein, 2012). As a result, pregnant women with malaria die 2-3 times more often than the general infected population (Shulman, 2003). In sub-Saharan Africa, malaria causes 20% of the cases of low infant birthweight, along with slow growth, spontaneous abortion, maternal anemia, and infant mortality (Robinson & Klein, 2012; Shulman, 2003; Steketee, Nahlen, Parise, & Menendez, 2001). Intriguingly, positive selection on a genetic variant of the gene *FLT1*, which reduces spontaneous abortions in cases of placental malaria, has been found for a malaria-endemic population in Tanzania (Muehlenbachs, Fried, Lachowitzer, Mutabingwa, & Duffy, 2008). This indicates that in the case of malaria resistance, selection mediated by pregnant women and their fetuses alone is sufficient for adaptive change in allele frequency in a population. Based upon this evidence, although genetic variants conferring general resistance to malaria experienced positive selection that could have been mediated by a broader subset of the population, pregnant women likely comprised an important portion of this selection.

During the 1918 influenza pandemic, ~50% of all infected pregnant women contracted pneumonia, and ~50% of this subset died (~27% total mortality for infected pregnant women), far more than the ~1% mortality for all individuals of reproductive age

with influenza (Harris, 1919; Taubenberger & Morens, 2006). Along with fetal abortion, this caused a 5-15% drop in birth rate the following spring (Bloom-Feshbach et al., 2011). This pattern is typical of other influenza pandemics (Pazos, Sperling, Moran, & Kraus, 2012). Mortality by influenza is heritable (Horby, Nguyen, Dunstan, & Baillie, 2012), so resistance to influenza may have been strongly selected for in recent human evolution, although this has been understudied.

Cholera causes diarrhea, vomiting, dehydration, and cramping, which can induce spontaneous abortion, preterm small birthweight babies, and maternal death (Carrera, 2007). Similar to influenza, smallpox, and dysentery, cholera decreases birthrates significantly during epidemic years (Hotelling & Hotelling, 1931; Woods, 2009), indicating it has strong potential as a selective agent in humans.

Many other infectious diseases are particularly dangerous for pregnant women. Among female Lassa Fever patients of childbearing years admitted to a hospital in Sierra Leone, death was significantly higher for pregnant women (25%) than non-pregnant women (13%) (Price, Fisher-Hoch, Craven, & McCormick, 1988). Tellingly, symptoms improved with delivery (Price et al., 1988). The Ebola virus killed more pregnant patients (95.5%) than the population average (77%) during an outbreak in the DRC (Mupapa et al., 1999). Some infectious agents, for example the parasite Toxoplasma gondii, cause disease only in pregnant women, who are likely to experience abortion (Robinson & Klein, 2012). Evidence from mice suggests that another parasite, Leishmania, also exploits immunological changes in pregnant women (C. Roberts & Walker, 2001). Finally, Varicella zoster, the chicken pox virus, causes pregnant women to develop more skin lesions and pneumonia at higher rates than the average adult with chicken pox (Harger et al., 2002).

Pregnant women are clearly especially vulnerable to infectious disease. Although many of these diseases also cause significant morbidity in non-pregnant adults, the dramatic impact on pregnant women makes it likely that selective effects would have been strongly mediated by this population, though the adaptive benefit of genetic resistance to infectious disease is felt across all life stages for both males and females. As researchers discover functional genetic variants in areas under selection in the human genome, we predict that many are likely to confer resistance to infectious diseases that severely impact pregnant women who lack resistance in addition to those causing high infant mortality.

Alternative Hypotheses and Avenues of Research

Although the evidence described in this chapter support the importance of pregnancy to recent selection in humans, alternative hypotheses could also explain some phenomena that we argue suggest selection in pregnancy. Take, for example, the differences in GDM prevalence across populations, and the inverse correlation with historical glycemic intake. When mothers born in energy-poor environments emigrate to energy-rich environments, fetal programming may contribute to the pattern as these women have heightened risk of GDM and type 2 diabetes (Hales & Barker, 2001). Maternal epigenetic modifications could be the mechanism underlying this programming to suit the early life environment. Another contributor could be the differences in patterns of adipose storage across populations—Asian women tend to have more central adiposity than women in other populations, and this is thought to increase insulin resistance (Raji, Seely, Arky, & Simonson, 2001). However, this proximate cause of increased GDM among Asians is not at odds with a history of natural selection acting on the trait.

Distinguishing among these competing explanations for the patterns we see could be a fruitful line of research. For example, first, one could conduct association studies in diverse ethnic populations to identify genetic loci linked to GDM risk. Second, these loci associated with GDM could be analyzed for signatures of recent selection in order to test whether selection has influenced GDM incidence across populations. Finally, one could test whether incidence of GDM among immigrants approaches that of the rest of the population across generations. GDM is reduced for South Asians born in the US compared to first generation immigrants, but it is still elevated above the level of European-Americans (Savitz et al., 2008), indicating fetal programming may explain a large fraction of differences in GDM risk, but is probably not the only factor.

Similar approaches could be used to test hypotheses of selection for resistance to preeclampsia, infectious disease, hypoxia and other reproductive factors. In a broad sense, this will require a better understanding of the axes of human variation—genetic and phenotypic. Next-generation sequencing data from diverse populations of humans will contribute to this understanding. However, the phenotypic data is just as critical. We need a clearer understanding of the susceptibility of pregnant women to infectious diseases and metabolic diseases across populations, and how this is mediated by nutritional status, UV radiation, hypoxia, and other external factors. Testing these hypotheses will be important both for evolutionary genetics and for improving care for human health across diverse ethnicities.

Conclusions

The field of human evolutionary genomics is in a period of transition. Currently, only a few examples of selection in response to environmental pressures felt by particular populations have been elucidated —malaria resistance, lactase persistence, etc. These examples were already under study prior to the development of evolutionary genomics, and the signatures of selection surrounding the genetic variants under selection merely served to substantiate strong adaptive hypotheses already presented. However, next-generation sequencing data, conducted in diverse populations, now provides the raw material to detect many more strong candidates for selection. Thus, the field of evolutionary genomics now has the potential to provide many new testable hypotheses of selection, which were not developed a priori. For example, a catalog of candidate variants for selection was recently published and one of these variants was experimentally characterized (Grossman et al., 2013).

At this turning point in the field, we seek to underscore that many aspects of human evolution are best understood by investigating the life-history bottleneck of pregnancy and birth from the perspective of both the mother and the infant. During pregnancy, nutritional, energetic, physical, and immunological requirements are constrained in the mother to support the fetus, concentrating selective forces upon the mother at a sensitive life stage. The pressures that have been most important in recent human evolution—infectious diseases from high population densities, adult dairy consumption from pastoralism, grain consumption from agriculture, changes in UV radiation and oxygen levels from moving to extreme latitudes and altitudes—have left genetic signatures of their selective impact. Although these selective factors may be felt across the lifespan, nowhere are they more

serious than during infancy and pregnancy. We should thus remain cognizant of these phases of life as next-generation sequencing now provides evolutionary genomicists with the data to generate many new testable hypotheses of why loci are under selection in humans.

Final Comments

This article was published in 2013 in the journal of *Trends in Genetics* (E. A. Brown, Ruvolo, & Sabeti, 2013). I wrote this perspective article under the supervision of Pardis Sabeti and Maryellen Ruvolo. Since the publication of this perspective, new research has been conducted on the genetics of gestational diabetes and preeclampsia, though large, well-powered studies remain to be done. A GWAS of gestational diabetes in European, Thai, and Afro-Caribbean women was conducted (Hayes et al., 2013). A GWAS of preeclampsia in European, Afro-Caribbean, and Hispanic women was also conducted (Zhao, Bracken, & DeWan, 2013). Two variants associated with preeclampsia in European women were replicated in Han Chinese women (J.-P. Wan et al., 2014; Ji-peng Wan et al., 2013), and expression levels of several candidate genes for preeclampsia were found to correlate with preeclampsia (Yong et al., 2014).

CHAPTER 4

IVD EXPRESSION AND LEUCINE ADAPTATION IN EAST ASIANS

ABSTRACT

This project uses new genomic tools, as well as computational and functional analysis, to develop a novel hypothesis of adaptation in the East Asian population. Combining the results of a high-power test for selection, the Composite of Multiple Signals (CMS) test (Figure 1) with a high-coverage dataset of expression quantitative trait loci (eQTLs) (Lappalainen et al., 2013) yields regulatory candidates for selection. Among these candidate loci are those associated with genes of known metabolic function, including the gene isovaleryl dehydrogenase (*IVD*), which lies within the peak of a CMS selected region on East Asian haplotypes (Figure 4) (Grossman et al., 2013). Functional and computational analyses reveal that multiple derived alleles on the same haplotype in Asian populations drive increased expression of *IVD* and increased efficiency of leucine catabolism. These genetic regulators of *IVD* and their relationship to phenotype and selection in Asian populations form the foundation of this investigation into a recent metabolic adaptation in modern human populations.

INTRODUCTION

Research Strategy: Selecting Candidate Adaptive Loci

The most important dimension of generating new hypotheses of adaptation is the identification of strong candidate loci, followed by functional and computational validation. This research employs a specific strategy to hone in on strong candidate loci for local

***IVD* eQTLs lie in the CMS peak of a selected region in East Asians.**

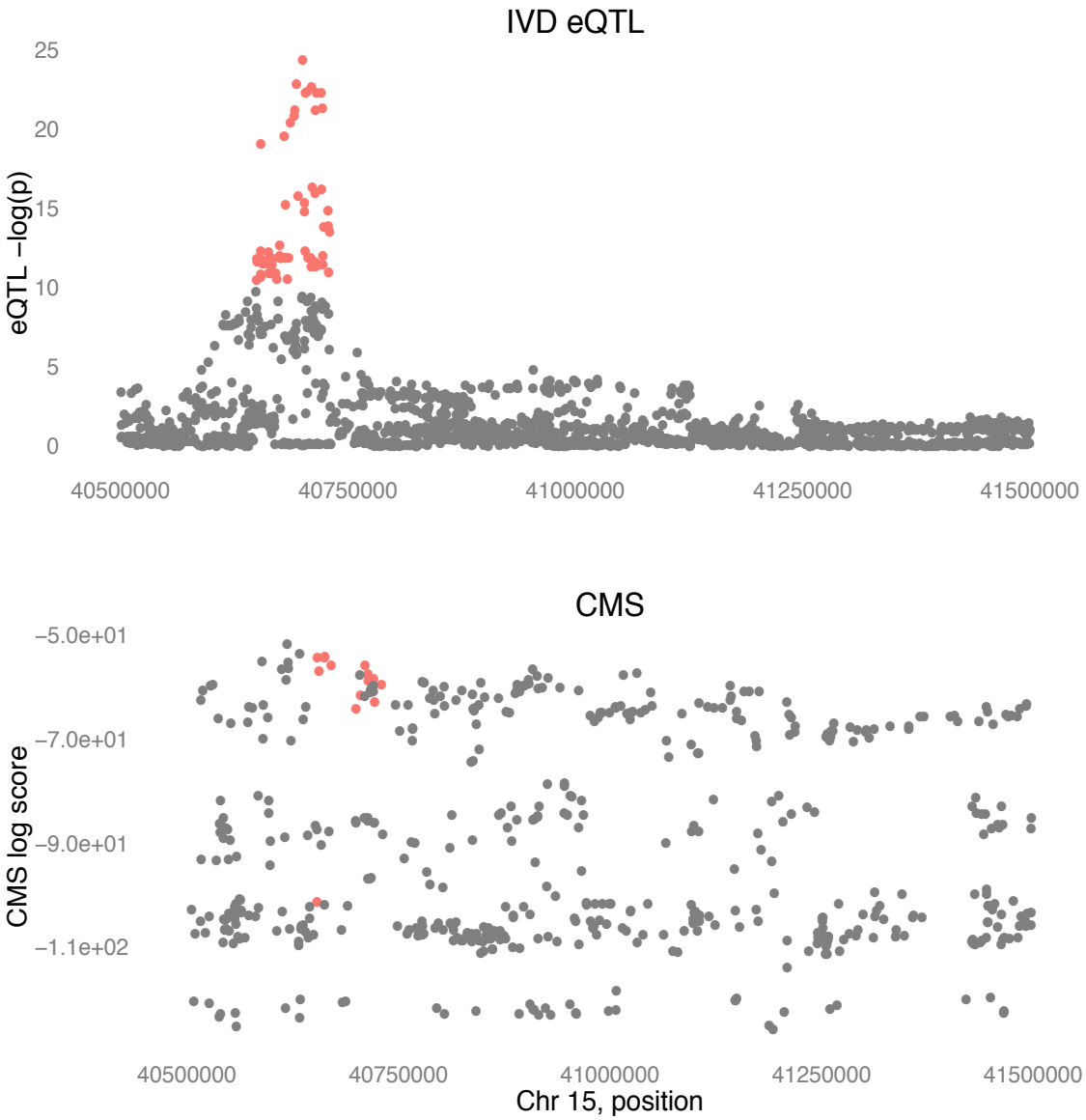


Figure 4. Variants in red that strongly associate with *IVD* expression in lymphoblastoid cell lines, plotted as the negative log of the p-value of association, are also strong candidates for selection in the region by the CMS test, plotted as the log of the CMS score.

adaptation in human evolution. Then, cell culture analyses are conducted to validate the regulatory function of candidate loci. Lastly, the relationship of these loci to selection in the region is analyzed by examining allele frequency and age, test statistics for selection, and haplotype structure in diverse populations. The product of this research is a novel hypothesis of dietary adaptation, which may be tested with more in depth laboratory assays and population-based studies.

This research begins by leveraging new genomic tools for studying selection to maximize the probability of selecting good candidates. First, regions under selection are drawn from a CMS analysis conducted on 1000 genomes resequencing data for Northern Europeans (EUR), East Asians (ESN), and Yoruba (YRI). These regions contain the strongest candidate loci for selection using the most high-coverage genetic maps available for these populations. Second, eQTLs lying within these regions are drawn from the Geuvadis dataset, which is the most high-powered RNA-seq study published, conducted on lymphoblastoid cell lines (LCLs) from Northern European and Yoruban individuals in the 1000 genomes project. Third, regions under selection containing eQTLs are paired down to those associated with genes of clear metabolic function. This is done because metabolism has been an important substrate for selection in recent human evolution, and narrowing candidates down to those of known function enables the generation of adaptive hypotheses. This process yields strong candidate adaptive loci, including the eQTLs for a gene called isovaleryl-CoA dehydrogenase, or *IVD* (gene function described below), in a region under selection in East Asians. The derived eQTL alleles at increased frequency in East Asians are associated with increased expression of *IVD*.

Isovaleryl Dehydrogenase and Leucine Metabolism

The gene, *IVD*, encodes a mitochondrial-matrix enzyme that catalyzes the third step in the digestion of leucine, an essential, branched-chain amino acid, converting isovaleryl-coenzyme A (CoA) to 3-methylcrotonyl-CoA (Figure 5) (Ikeda & Tanaka, 1983). Coding mutations in *IVD* cause acidosis through the build-up of alternative acidic products of isovaleryl-CoA in the blood and urine (Tanaka, 1990). Isovaleric acidemia results in vomiting, diarrhea, lethargy, seizures, neurotoxicity, and death (Newman, Wilson, Callaghan, & Young, 1967; Tanaka, 1990; Tanaka & Budd, 1966). Most severe cases manifest shortly after birth in infancy, but some milder cases manifest later in life, and may be triggered by fasting, illness, physical exertion, or heavy consumption of dietary protein, all states that increase protein catabolism (Collins, Umpleby, Boroujerdi, Leonard, & Sonksen, 1987; Feinstein & O'Brien, 2003; Millington, Roe, Maltby, & Inoue, 1987). While isovaleric acidemia was the first acidemia discovered in humans, mutations in other enzymes along the leucine-catabolism pathway also cause build-up of acidic intermediates, and result in similar suites of symptoms (Holzinger et al., 2001; Lehnert, Scharf, & Wendel, 1985).

Understanding *IVD* function and expression requires a thorough understanding of leucine's role in the diet, and its role in building proteins, producing energy, and regulating metabolic pathways. Many protein-rich foods, such as eggs, soy, and dairy, are plentiful sources of leucine (Figure 6), which makes up over 20% of all human protein consumption, the most of any amino acid (Li, Yin, Tan, Kong, & Wu, 2011). Leucine is a common building-block for many human proteins, but cannot be synthesized by vertebrates, hence the heavy consumption by humans. Furthermore, when intracellular levels of leucine are reached

Leucine catabolism pathway and signaling effects

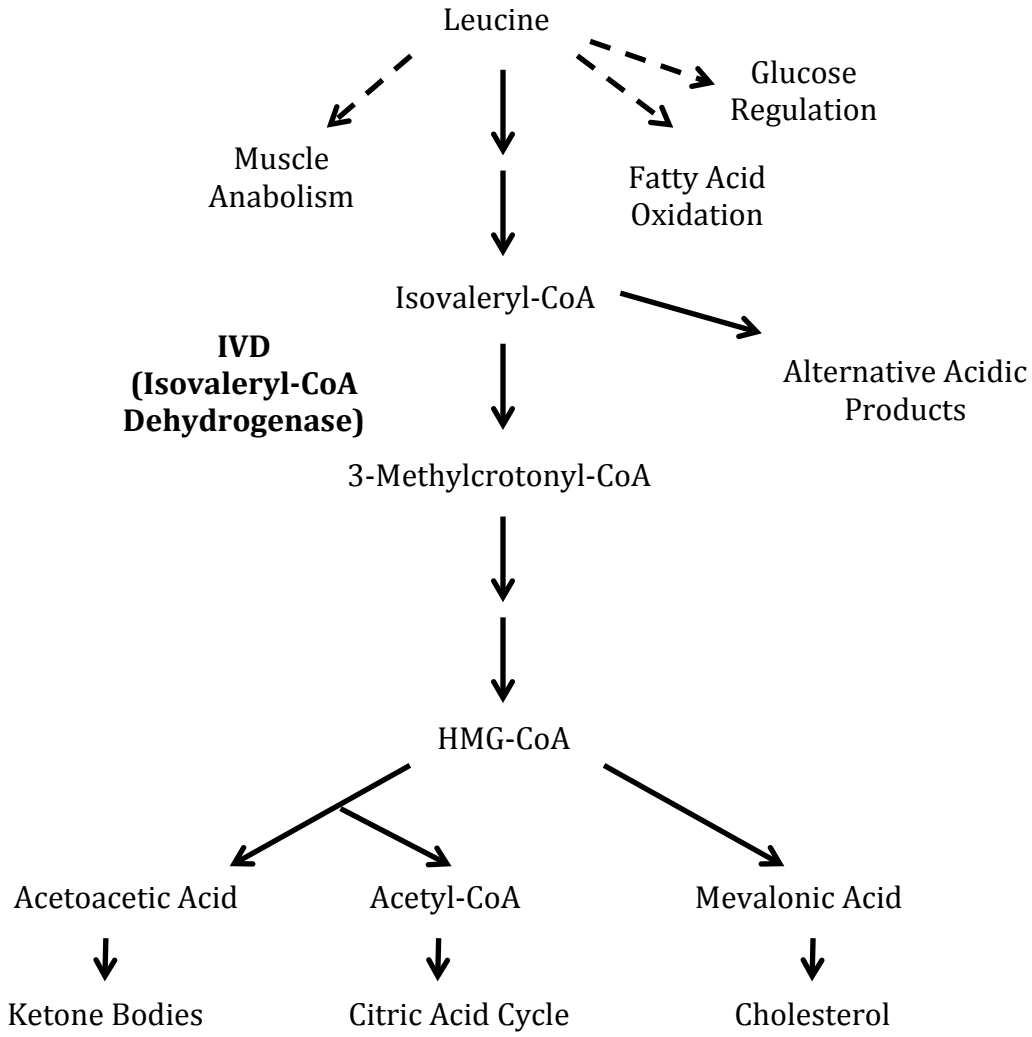


Figure 5. Leucine undergoes catabolism to become HMG-CoA, which can be used for ATP synthesis, ketone body synthesis, or cholesterol synthesis. The IVD enzyme catalyzes the third step in leucine catabolism, converting isovaleryl-CoA to 3-methylcrotonyl-CoA. Leucine also stimulates muscle anabolism and fatty acid oxidation through the mTOR-pathway, and insulin release from the pancreas.

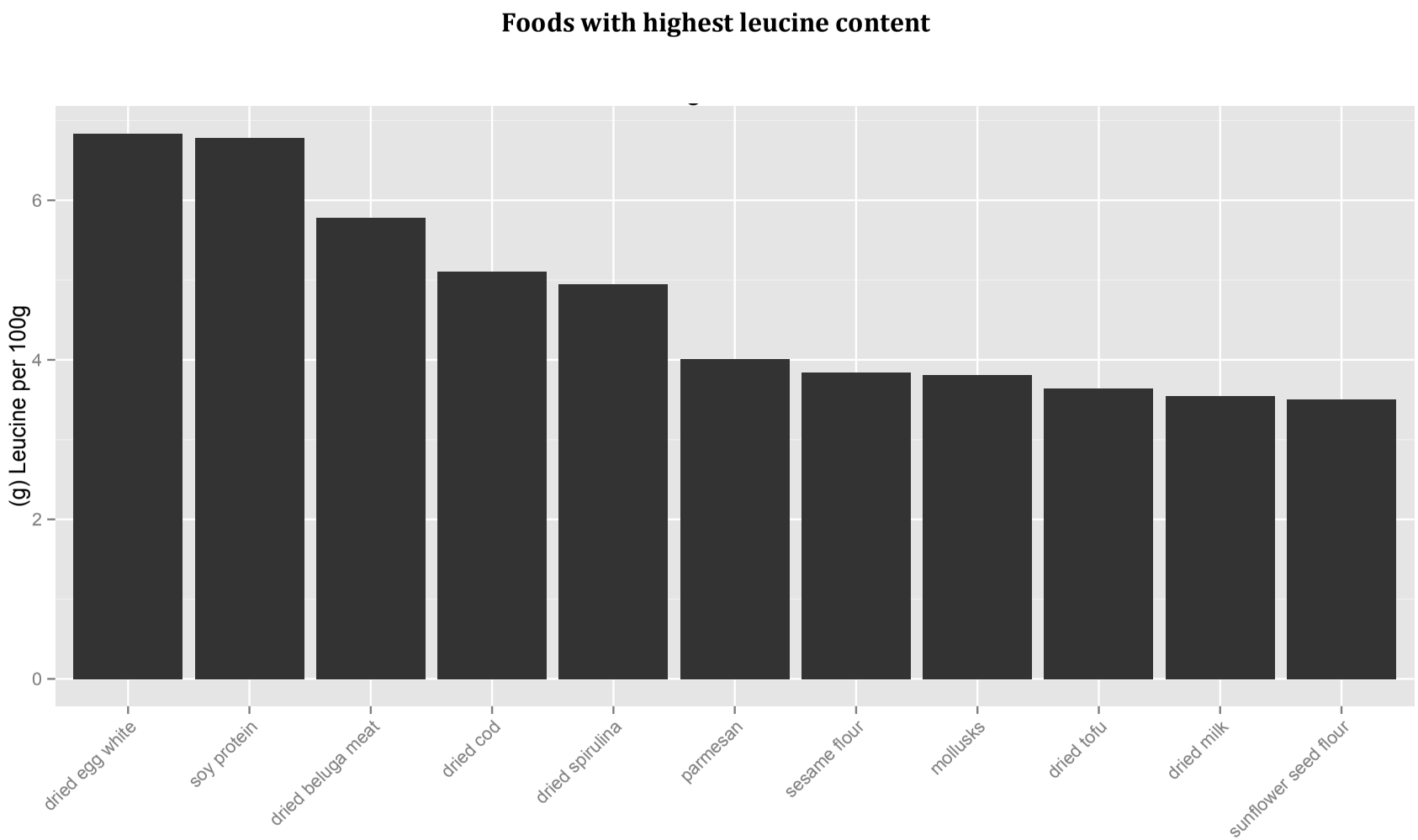


Figure 6. The foods with highest leucine content from the United States Department of Agriculture, plotted in grams of leucine per 100 grams of food.

above that necessary merely for protein synthesis, it feeds into important catabolic and signaling pathways (Ahlborg, Felig, Hagenfeldt, Hendler, & Wahren, 1974; Anthony et al., 2000).

Along the catabolic pathway of leucine, it is first oxidized into an organic acid, 2-oxo-isocaproic acid, and maintains this acidic state through several subsequent steps of catabolism (Figure 5). As described above, mutations that disrupt the functioning of the enzymes that mediate these steps of leucine degradation cause acidosis. After five steps of leucine catabolism, it is converted into 3-hydroxy-3-methyglutaryl-CoA (HMG-CoA), which is a central component of many human metabolic pathways.

HMG-CoA, the product of leucine catabolism, is shuttled into different pathways depending on the energetic and molecular state of the cell. It can be broken down into acetyl-CoA, which fuels the citric acid cycle and ATP-energy production in the matrix membrane of mitochondria (Lopes-Cardozo et al., 1975). Alternatively, HMG-CoA can be converted to acetoacetic acid, which is used by mitochondria to synthesize ketone bodies—ketogenesis—in the liver (25% of all ketogenesis), skeletal muscle, and adipose tissue (Katz & Bergman, 1969; Lopes-Cardozo et al., 1975). Ketogenesis typically occurs in conditions of hypoglycemia, in order to provide an alternative fuel to glucose for use in the brain and other tissues. However, ketogenesis also occurs in situations of hyperglycemia, as in diabetes, in order to store excess blood glucose, which is toxic to tissues. Lastly, HMG-CoA can be reduced into mevalonic acid, which is converted in the cytosol into a variety of sterols, principally cholesterol (Panini, Schnitzer-Polokoff, Spencer, & Sinensky, 1989). In vertebrates, liver cells, again, are the largest producers of cholesterol (Ott & Lachance,

1981). Cholesterol acts as a negative feedback inhibitor of the enzyme HMG-CoA reductase, thus controlling excess cholesterol production (Panini et al., 1989).

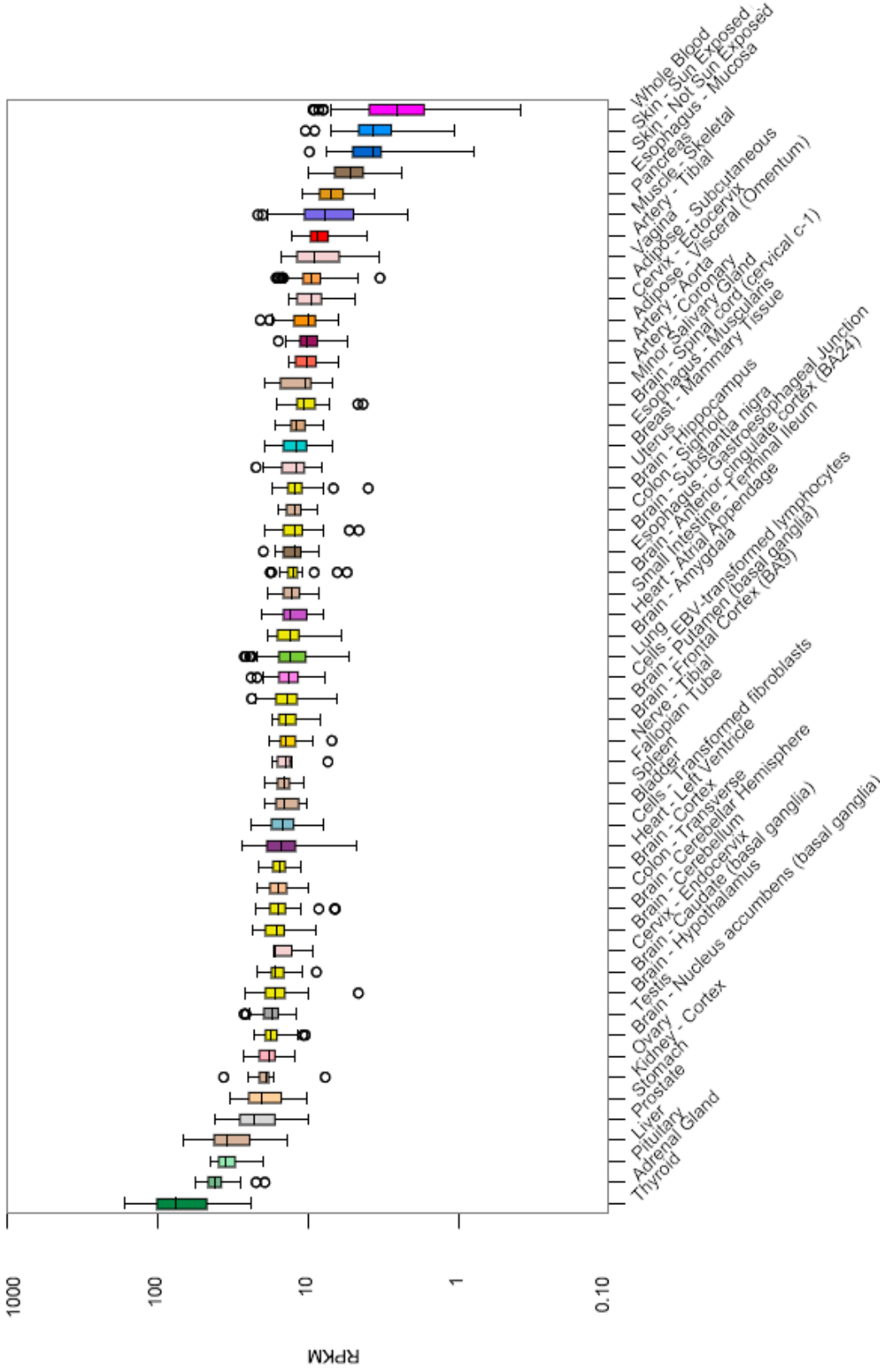
IVD expression mirrors the roles leucine plays in the body. As leucine is such a critical component of protein synthesis, *IVD* is expressed moderately to highly in most tissues (Figure 7). Reflecting the importance of leucine catabolism to cell metabolism, *IVD* is expressed at high levels in metabolic tissues, principally liver, but also stomach and kidney. Interestingly, rats on a fat-free diet increase *IVD* expression in heart, indicating a fat-free diet may upregulate catabolism of leucine for ATP synthesis (Nagao, Parimoo, & Tanaka, 1993). Similarly, a high-fat diet decreased expression of *IVD* in the heart, but expression in liver was unaffected. Expression is also high in several brain tissues, perhaps because ketone bodies are a critical energy resource for the brain during glucose scarcity, and ketogenesis does occur in brain cells, such as astrocytes (Guzmán & Blázquez, 2004). But, strikingly, *IVD* is most highly expressed in the thyroid, adrenal, and pituitary glands, possibly reflecting leucine's importance as a signaling molecule. For example, estrogen injection into the inguinal region of ovariectomized rats reduced *IVD* expression in the pituitary (Blake, Brown, Duncan, Hunsucker, & Helmke, 2005). Leucine has important signaling roles in adipocytes, myocytes, and pancreatic β-cells, where *IVD* is expressed at moderate levels.

One signaling function of leucine is activating the mammalian target of rapamycin- (mTOR-) pathway in adipose tissue and skeletal muscle (Li et al., 2011). This pathway instigates protein translation, mitochondrial biogenesis, and fatty acid oxidation (Duan et al., 2015). Specifically, leucine shifts energy from adipocytes to myocytes by inhibiting expression of fatty acid synthase and peroxisome proliferator-activator gamma (*PPARG*),

Figure 7. *IVD* expression across tissues

IVD expression as measured by RNA-seq in diverse tissue types in the gTEX database, plotted by log transformed values of RPKM, Reads Per Kilobase of transcript per Million mapped reads.

Figure 7 (Continued)



which promote adipocyte differentiation and lipogenesis, and activating mitochondrial biogenesis and fatty acid oxidation to fuel protein anabolism (Sun & Zemel, 2007). In both rats and humans, leucine supplementation reduces body fat and reduces muscle loss with age (Yang, Chi, Burkhardt, Guan, & Wolf, 2010). In mice fed a high-fat diet, leucine supplementation attenuates body weight gain and prevents hyperglycemia (Y. Zhang et al., 2007). Leucine may stimulate this mTOR protein translation pathway by activating both ribosomal protein S6 kinase 1(S6K1) and eukaryotic initiation factor 4E binding protein 1 (4E-BP1) (Lynch, 2001; Tsukiyama-Kohara et al., 2001).

Another related pathway stimulated by leucine is insulin release from the pancreas. This occurs partially through the mTOR-pathway, believed to stimulate insulin production in pancreatic β-cells (Yang et al., 2010). In addition, leucine is an allosteric activator of glutamate dehydrogenase (Gylfe, 1976), which produces ATP that in turn stimulates insulin production. Finally, leucine and its first metabolic product 2-oxo-isocaproic acid are both believed to inhibit ATP-dependent potassium channel currents, thus increasing Ca^{2+} concentrations in the cells, stimulating insulin release (Henquin et al., 1994). By increasing insulin production from pancreatic islets, leucine aids glycemic control in both humans with diabetes and diabetic and obese rodent models (Guo, Yu, Hou, & Zhang, 2010; Li et al., 2011). However, given the increased insulin levels, evidence from rodent models is also conflicted about whether leucine aids or reduces insulin sensitivity. In addition, increased insulin production by pancreatic β-cells may also contribute to protein anabolism as insulin binding to myocyte receptors may promote amino acid uptake.

Surprisingly, the benefit of dietary leucine to controlling blood glucose levels seems to reverse in pregnancy. Leucine supplementation of pregnant women results in higher

birth weight babies and also prevents fetal growth defects (Roos et al., 2007; Suryawan et al., 2008). However, supplementing pregnant rats with leucine inhibits fetal pancreatic β-cell development, which reduces blood insulin levels and increases blood glucose levels in fetuses (Rachdi, Aïello, Duvillié, & Scharfmann, 2012). Furthermore, these rat pups are larger at birth with impaired glucose tolerance after four weeks of life, and pancreatic β-cell number does not recover to normal levels during infancy despite β-cell's continued proliferation throughout infancy (Rachdi et al., 2012). Evidence suggests this discrepancy between the impact of leucine on glucose control through most of life, and the impact on fetal pancreatic development may be mediated by the mTOR-pathway: MicroRNA-7 has been shown to have similar discrepancies during pregnancy, and also acts through the mTOR-pathway (You Wang, Liu, Liu, Naji, & Stoffers, 2013).

The importance of leucine signaling and catabolism in pregnancy is also reflected in *IVD* expression. *IVD* expression increases 3.7-fold in the pancreatic β-cells of pregnant mice compared to before pregnancy, as measured by RNA-seq (Kim et al., 2010) and microarray experiments (Layden et al., 2010; Rieck et al., 2009). Upregulation of leucine catabolism in β-cells during pregnancy may attenuate over-stimulation of mTOR during pregnancy to avoid injuring fetal pancreatic development and glycemic control. On the other hand, feeding rats low protein diets during pregnancy reduces mTOR signaling and also induces hyperglycemia in adult offspring (Alejandro et al., 2014), indicating that adequate exposure to leucine *in utero* may be critical for normal metabolic development.

Genetic regulation of IVD *and phenotype*

IVD's function in leucine catabolism, and leucine's many roles as a building block for protein, a source of energy and cholesterol, and a stimulator of protein synthesis and

insulin production, are important when considering the relationship of genetic variation in *IVD* to phenotype. Mild, sub-clinical cases of isovaleric acidemia have been noted (Vockley & Ensenauer, 2006) indicating that epistatic or environmental factors, such as diet, impact the phenotype caused by *IVD* dysfunction, though such cases have not been well-studied. However, variation in *IVD* enzymatic activity does alter phenotype at the cellular level. Two large GWAS of blood metabolites in the British population detected variants in the *IVD* region that associate strongly with plasma levels of isovaleryl-CoA (normalized to propionyl-CoA, a metabolite of isoleucine and valine catabolism) (Shin et al., 2014; Suhre et al., 2011). These variants overlap with *IVD*'s eQTL peak of association in European individuals (Lappalainen et al., 2013) (Figure 8). Therefore, genetic regulators of *IVD* expression modify the efficiency of leucine catabolism, with potential ramifications for leucine as a source of energy for the cell or as a signaling molecule.

Hypothesis

The goal in this research is to find likely candidates for metabolic adaptation in diverse human populations. Overlapping the CMS regions containing high-probability candidates for strong, positive selection with eQTLs detected in lymphoblastoid cell lines (LCLs) yields a region containing multiple eQTLs for expression of isovaleryl-CoA dehydrogenase (*IVD*) within the top 100 CMS ranked SNPs for selection in region262 in East Asians. As the derived variants of these functional candidates are associated with increased expression of *IVD*, and are at increased frequency in East Asians, the hypothesis is that one or more enhances *IVD* expression and activity and was selected in East Asia for increasing efficiency of leucine catabolism.

***IVD* eQTLs overlap isovaleryl-CoA GWAS hits**

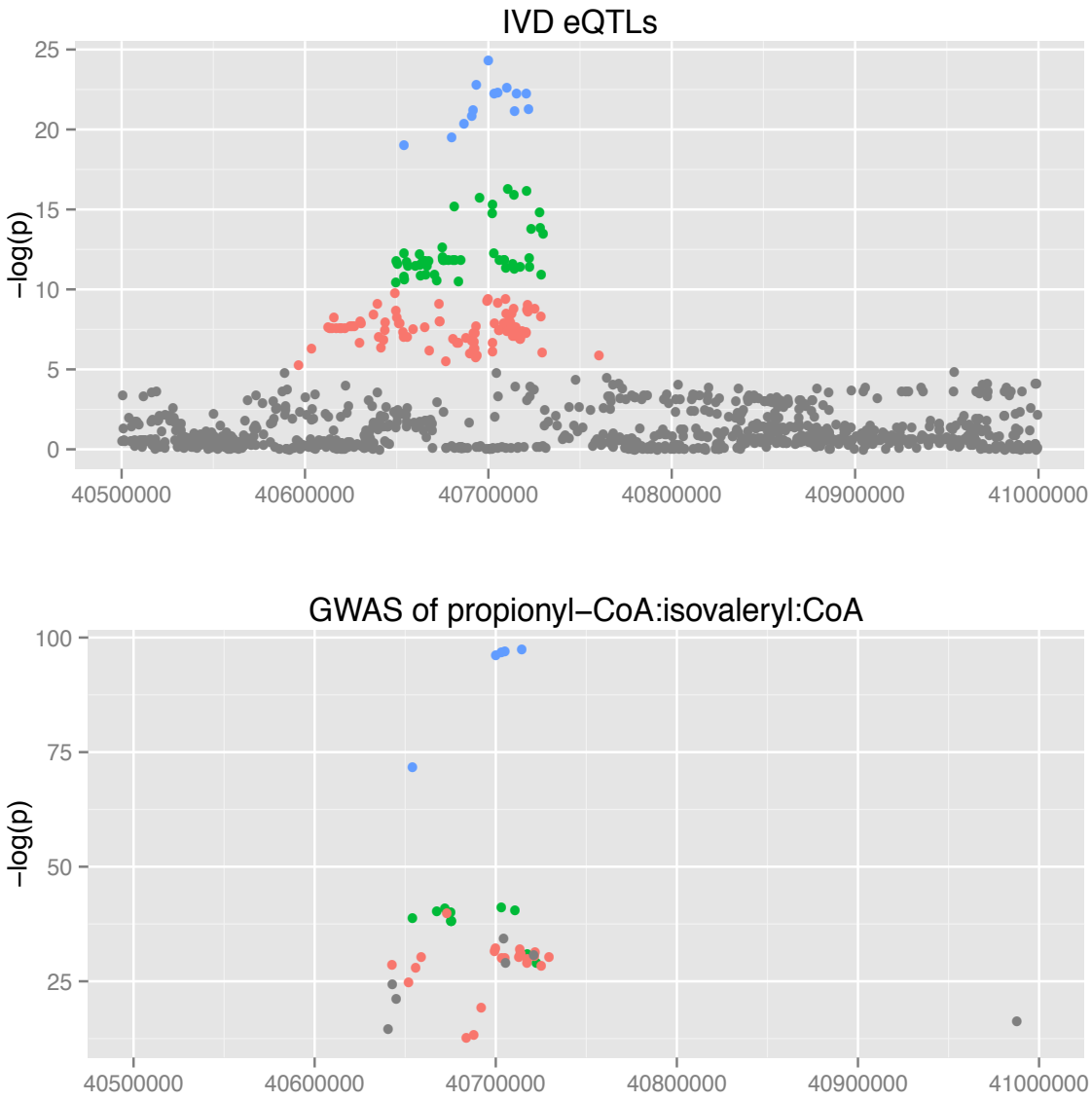


Figure 8. *IVD* eQTLs (Lappalainen *et al.* 2013), plotted in blue, green, and red for strength of association with *IVD* expression also associate with isovaleryl-CoA levels (normalized to propionyl-CoA) in a GWAS for blood metabolite levels (Shin *et al.* 2014), plotted in the same colors as above.

The hypothesis is tested by 1) luciferase assays to test enhancer activity of ancestral and derived candidate alleles, and 2) computational analyses of the haplotypes surrounding functional variants. Functional and computational interrogation of candidates shows that East Asian populations have an increased prevalence of variants for more efficient leucine catabolism. Furthermore, these variants tag a positively selected haplotype in these populations. This research opens new avenues to pursue the biological impact of a novel dietary adaptation in humans.

MATERIALS AND METHODS

Selection of Candidate Variants and Regions

Four candidate variants (rs2075624, rs12593066, rs11633883, rs17733719) were selected for testing based upon their association with *IVD* expression in the Geuvadis dataset and falling within the top 25 ranked SNPs for region 262 in CMS for East Asian haplotypes. In addition, six other candidate variants were selected for testing based on their linkage ($r^2 > 0.9$ in the 1000 genomes GBR population) to the top ranked eQTL for *IVD* from the Geuvadis dataset, rs10518693, detected in the European populations (Table 5). Regions surrounding candidate variants were defined using ENCODE data for histone acetylation marks and DNase I hypersensitivity regions, down to a minimum of 0.7kb (Figure 9). In the cases of rs10518693 and rs17733719-rs8033249 (one region), region sizes were curtailed to exclude repetitive Alu elements that caused recombination during the cloning process and interfered with cloning efficiency. However, in these cases the regions tested were ultimately still 1kb and 3kb in length.

Table 5. *IVD* eQTL candidates for selection

Variant	Position (hg19)	eQTL -log10	CMS rank	r^2 with top eQTL rs10518693 (GBR)
rs2075624	40710723	16.3	8	0.458
rs11633883	40714401	21.2	23	0.977
rs12593066	40714019	15.9	15	0.458
rs17733719	40720786	22.3	20	0.933
rs10518693	40700022	24.3	91	-
rs9635324	40703211	22.2	-	1
rs2289329	40705159	22.3	57	1
rs11638033	40710108	22.6	-	1
rs12902310	40715427	22.2	-	0.977
rs8033249	40721939	21.3	68	0.912

Regions cloned for candidate eQTLs

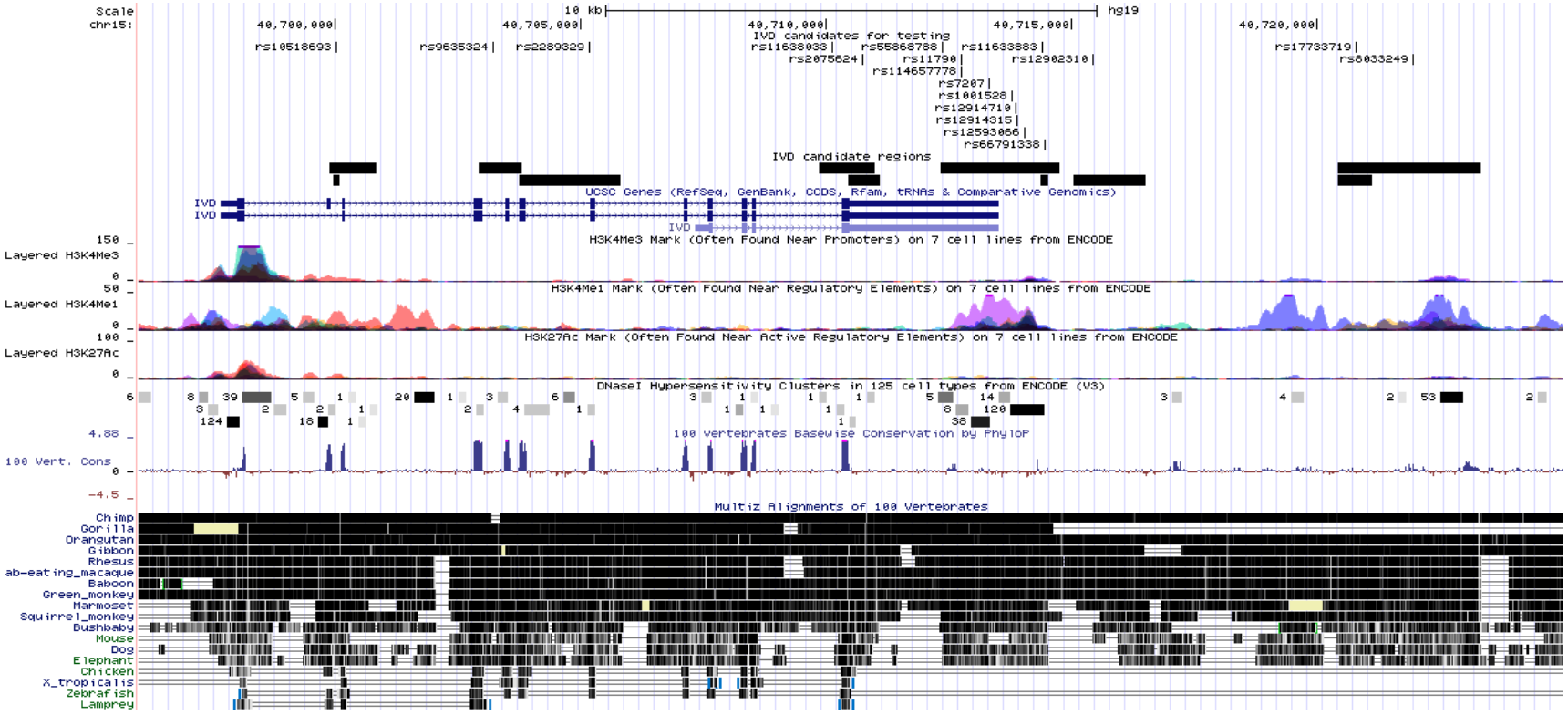


Figure 9. Regions surrounding tested candidate variants shown in black in the UCSC genome browser along with histone methylation and acetylation marks, DNaseI hypersensitivity clusters, and conservation across vertebrate genomes.

Cloning of candidate regions

Primers to amplify candidate regions were designed in Primer3 plus (Table 6). These regions were amplified by PCR from the genomic DNA of Han Chinese individual NA18576 (Coriell Biorepository) in 4 x 50µL PCR reactions each with Phusion HF polymerase in 2x GC-buffer, using primers containing 5'-SfiI restriction enzyme cut sites. PCR products were column purified using the Qiagen QIAquick PCR purification kit, or in the presence of smaller non-specific bands, products were gel purified from 0.7% agarose gels using the Qiagen gel purification kit. Then, SfiI was used for restriction digestion according to the NEB suggested protocol, and ligation into gel-purified, SfiI-digested pgl4.23 firefly luciferase vector was set up using T4 ligase and buffer according to NEB protocol. Finally, high-efficiency transformation of ligation products into NEB10β cells was conducted according to NEB instructions. 100µL of cells were plated onto LB plates containing 100ng/µL Amp, and colonies grew overnight at 37°C. The next day colonies were tested for presence of insert by boil prep PCR in 25µL reactions with *Taq* polymerase using primers aligned to the vector backbone. Colonies containing inserts of appropriate size were sequenced with Sanger sequencing at Eton Bio Labs to verify presence and fidelity of insert, as well as ancestral versus derived state of the candidate variant. Colonies containing correct constructs were selected into 50mL LB broth with 100ng/µL Amp, and grown overnight, shaking at 37°C for 12-16 hours. Plasmid construct DNA was purified from these colonies using the Qiagen Plasmid *Plus* DNA purification kit. DNA concentration was tested using the Nanodrop 2000 instrument, and when necessary DNA was concentrated to >1µg/µL by spinning on the SpeedVac.

Table 6. Primers used in study

Target	Forward	Reverse	Region Size
rs2075624	ATCGGCAGAGCCTTCAATGCA	GTGCTGGGATTACAAGCATGA	657
rs12593066 rs11633883	GGTCACACTATCAGGGCAGCTC	GCCTATCCTCTATTACCATC	2416
rs17733719	CTGCTGCTTGAGACTGCTTG	CACACAATGTGAGCTCTGAGG	694
rs10518693	CAAGGCCCAGGAGATCGATC	CCAGTCTGGGCCTCTTTCAG	967
rs10518693	AGTCGGGGGCAGTCAGG	CACATGTGACTGGCTGCCTT	150
rs9635324	GCATCAACCAGCTTGTACGC	GATCACTAATGGCCCTGATGCTG	883
rs2289329	AGCCAACATCCTGCCCTTAG	CTGACATGGTGGAGCACTGT	2078
rs11638033	ATGCTATGTGGGCTACCTGC	GAAATCTGAAACACTTGTGGTTCCA	1134
rs12902310	AATTGATCCACCGGCCTCAG	CTTCTGCTGCTTCCCTCCTTC	1469
rs11733719 rs8033249	CTAGGAGAGTCAGCTGGGGTC	AGTGGAGCCAGGGTTTACAT	2902

Site Directed Mutagenesis Primers			
Target	**Forward**	**Reverse**	**Conversion Type**
rs5586878 - to G	GCTGCTGAGTCCGCAGGGGGGGCAGAGCAGAGGACAGCGTGC	GCACGCTGTCCTCTGCTCTGCCCCCCCTGCGGACTCAGCAGC	Polymorphism
rs11790 C to T	GGCCCAGGTCCTATTCCTGTCCTCCAGGCCG	CGGCCTGGAGGACAGGAATAGGACCTGGGCC	Polymorphism
rs114657778 G to C	CCCGTCCTCCAGCCCGTTCTTTCATGAGGG	CCCTCATGAAAGAACGGGCTGGAGGACGGG	Polymorphism
rs7207 C to T	CCTGCTATGTTGGAGATGAATGTGACTAAAAGGGCCATC	GATGGCCCTTTTAGTCACATTCATCTCCAACATAGCAGG	Polymorphism
rs1001528 G to A	GAGCACACCCCTTATCAAAATTTGGCAACCTAG	CTAGGTTGCCAAATTTTGATAAGGGGTGTGCTC	Polymorphism

Table 6 (Continued)

rs12914710 A to G	CCCCAAGGGTGAGTCTGC AAGGCAATC	GATTGCCTTGCAGACTCACCC TTGGGG	Polymorphism
rs12914315 G to A	CAACCACCATCTGAGTCC TAAAGCAGGCTTCCCC	GGGGAAGCCTGCTTTAGGACT CAGATGGTGGTTG	Polymorphism
rs12593066 T to C	GGAGGCAAGGTTTCGCG CAGGGCCCCC	GGGGGCCCTGCGCGAAACCTT GCCTCC	Polymorphism
rs11633883 G to A	GGGATTTCAGGACTCAA CGACGTTTTGTTTTAGCC	GGCTAAAACAAAACGTCGTTG AGTCCTGAAATCCC	Polymorphism
rs66791338 - to GAAAG	GCCAATAAGGACATGAA AGGGAAGAGGGGTTGGG GGAAGCC	GGCTTCCCCCAACCCCTCTTC CCTTTCATGTCCTTATTGGC	Polymorphism
rs66791338 region	CATTTCTCTGGCCTAACT GGCCGTGACCACCCACAG CTGAC	GTCAGCTGTGGGTGGTCACGG CCAGTTAGGCCAGAGAAATG	Deletion
rs66791338 region	CATTTCTCTGGCCTAACT GGCCGATTTCTGAAAAC CGCCCCTTTG	CAAAGGGGCGGTTTTCAGAA ATCGGCCAGTTAGGCCAGAGA AATG	Deletion
rs66791338 region	CATTTCTCTGGCCTAACT GGCCGGCAGTCCCTTTCA CTTCCTTG	CAAGGAAGTGAAAGGGACTG CCGGCCAGTTAGGCCAGAGAA ATG	Deletion
rs66791338 region	CATTTCTCTGGCCTAACT GGCCGGTCACACTATCAG GGCAGCTC	GAGCTGCCCTGATAGTGTGAC CGGCCAGTTAGGCCAGAGAAA TG	Deletion
rs66791338 region	CATTTCTCTGGCCTAACT GGCCGGAATGGGACTAG ACCTGGTGTCAAC	GTTGACACCAGGTCTAGTCCC ATTCCGGCCAGTTAGGCCAGA GAAATG	Deletion
rs11633883 A to G	GGGATTTCAGGACTCGA CGACGTTTTGTTTTAGCC	GGCTAAAACAAAACGTCGTCG AGTCCTGAAATCCC	Polymorphism
rs66791338 GAAAG to -	GCCAATAAGGACATGGA AGAGGGGTTGGGGGAAG CC	GGCTTCCCCCAACCCCTCTTC CATGTCCTTATTGGC	Polymorphism
rs11638033 A to D	GGGGTTCGGTTCTATCT ACGGTTTC	GAAACCGTAGATAGAACCGA ACCCC	Polymorphism
rs2289329 m1	GGCTTTAGCACCTCTAAG AAGCTGG	CCAGCTTCTTAGAGGTGCTAA AGCC	Mutation

Table 6 (Continued)

rs2289329 m2	GAAAGCCTCTGGGTTAG AGAGGCTTGG	CCAAGCCTCTCTAACCCAGAG GCTTTC	Mutation
rs2289329 m3	CTGCTTGAGTATGAGTA TTTTCTCCC	GGGAGAAAATACTCATACTCA AGCAG	Mutation
rs2289329 m4	CAATAGGAAGGTGAGTG CCTTCTTCCC	GGGAAGAAGGCACTCACCTTC CTATTG	Mutation
rs2289329 A to D	GGCTAATCTGCAACCAG GACCACC	GGTGGTCCTGGTTGCAGATTA GCC	Polymorphism
rs9635324 D to A	CATTGAACAACAGACAG ACAACATTTGAAGAGAA CTCTAAG	CTTAGAGTTCTCTTCAAATGT TGTCTGTCTGTTGTTCAATG	Polymorphism
rs9635324 m1	GAGAAGTATCTCCCGAA GGTGAGGAAATGG	CCATTTCCTCACCTTCGGGAG ATACTTCTC	Mutation
rs10518693 A to D	GGTAAATGAAGTCTCTC TAAGAATGC	GCATTCTTAGAGAGACTTCAT TTACC	Polymorphism
rs17733719 D to A	GCTTGGAACCTGGCCACA CACATCC	GGATGTGTGTGGCCAGGTTCC AAGC	Polymorphism
rs8033249 D to A	GGGGGAATCTCAAGGCT TCGCC	GGCGAAGCCTTGAGATTCCCC C	Polymorphism

Gblocks: 150bp	**Sequence**
rs66791338 Ancestral	CTTTGCCTGGGGTTCAAGGCCCCCAGTTTGGGATTTCAGGACTCAACGACGTT TTGTTTTAGCCAATAAGGACATGAAAGGGAAGAGGGGTTGGGGGAAGCCTATC CTCTATTACCATCATTTCATAAAAGGCTTTTTTTTTTTTGAGACGGAGTC
rs66791338 Ancesral (alt)	CTTTGCCTGGGGTTCAAGGCCCCCAGTTTGGGATTTCAGGACTCGACGACGTT TTGTTTTAGCCAATAAGGACATGAAAGGGAAGAGGGGTTGGGGGAAGCCTATC CTCTATTACCATCATTTCATAAAAGGCTTTTTTTTTTTTGAGACGGAGTC
rs66791338 Derived	CTTTGCCTGGGGTTCAAGGCCCCCAGTTTGGGATTTCAGGACTCGACGACGTT TTGTTTTAGCCAATAAGGACATGGAAGAGGGGTTGGGGGAAGCCTATCCTCTA TTACCATCATTTCATAAAAGGCTTTTTTTTTTTTGAGACGGAGTC
rs66791338 Derived (alt)	CTTTGCCTGGGGTTCAAGGCCCCCAGTTTGGGATTTCAGGACTCAACGACGTT TTGTTTTAGCCAATAAGGACATGGAAGAGGGGTTGGGGGAAGCCTATCCTCTA TTACCATCATTTCATAAAAGGCTTTTTTTTTTTTGAGACGGAGTC

Cell Culture

LCLs were maintained in suspension at $2*10^6$ cells/mL density in RPMI with 15% FBS at 37°C in 5.0% CO_2. Media was changed every two to three days by removing 75% of cells and media, and replacing with prewarmed RPMI with 15% FBS. One day prior to transfections, media was changed on LCLs to bring the concentration to $5*10^5$ cells/mL, so LCLs would be in log-phase growth at the time of transfection. LCLs were discarded after 3 months of growth.

HEK293s were maintained in 10mL DMEM with 10% FBS on a $10cm^2$ plate. Every two days when cells achieved 90% confluence, HEK293s were split using 1mL 0.05% trypsin, and plated 1:10 in fresh, prewarmed media on a fresh $10cm^2$ plate. HEK293s were discarded after p=30.

HepG2s were maintained in 10mL MEM with 10% FBS on a $10cm^2$ plate. Every four to five days when cells achieved 75% confluence, HepG2s were split using 1mL 0.25% trypsin, and plated 1:10 in fresh, prewarmed media on a fresh $10cm^2$ plate. HepG2s were discarded after p=30.

Luciferase Assays

LCLs at a concentration of $5*10^7$ cells/mL were transfected in 10µL reactions with the pgl4.23 firefly luciferase constructs (1µg/rxn) and the pgl4.74 renilla control plasmid (100ng/rxn) using the Neon electroporation system. Following electroporation, cells were added immediately to 100µL RPMI containing 15% FBS in a black-sided 96-well plate, prewarmed to 37°C. Transfected LCLs in the 96-well plate were grown for 24 hours at 37°C in an incubator with 5.0% CO_2. Then, the DualGlo assay was performed according to manual instructions, and luminescence was read on the Spectrophotomax L at a

wavelength of 470nm. Firefly luminescence was normalized to renilla luminescence for each well to control for variation in transfection efficiency. Five to seven technical replicates were performed per construct per day of testing and three biological replicates were performed per construct to confirm significant results by testing constructs on three separate days.

Where specified, constructs were instead transfected and tested in Hek293 cells and HepG2 cells using the Lipofectamine 2000 and 3000 systems, respectively. In these cases cells were transfected after they reached 50% confluence in the 96-well plate, 24 hours after plating at $2.5*10^4$ cells/mL (Hek293) or $5*10^4$ cells/mL (HepG2). Each well was transfected with 100ng pgl4.23 luciferase test construct and 10ng pgl4.74 renilla control plasmid. DualGlo assays were conducted as described above.

Site Directed Mutagenesis of Test Constructs

For the region containing candidates rs12593066 and rs11633883, ten variants differed between the ancestral and derived haplotypes cloned into the construct. So, site directed mutagenesis was performed on the derived haplotype construct to convert each site individually to the alternative allele, keeping the rest the same. The QuikChange protocol was followed, using the Phusion HF Polymerase in GC-Buffer. Complementary forward and reverse primers were designed with at least 10bp on either side of the target site, ending with 3' G or C on both primers (Table 6). Colonies were screened using boil prep PCR as described above. At least 50% efficiency was achieved for all reactions. Site directed mutagenesis was also performed just on the candidate variant to get the alternative allele in cases for which only one version of the construct was achieved during cloning. In addition, site directed mutagenesis was performed to successively delete 400bp

regions from the construct containing the rs12593066 and rs11633883 candidate eQTLs, after the rs66791338, 5bp indel was determined to be the causal variant in the region. For the large-deletion site directed mutagenesis reactions, primers were designed 25bp long with homology on either side of the deletion for the complementary forward and reverse primers. The protocol described in (Makarova, Kamberov, & Margolis, 2000) was followed using Phusion HF polymerase with GC-buffer.

Association with Gene Expression in Diverse Tissue Types

After identifying two functional regulatory variants using luciferase assays, their association with gene expression was examined further in the gTEX database. While the power for detecting associations in this dataset is less than in the Geuvadis dataset due to smaller samples sizes, it includes diverse tissue types.

Transcription Factor Motif Enrichment

PWMEnrich was used in R to identify candidate transcription factors for binding to the rs66791338 and rs10518693 functional variants. PWMEnrich scans position weight matrices (PWMs) from large experimentally determined datasets (TRANSFAC, JASPAR, Uniprobe, Human DNA Interactome, Jolma et al. 2013) collected by MotifDb to detect potential binding partners for inputted DNA sequences. The PWMs are based on published ChIP-seq, SELEXA, and microarray experiments. For rs66791338 the analysis was run on a 28bp region containing the insert, and on the same region with the deletion, 23bp. For rs10518693 the analysis was run on 25bp regions containing the ancestral or derived single nucleotide polymorphism (SNP). Enrichment for transcription factor binding is determined by comparing the calculated value of the PWM for the inputted sequence to the genomic distribution of PWMs across 500 human promoter sequences. In addition,

PWMEnrich calculates a differential motif score for likelihood of binding one inputted motif over the other, which is the difference of log-normal z-scores between the two inputs.

Expression data was collected for all candidate transcription factors across tissues and cell lines using RNA-seq of coding RNA from 95 human individuals across 27 different tissues (Fagerberg et al., 2013); RNA-seq from 16 human tissues in Illumina Body Map (Asmann et al., 2012; Barbosa-Morais et al., 2012; Derrien et al., 2012); RNA-seq from ENCODE cell lines (Djebali et al., 2012). In addition, the EMBL-EBI Expression Atlas was used to find RNA-seq or microarray experiments in which *IVD* is significantly up- or down-regulated in order to compare to significant changes in expression of transcription factor binding candidates.

Haplotype Analysis and Frequencies

Haploview was used to analyze linkage disequilibrium across the region of interest, and surrounding functional variants and high-scoring loci for selection, in diverse phase 3, 1000 genomes populations. Haplotypes from 1000 genomes phase 3 phased data were also plotted for diverse populations using the HapViz program (Shervin Tabrizi). These haplotypes were sorted by the long haplotype in East Asian populations, using the highest scoring CMS variant in the region, rs12081, as a proxy. In East Asians and other populations, haplotypes were also sorted by functional variants of interest to establish the relationship of functional variants to the long haplotype, and establish the length of the haplotypes in the region. The Bifurcator program (Ben Fry) was also used to visualize haplotype length for ancestral and derived alleles for rs12081, as a tag for the long haplotype in East Asians, as well as for functional alleles. The Diaspora program (Ben Fry)

was used to map frequencies of functional haplotypes and alleles for 1000 genomes phase 3 populations around the world.

Allele Dating

The haplotype-based method described in (Stephens et al., 1998) was used to estimate the ages of derived functional variants in the Yoruba population, as the alleles do not appear on a selected haplotype in these populations. First haplotype length was visualized for this population by observing the chunk of alleles with no recombination surrounding the derived functional variants. Then, the fraction of haplotypes lost with the addition of more SNPs on each side of the derived variants were averaged on each side of the allele. This generated an estimate of haplotype age on either side of the functional derived alleles using a 25-year generation time.

RESULTS

Luciferase Assays

The 2.5kb region containing the candidates rs12593066 and rs11633883 demonstrated a statistically significant increase, 2.5-fold, in luminescence for the derived construct compared to the ancestral (Figure 10a). This was replicated in HepG2 and Hek293 cells, transfected with the same constructs (Figure 10b,c). To pinpoint which of the ten variants that distinguish the ancestral and derived haplotypes were responsible for the difference in expression level for derived versus ancestral, site-directed mutagenesis was conducted to change each variant from the derived version to the ancestral version, one at a time. This analysis revealed that rs66791338, a 5-bp derived deletion, located 30-bp

Derived candidate region drives 2.5-fold increase over ancestral

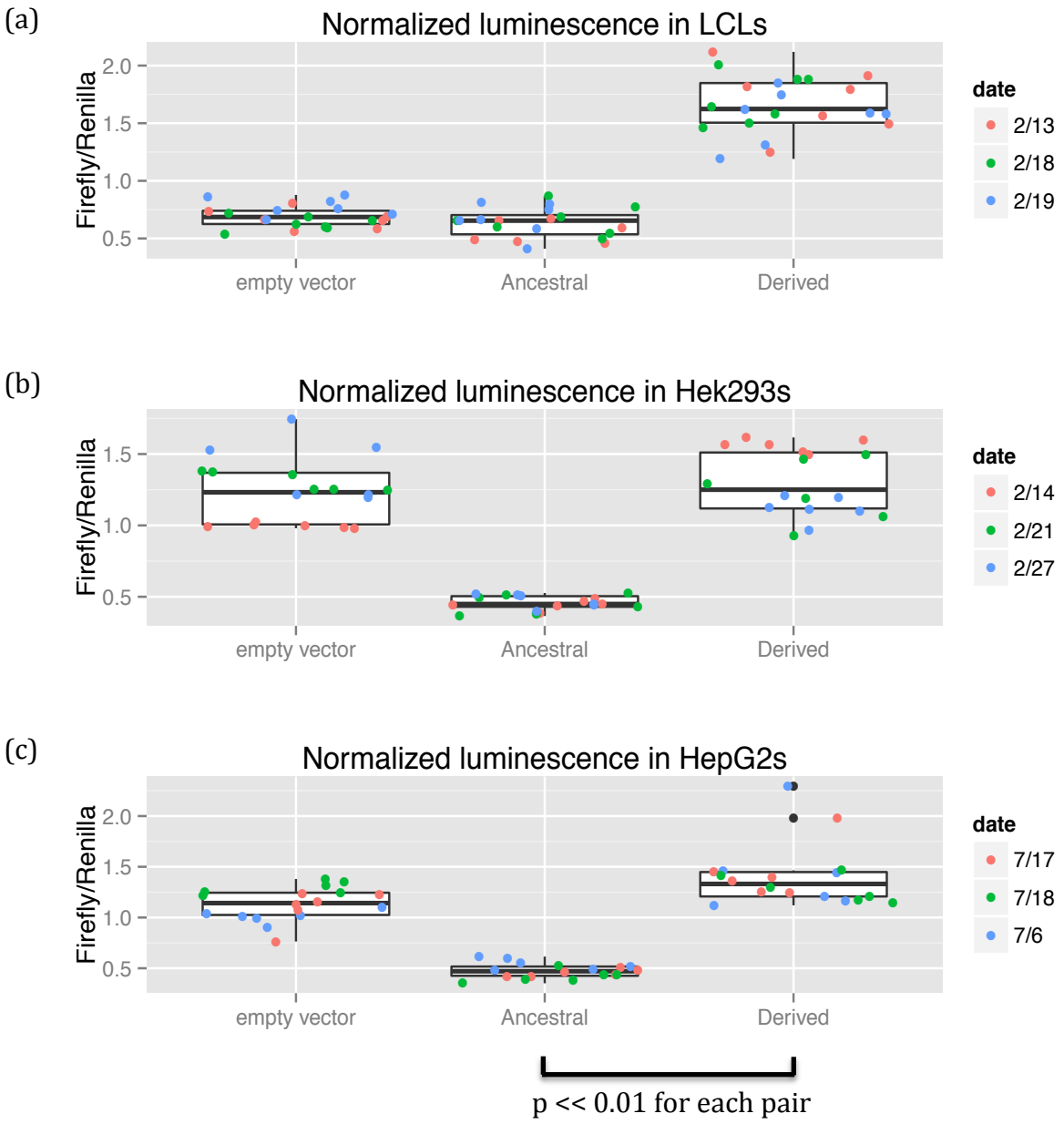


Figure 10. One candidate region drove a 2.5-fold increase in expression by the derived haplotype compared to the ancestral haplotype. This region contains the identified candidate eQTLs rs12593066 and rs11633883, along with eight other variants that differ between the ancestral and derived haplotypes.

downstream from the rs11633883 candidate was responsible for the expression difference between the derived and ancestral versions (tested in Hek293 cells) (Figure 11).

To probe the importance of surrounding context to the rs66791338 variant, a construct was cloned containing only 150bp of the surrounding region, which included just rs66791338 and rs11633883. Constructs were created with all four combinations of ancestral and derived alleles for these variants, and they were tested in HepG2 and Hek293 cells. This 150bp region drove a difference in expression for derived versus ancestral constructs of a much smaller magnitude (1.4-fold) (Figure 12). Constructs with insert sizes ranging in 400bp increments from 600bp to 2.5kb were then created to establish boundaries of the genomic context critical to drive changes in expression. While the effect size does attenuate with decreasing size of insert—the difference in expression between the ancestral insert of 1.6-kb is significantly increased over the ancestral insert of 2.5-kb—by far the biggest magnitude of change in effect size occurs between the 150-bp and 600-bp insert constructs (Figure 13), indicating potential binding of a factor that modulates the primary expression signal within the 600-bp region.

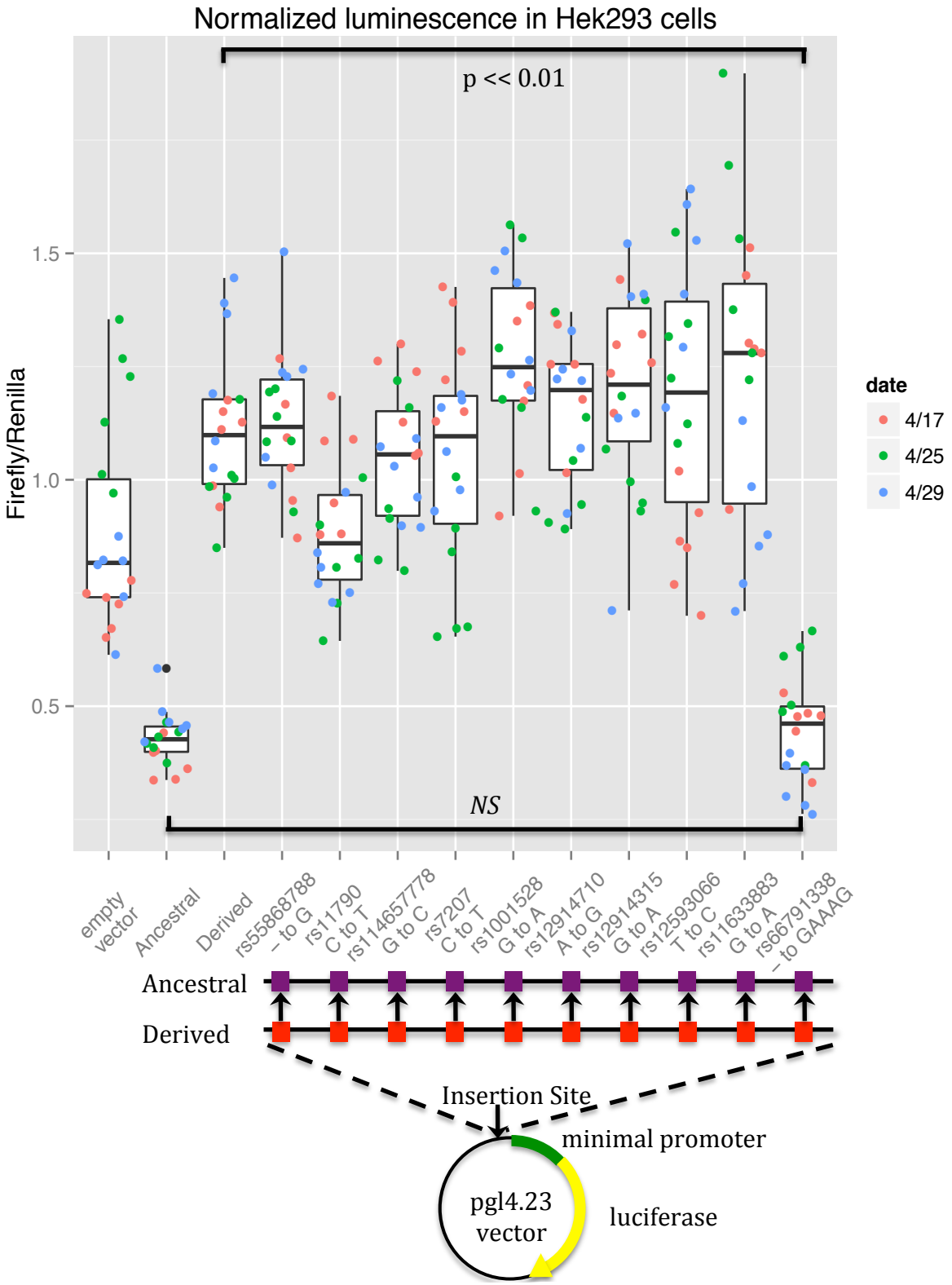
In addition, in both HEK293 and HepG2 cells the rs11633883 ancestral and derived alleles interacted with the rs66791338, such that pairing the ancestral rs11633883 ancestral allele with the rs66791338 derived allele slightly lowered expression from the highest derived, derived level (Figure 12)**.** Also, pairing the rs11633883 derived allele with the rs66791338 ancestral allele lowered expression even below the ancestral, ancestral level. These experiments demonstrate that both rs66791338 and rs11633883 are important to control expression levels in the Hek293 and HepG2 cells. However, when these four variations of alleles were tested on 2.5-kb inserts in LCLs, rs11633883 exhibited

Figure 11. Site Directed Mutagenesis identifies 5 bp indel driving expression change

Site directed mutagenesis revealed that changing a single variant, the 5-bp indel rs66791338, from derived to ancestral drives a level of luminescence that is not significantly different from the ancestral haplotype and is highly significantly different from the derived haplotype ($p << 0.01$).

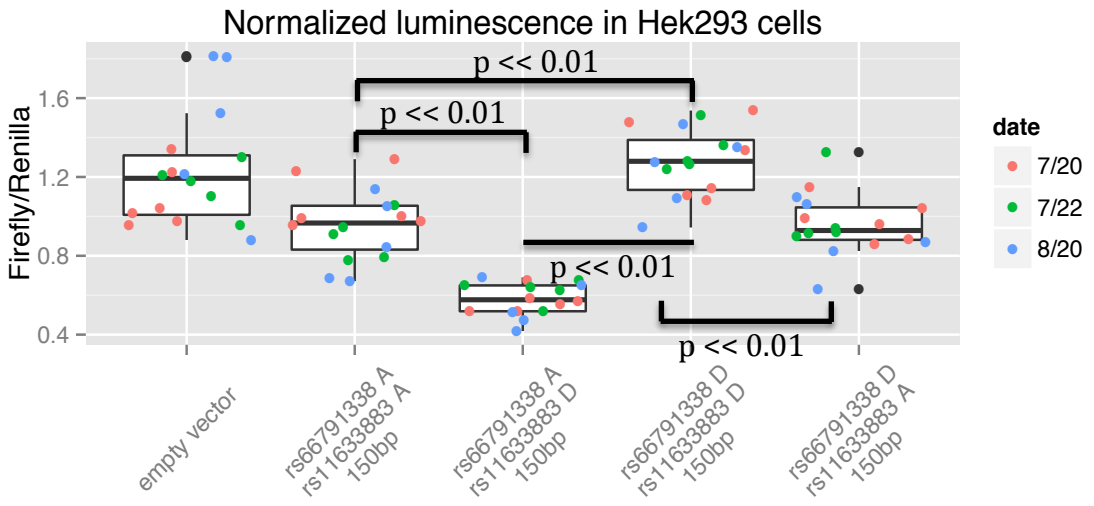
Figure 11 (Continued)

Site Directed Mutagenesis identifies 5-bp indel driving expression change



150-bp regions surrounding the 5-bp indel drive smaller changes in expression

(a)



(b)

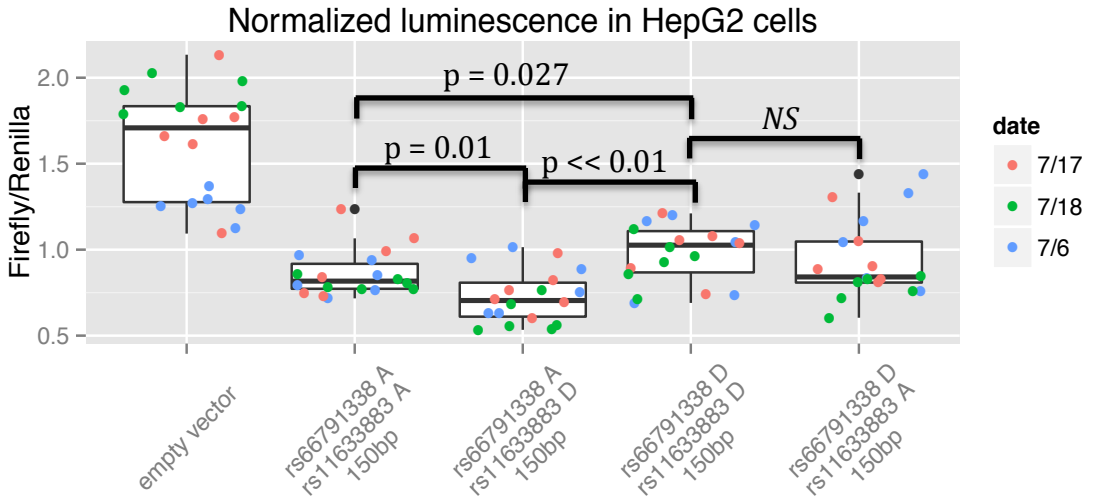


Figure 12. Testing only a 150-bp region surrounding rs66791338 and rs11633883 drove only a 1.36-fold change in expression between ancestral and derived haplotypes. Furthermore, ancestral rs11633883 reduced expression when paired with rs66791338 derived, and derived rs11633883 reduced expression when paired with rs66791338 ancestral.

Tests of different region sizes surrounding the 5-bp indel reveals largest increase in effect size for regions with at least 400 bp upstream of the indel

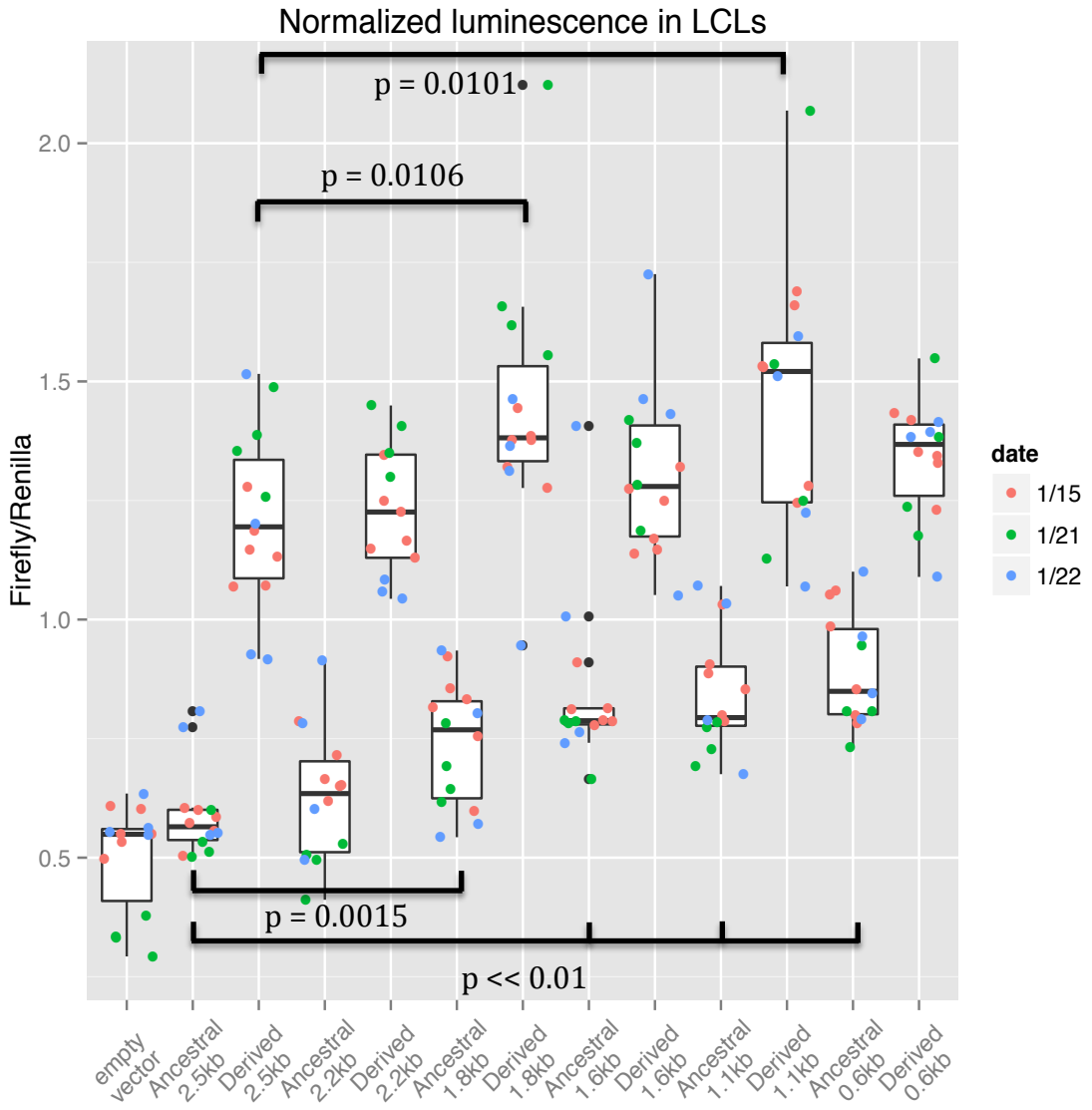


Figure 13. Regions surrounding rs66791338 of at least 600 bp, including at least 400 bp upstream of the variant, drive large fold-changes in expression.

no significant impact on expression of derived rs66791338, and significantly increased expression of ancestral rs66791338, giving evidence of cell type specificity in the regulation of expression by this variant (Figure 14).

Of the six additional regions containing seven candidate variants with r^2>0.9 to the top associated eQTL for *IVD* in the European population of the Geuvadis dataset (rs10518693) and the top associated GWAS hit for isovaleryl-CoA blood metabolite levels (rs9635324; r^2 = 1 with rs10518693 in GBR), only rs10518693, tested in a 1kb region, exhibited a significant difference between ancestral and derived alleles tested in LCLs (Figure 15). Notably, this variant had no impact on expression, neither when tested with only 150 bp of genomic DNA in a luciferase assay, nor with 150 bp in a high-throughput assay to test regulatory function (Tewhey, RS; unpublished data).

Association of Functional Variants with Gene Expression in the gTEX Database

The two functional variants, rs10518693 and rs66791338 both associate with *IVD* gene expression in the gTEX dataset with p < 1E-5. However, they exhibit association in different tissue types. rs10518693 is significantly associated with *IVD* gene expression in esophagus muscularis (p = 3E-6), while rs7207 and rs1001528 (linked proxies for rs66791338) associate with *IVD* gene expression in skeletal muscle (p = 4.6E-6 for rs7207 and p = 2.1E-6 for rs1001528). In addition, rs10518693 associates with expression of a nearby gene of unknown function *DISP2* in skin and another transcript on the opposite strand complementary to *DISP2* of unknown function, *RP11-64K12.4*, in subcutaneous adipose tissue. The proxies for rs66791338 also associate with *DISP2* in skin and subcutaneous adipose tissue. The associations of functional *IVD* eQTLs with *DISP2* and *RP11-64K12.4* expression do not replicate in the Geuvadis dataset. Two important factors

LCLs show a different pattern of interaction for rs66791338 (the 5-bp indel) and rs11633883

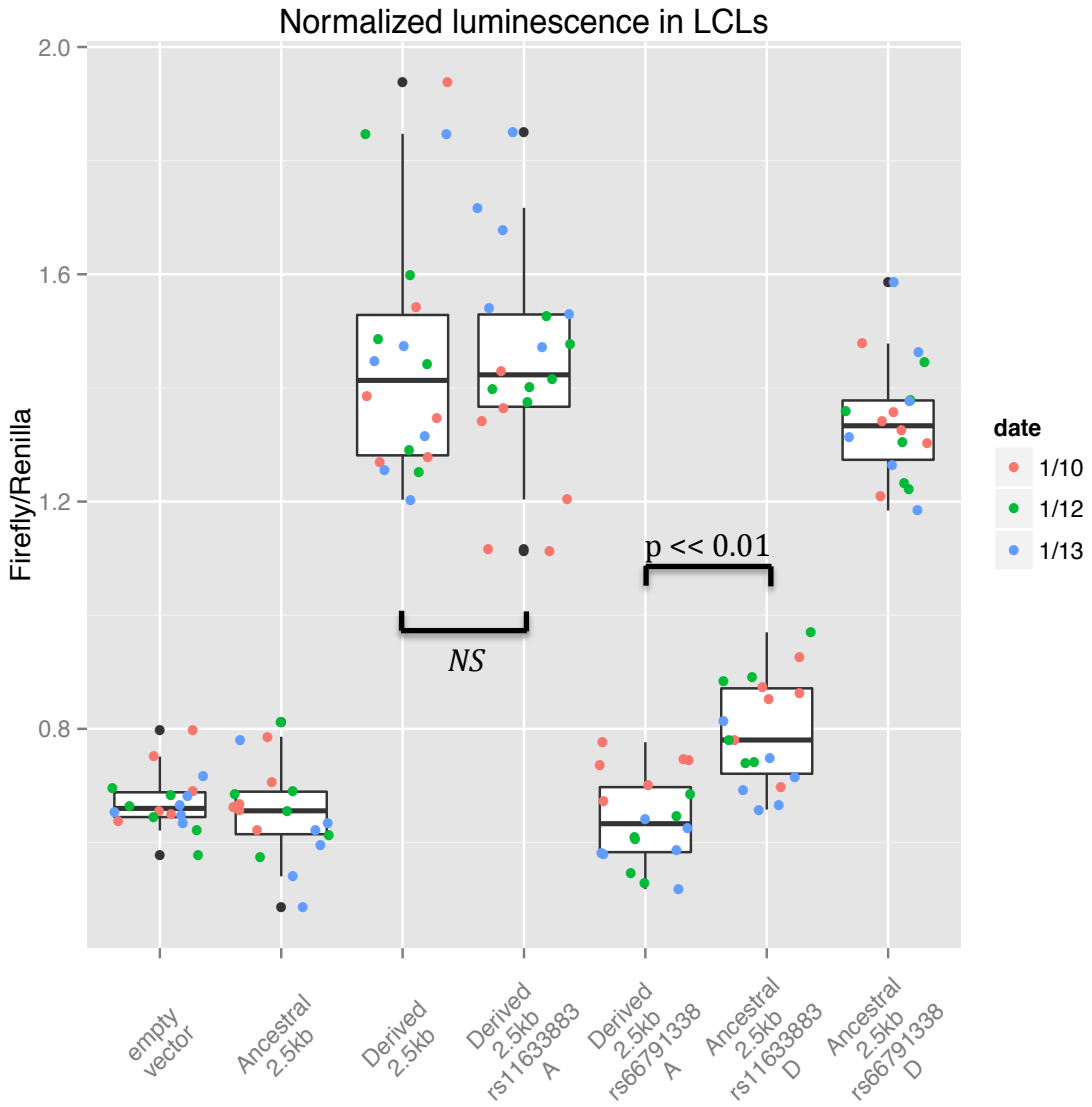
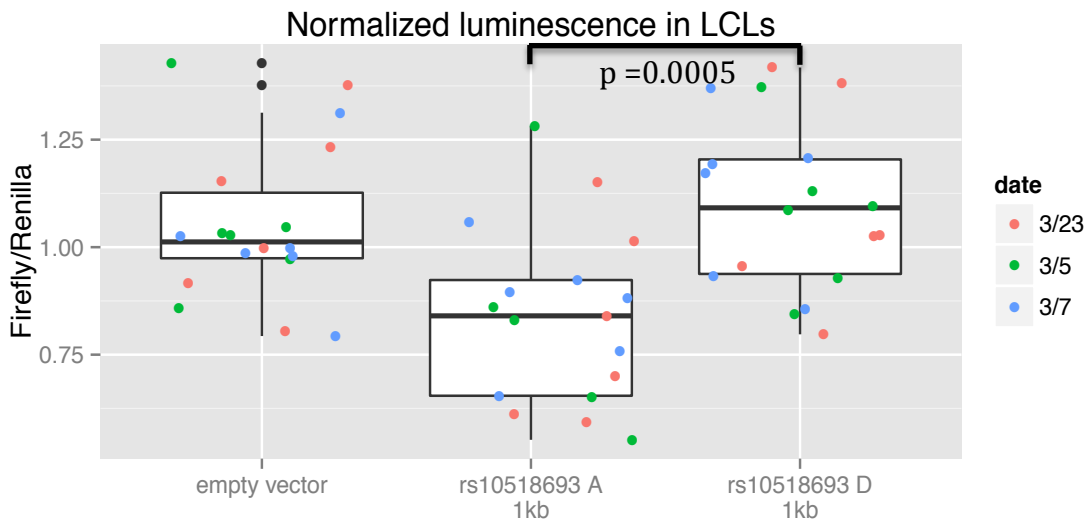


Figure 14. In LCLs using 2.5 kb region sizes, rs11633883 derived allele increases expression on the ancestral haplotype, whereas rs11633883 ancestral allele has no impact on expression of the derived haplotype.

Top *IVD* eQTL, rs10518693, only shows an increase in expression for derived over ancestral using 1 kb of surrounding context

(a)



(b)

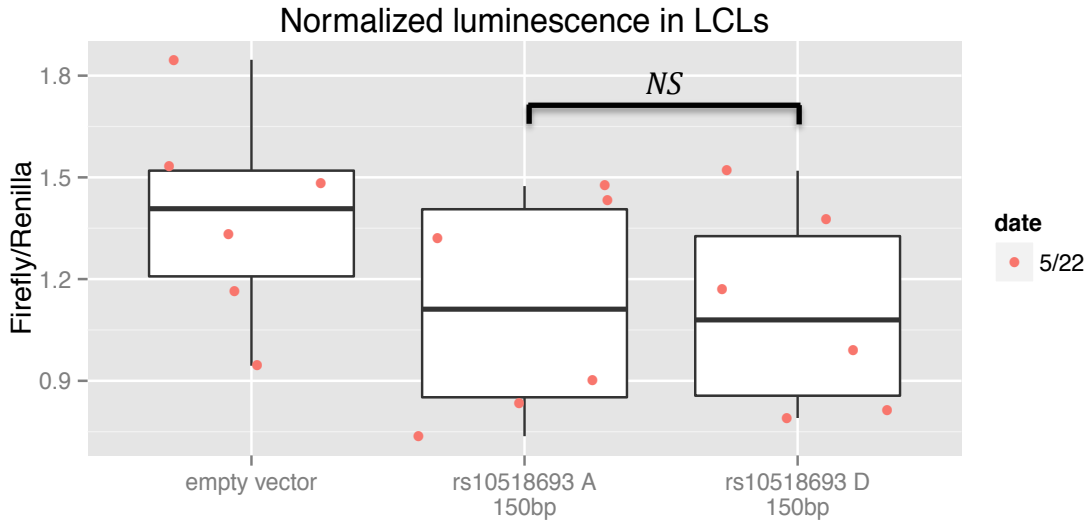


Figure 15. Testing a 1-kb region surrounding the top eQTL for *IVD*, rs10518693, shows a 1.25-fold increase in luminescence for the derived allele over the ancestral allele. Testing these alleles using 150-bp region reveals no change in expression.

limit this analysis of tissue specificity in gene expression. First, the gTEX dataset has a minimum of just 60 samples in each tested tissue type (*e.g.*, 143 in skeletal muscle, 111 in subcutaneous adipose, and 114 in skin), whereas the Geuvadis dataset has 462 LCL lines, including 373 in the EUR population where the *IVD* association in LCLs was detected. Second, important tissues of *IVD* expression and function such as liver, kidney, and pancreas are not included in the gTEX data.

Transcription Factor Motif Enrichment

Analysis by the PWMEnrich program for transcriptional binding motifs yielded several candidates for differential binding to the ancestral versus derived alleles of rs10518693 and rs66791338 (Tables 7-10; Figures 16-19). Since affinity for binding motifs is also modulated by abundance of transcription factors in the nucleus, differences in expression across tissue types and significant changes in expression across experimental conditions from RNAseq and microarray experiments were observed for correlation with differences in *IVD* expression to examine likelihood of the candidates. In the cases of these loci, binding candidates with the ancestral alleles are predicted to act as repressors, while binding candidates with the derived alleles are predicted to act as enhancers. Analyzing binding candidates according to these criteria narrows the field to a few candidates for each functional variant (Table 11-12; Figures 20-21).

Haplotype Analyses

To visualize the haplotype surrounding the derived functional variants downstream of *IVD*, haplotypes were plotted for the 150kb region in 1000 genomes phase 3 populations (Figure 22). These plots reveal a high frequency long-range haplotype in the East Asian population, not present in other populations. The long-range haplotype is also clearly

Table 7. Candidate transcription factors for binding rs66791338 insertion

p < 0.05 for PWM motif enrichment				
Candidate Transcription Factor	**Binding -log10(p)**	**Binding Raw Score**	**Differential Binding Score**	**Differential Binding Rank**
TCF7L2	3.88	1354.25	8.03	1
LEF1	3.78	580.52	7.68	2
ESRRA	3.04	18.04	3.97	11
ESRRA	2.90	3.49	3.88	13
RARG	2.81	98.75	4.69	4
TPI1	2.71	29.59	4.26	6
RAN	2.29	42.99	4.71	3
DAB2	2.29	29.21	3.07	19
ARFGAP1	2.25	16.93	-0.49	1164
RARA	2.24	1.56	4.39	5
CSNK2B	2.24	20.03	4.14	7
DUSP22	2.23	19.64	3.41	16
CLK1	2.18	16.30	3.32	17
ZSCAN31	2.16	12.52	3.09	18
PRKRIR	2.12	31.35	3.66	15
RAX	2.11	18.99	4.04	8
DIS3	2.05	18.35	0.19	350
RUVBL1	1.98	23.07	4.01	9
KIAA0907	1.96	10.84	1.31	70
ZBTB25	1.93	19.78	3.68	14
RFC3	1.90	14.60	2.40	29
RARG	1.87	0.23	3.98	10
NR4A1	1.83	5.83	-0.33	1142
OLIG3	1.82	13.72	2.44	27
SRF	1.82	4.89	3.91	12
IRX2	1.79	7.39	2.69	24
EBF1	1.74	7.91	0.62	158
ZNF160	1.71	8.74	-0.19	1065
THAP5	1.70	12.79	1.67	54
ESRRG	1.66	0.01	0.78	124
RBM42	1.64	14.41	1.52	61
ZNF766	1.64	4.89	1.23	78
BCL11A	1.58	10.07	1.55	60
RNF138	1.52	5.45	2.26	31
SPATS2	1.52	6.01	2.75	21
KIF22	1.51	4.68	0.67	147
SOX14	1.51	7.98	3.05	20
ADARB1	1.48	7.81	1.39	64
TCEAL6	1.47	8.23	2.70	23

Table 7 (Continued).

IRX5	1.47	4.76	2.30	30
CBX3	1.44	3.59	-0.34	1146
ELF5	1.44	5.85	0.44	217
MEX3C	1.40	8.95	0.62	156
SP1	1.39	51.02	-0.24	1102
RAB14	1.37	3.57	-0.29	1124
ELK1	1.36	3.85	0.92	104
NELFB	1.36	4.83	-0.17	1026
ZMAT2	1.34	6.03	-0.20	1069
PDLIM5	1.33	9.07	1.79	47
ZDHHC5	1.33	6.47	-0.20	1072
MRPL2	1.33	6.20	0.08	472
MCTP2	1.31	5.23	2.24	32
STAT2::STAT1	1.30	4.03	1.59	58

Table 8. Candidate transcription factors for binding rs66791338 deletion

p < 0.05 for PWM motif enrichment				
Candidate Transcription Factor	**Binding -log10(p)**	**Binding Raw Score**	**Differential Binding Score**	**Differential Binding Rank**
SMUG1	3.04	29.94	4.52	1
ARFGAP1	2.46	24.11	0.49	85
PAX5	2.14	3.99	4.45	2
NR4A1	1.97	7.33	0.33	129
DIS3	1.97	17.16	-0.19	940
TFAM	1.89	19.04	3.05	6
REST	1.87	0.46	3.93	3
EXOSC3	1.86	10.41	1.35	28
ZNF160	1.80	10.25	0.19	206
LAS1L	1.72	14.43	2.74	8
PAX2	1.66	0.17	3.57	4
PAX1	1.66	0.02	3.38	5
PPP5C	1.61	4.71	2.01	16
CBX3	1.59	4.57	0.34	125
MYLK	1.58	10.08	2.25	10
SFT2D1	1.52	8.67	2.82	7
ESRRA	1.51	0.24	1.96	19
SP1	1.50	65.76	0.24	169
RAB14	1.50	4.64	0.29	147
YY2	1.48	5.50	2.34	9
EBF1	1.47	5.53	-0.62	1113
HCLS1	1.46	5.73	0.97	44
AVEN	1.44	7.64	1.97	18
NELFB	1.44	5.52	0.17	245
ETV6	1.43	0.15	0.84	52
ZMAT2	1.43	7.05	0.20	202
ZDHHC5	1.42	7.58	0.20	199
KIAA0907	1.39	5.75	-1.31	1201
MED30	1.36	7.91	0.66	62
ETFB	1.34	4.92	0.36	116
ESRRG	1.32	0.00	-0.78	1130
NONO	1.32	4.93	0.30	139
ESRRA	1.32	0.03	1.96	19

Table 9. Candidate transcription factors for binding rs10518693 ancestral

p < 0.05 for PWM motif enrichment

Candidate Transcription Factor	Binding -log10(p)	Binding Raw Score	Differential Binding Score	Differential Binding Rank
RBM22	259.86	3.67	-1.8E-03	474
MRPS25	160.25	3.65	-3.7E-04	506
PPP1R10	55.73	3.24	-4.5E+00	1
RFX4	217.18	3.01	1.0E-05	562
ZNF830	101.40	2.69	1.4E-05	563
TRMT1	125.77	2.67	1.2E-06	549
DTL	40.11	2.40	6.4E-05	576
MYOD1	27.27	2.33	2.0E-05	566
TCEAL2	31.03	2.23	2.2E-05	567
DLX6	21.33	2.22	5.3E-03	743
BARHL2	12.39	2.06	7.0E-08	544
NFIL3	12.53	1.90	3.5E-03	719
BARHL2	9.43	1.86	6.6E-05	578
HHEX	14.85	1.86	1.4E-03	662
RFC2	12.72	1.83	5.8E-05	575
CNOT6	13.04	1.81	1.5E-04	590
RIOK2	13.89	1.78	8.7E-03	778
CBFB	9.95	1.75	-9.4E-03	407
BARHL2	7.78	1.73	2.7E-07	547
MAGEF1	7.70	1.73	-6.0E-04	502
E2F2	0.02	1.73	-2.4E+00	4
CSTF2	39.34	1.72	4.7E-04	619
BARHL2	7.85	1.71	1.6E-03	667
SMAD2	9.01	1.70	6.1E-04	629
FOXD3	11.36	1.69	2.3E-06	552
RNF138	6.60	1.65	1.2E-03	654
TSNAX	14.78	1.63	-2.9E-02	318
DUSP22	8.83	1.62	3.6E-03	720
SOX13	8.73	1.60	-2.8E+00	3
ZBTB43	6.57	1.60	1.0E-02	801
ESX1	4.74	1.58	-1.4E-01	165
LARP1	9.42	1.55	5.5E-04	625
GLTPD1	6.18	1.54	-2.9E+00	2
NOTO	5.14	1.48	1.3E-09	540
CBX3	3.72	1.46	-2.3E+00	5
HNRNPA0	7.15	1.44	2.3E-02	868
HSF4	1.27	1.40	-3.2E-01	87
MECP2	4.53	1.40	2.1E-03	683
LMX1B	3.99	1.34	-8.1E-04	492

Table 9 (Continued).

MEX3C	7.64	1.34	1.2E-03	656
ZNF3	6.38	1.33	4.3E-03	734
TFE3	6.42	1.33	2.0E-04	596
TRIP10	4.81	1.32	6.3E-03	751
HSF1	0.04	1.32	-3.6E-01	80
HOXD8	6.00	1.31	2.9E-03	706
METTL21B	7.17	1.30	1.9E-02	847

Table 10. Candidate transcription factors for binding rs10518693 ancestral

p < 0.05 for PWM motif enrichment

Candidate Transcription Factor	Binding -log10(p)	Binding Raw Score	Differential Binding Score	Differential Binding Rank
RBM22	259.59	3.67	-1.82E-03	816
MRPS25	160.22	3.65	-3.70E-04	784
RFX4	217.18	3.01	1.04E-05	728
ZNF830	101.40	2.69	1.40E-05	727
TRMT1	125.77	2.67	1.17E-06	741
DTL	40.12	2.40	6.37E-05	714
MYOD1	27.27	2.33	2.04E-05	724
ING3	8.51	2.25	3.77E+00	1
TCEAL2	31.03	2.23	2.20E-05	723
DLX6	21.40	2.22	5.29E-03	547
GLYCTK	43.37	2.18	2.70E+00	4
ETS1	15.89	2.14	3.44E+00	2
BARHL2	12.39	2.06	6.96E-08	746
NFIL3	12.56	1.90	3.46E-03	571
BARHL2	9.43	1.86	6.62E-05	712
HHEX	14.86	1.86	1.41E-03	628
RFC2	12.72	1.83	5.81E-05	715
CNOT6	13.05	1.81	1.54E-04	700
RIOK2	13.98	1.78	8.71E-03	512
CBFB	9.89	1.74	-9.38E-03	883
BARHL2	7.78	1.73	2.68E-07	743
MAGEF1	7.69	1.73	-6.02E-04	788
CSTF2	39.35	1.72	4.66E-04	671
BARHL2	7.86	1.71	1.61E-03	623
SMAD2	9.01	1.70	6.10E-04	661
FOXD3	11.36	1.69	2.27E-06	738
RNF138	6.60	1.65	1.17E-03	636
DUSP22	8.85	1.62	3.57E-03	570
TSNAX	14.41	1.61	-2.86E-02	972
ZBTB43	6.61	1.61	1.04E-02	489
WDR83	8.46	1.58	1.98E+00	10
FOSL1	8.09	1.55	2.04E+00	8
LARP1	9.42	1.55	5.54E-04	665
MEF2BNB-MEF2B	12.23	1.53	2.31E+00	6
ESX1	4.46	1.52	-1.37E-01	1125
EDN1	9.48	1.52	2.38E+00	5
NOTO	5.14	1.48	1.26E-09	750
CREB3	0.21	1.48	2.89E+00	3
HNRNPA0	7.28	1.45	2.26E-02	422

Table 10 (Continued).

MECP2	4.53	1.40	2.07E-03	607
XBP1	0.10	1.38	1.65E+00	19
LMX1B	3.99	1.34	-8.09E-04	798
MEX3C	7.65	1.34	1.22E-03	634
ZNF3	6.41	1.34	4.27E-03	556
TFE3	6.42	1.33	1.98E-04	694
TRIP10	4.83	1.33	6.32E-03	539
BAX	2.82	1.32	1.13E+00	37
METTL21B	7.28	1.31	1.90E-02	443
HOXD8	6.02	1.31	2.93E-03	584
MAGOH	3.26	1.31	1.24E+00	34

Figure 16. Position Weight Matrices for all candidate transcription factors ($p < 0.05$), rs66791338, Deletion

PWMs from PWMenrich for all transcription factors with $p < 0.05$ for binding to the queried sequence.

Figure 16 (Continued)

Position Weight Matrices for all candidate transcription factors (p < 0.05)

rs66791338
Deletion

ATAAGGACATGGAAGAGGGGTTG

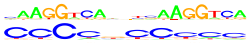
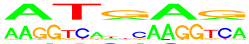
Rank	Target	PWM	Motif ID	Raw score	P–value
1	SMUG1		Hsapiens–hPDI–SMUG1	29.9	0.000918
2	ARFGAP1		Hsapiens–hPDI–ARFGAP1	24.1	0.00347
3	PAX5		Hsapiens–jolma2013–PAX5	3.99	0.00724
4	NR4A1		Hsapiens–hPDI–NR4A1	7.33	0.0106
5	DIS3		Hsapiens–hPDI–DIS3	17.2	0.0107
6	TFAM		Hsapiens–hPDI–TFAM	19	0.0128
7	REST		Hsapiens–JASPAR_CORE–REST–MA0138.2	0.456	0.0135
8	EXOSC3		Hsapiens–hPDI–EXOSC3	10.4	0.0139
9	ZNF160		Hsapiens–hPDI–ZNF160	10.2	0.016
10	LAS1L		Hsapiens–hPDI–LAS1L	14.4	0.0189
11	PAX2		Hsapiens–jolma2013–PAX2	0.171	0.0217
12	PAX1		Hsapiens–jolma2013–PAX1	0.0154	0.0221
13	PPP5C		Hsapiens–hPDI–PPP5C	4.71	0.0247
14	CBX3		Hsapiens–hPDI–LOC653972	4.57	0.0257
15	MYLK		Hsapiens–hPDI–MYLK	10.1	0.026
16	SFT2D1		Hsapiens–hPDI–SFT2D1	8.67	0.0299
17	ESRRA		Hsapiens–jolma2013–ESRRA–6	0.241	0.031
18	SP1		Hsapiens–JASPAR_CORE–SP1–MA0079.2	65.8	0.0316
19	RAB14		Hsapiens–hPDI–RAB14	4.64	0.0318
20	YY2		Hsapiens–jolma2013–YY2–2	5.5	0.0331
21	EBF1		Hsapiens–hPDI–EBF1	5.53	0.034
22	HCLS1		Hsapiens–hPDI–HCLS1	5.73	0.0346
23	AVEN		Hsapiens–hPDI–AVEN	7.64	0.0363
24	NELFB		Hsapiens–hPDI–COBRA1	5.52	0.0367
25	ETV6		Hsapiens–jolma2013–ETV6	0.147	0.0369
26	ZMAT2		Hsapiens–hPDI–ZMAT2	7.05	0.0374
27	ZDHHC5		Hsapiens–hPDI–ZDHHC5	7.58	0.0381
28	KIAA0907		Hsapiens–hPDI–KIAA0907	5.75	0.0403
29	MED30		Hsapiens–hPDI–THRAP6	7.91	0.0441
30	ETFB		Hsapiens–hPDI–ETFB	4.92	0.0462
31	ESRRG		Hsapiens–jolma2013–ESRRG–2	0.00148	0.0473
32	NONO		Hsapiens–hPDI–NONO	4.93	0.0476
33	ESRRA		Hsapiens–jolma2013–ESRRA–5	0.0318	0.0482
34	MRPL2		Hsapiens–hPDI–MRPL2	5.81	0.0512
35	XG		Hsapiens–hPDI–XG	3.99	0.0527
36	ESRRG		Hsapiens–jolma2013–ESRRG–3	1.31	0.054
37	MYC::MAX		Hsapiens–JASPAR_CORE–MYC::MAX–MA0059.1	0.973	0.0541
38	ELF5		Hsapiens–jolma2013–ELF5–2	3.33	0.0563
39	ESRRA		Hsapiens–hPDI–ESRRA	3.76	0.0571
40	GRHL1		Hsapiens–jolma2013–GRHL1	0.17	0.0582
41	YY1		Hsapiens–JASPAR_CORE–YY1–MA0095.1	3.15	0.0586
42	MESP1		Hsapiens–jolma2013–MESP1	5	0.0596
43	PTCD1		Hsapiens–hPDI–PTCD1	3.35	0.0597
44	KIF22		Hsapiens–hPDI–KIF22	3.36	0.0602
45	ESRRA		Hsapiens–jolma2013–ESRRA–2	0.00589	0.0613
46	SNRPB2		Hsapiens–hPDI–SNRPB2	3.95	0.0632
47	NCBP2		Hsapiens–hPDI–NCBP2	2.83	0.0638
48	PRDX5		Hsapiens–hPDI–PRDX5	1.62	0.0691
49	TFE3		Hsapiens–hPDI–TFE3	4.09	0.0696
50	CREB3L1		Hsapiens–jolma2013–CREB3L1–4	0.0123	0.0707
51	ECSIT		Hsapiens–hPDI–ECSIT	2.84	0.0727
52	MEX3C		Hsapiens–hPDI–RKHD2	4.67	0.0737
53	MSRB3		Hsapiens–hPDI–MSRB3	2.77	0.0771
54	MZF1_5–13		Hsapiens–JASPAR_CORE–MZF1_5–13–MA0057.1	3.93	0.0772
55	ZNF766		Hsapiens–hPDI–ZNF766	2.15	0.0782
56	PITX1		Hsapiens–hPDI–PITX1	2.78	0.0795
57	NKX6–1		Hsapiens–jolma2013–NKX6–1–2	1.77	0.0809
58	POLE3		Hsapiens–hPDI–POLE3	2.65	0.0828
59	ETS1		Hsapiens–JASPAR_CORE–ETS1–MA0098.1	2.99	0.0846
60	ELF5		Hsapiens–jolma2013–ELF5	1.49	0.0854

Figure 17. Position Weight Matrices for all candidate transcription factors (p < 0.05), rs66791338, Insertion

PWMs from PWMenrich for all transcription factors with p < 0.05 for binding to the queried sequence.

Figure 17 (Continued)

Position Weight Matrices for all candidate transcription factors (p < 0.05)

rs66791338
Insertion

ATAAGGACATGAAAGGGAAGAGGGGTTG

Rank	Target	PWM	Motif ID	Raw score	P–value
1	TCF7L2		Hsapiens–JASPAR_2014–TCF7L2–MA0523.1	1350	0.000132
2	LEF1		Hsapiens–jolma2013–LEF1	581	0.000166
3	ESRRA		Hsapiens–jolma2013–ESRRA–5	18	0.000906
4	ESRRA		Hsapiens–jolma2013–ESRRA–2	3.49	0.00127
5	RARG		Hsapiens–jolma2013–RARG–6	98.7	0.00155
6	TPI1		Hsapiens–hPDI–TPI1	29.6	0.00196
7	RAN		Hsapiens–hPDI–RAN	43	0.0051
8	DAB2		Hsapiens–hPDI–DAB2	29.2	0.00513
9	ARFGAP1		Hsapiens–hPDI–ARFGAP1	16.9	0.00565
10	RARA		Hsapiens–jolma2013–RARA	1.56	0.0057
11	CSNK2B		Hsapiens–hPDI–CSNK2B	20	0.00576
12	DUSP22		Hsapiens–hPDI–DUSP22	19.6	0.00588
13	CLK1		Hsapiens–hPDI–CLK1	16.3	0.00655
14	ZSCAN31		Hsapiens–hPDI–ZNF323	12.5	0.00689
15	PRKRIR		Hsapiens–hPDI–PRKRIR	31.3	0.00765
16	RAX		Hsapiens–hPDI–RAX	19	0.0078
17	DIS3		Hsapiens–hPDI–DIS3	18.4	0.00883
18	RUVBL1		Hsapiens–hPDI–RUVBL1	23.1	0.0104
19	KIAA0907		Hsapiens–hPDI–KIAA0907	10.8	0.0109
20	ZBTB25		Hsapiens–hPDI–ZBTB25	19.8	0.0116
21	RFC3		Hsapiens–hPDI–RFC3	14.6	0.0125
22	RARG		Hsapiens–jolma2013–RARG–3	0.234	0.0135
23	NR4A1		Hsapiens–hPDI–NR4A1	5.83	0.0148
24	OLIG3		Hsapiens–hPDI–OLIG3	13.7	0.0151
25	SRF		Hsapiens–JASPAR_2014–SRF–MA0083.2	4.89	0.0152
26	IRX2		Hsapiens–jolma2013–IRX2	7.39	0.0161
27	EBF1		Hsapiens–hPDI–EBF1	7.91	0.0184
28	ZNF160		Hsapiens–hPDI–ZNF160	8.74	0.0194
29	THAP5		Hsapiens–hPDI–THAP5	12.8	0.0201
30	ESRRG		Hsapiens–jolma2013–ESRRG–2	0.00984	0.0217
31	RBM42		Hsapiens–hPDI–MGC10433	14.4	0.0227
32	ZNF766		Hsapiens–hPDI–ZNF766	4.89	0.0229
33	BCL11A		Hsapiens–hPDI–BCL11A	10.1	0.0263
34	RNF138		Hsapiens–hPDI–RNF138	5.45	0.0302
35	SPATS2		Hsapiens–hPDI–SPATS2	6.01	0.0303
36	KIF22		Hsapiens–hPDI–KIF22	4.68	0.0306
37	SOX14		Hsapiens–hPDI–SOX14	7.98	0.0308
38	ADARB1		Hsapiens–hPDI–ADARB1	7.81	0.0334
39	TCEAL6		Hsapiens–hPDI–TCEAL6	8.23	0.0339
40	IRX5		Hsapiens–jolma2013–IRX5	4.76	0.0343
41	CBX3		Hsapiens–hPDI–LOC653972	3.59	0.0359
42	ELF5		Hsapiens–jolma2013–ELF5–2	5.85	0.0362
43	MEX3C		Hsapiens–hPDI–RKHD2	8.95	0.0394
44	SP1		Hsapiens–JASPAR_CORE–SP1–MA0079.2	51	0.0403
45	RAB14		Hsapiens–hPDI–RAB14	3.57	0.0424
46	ELK1		Hsapiens–JASPAR_CORE–ELK1–MA0028.1	3.85	0.0432
47	NELFB		Hsapiens–hPDI–COBRA1	4.83	0.0433
48	ZMAT2		Hsapiens–hPDI–ZMAT2	6.03	0.0457
49	PDLIM5		Hsapiens–hPDI–PDLIM5	9.07	0.0463
50	ZDHHC5		Hsapiens–hPDI–ZDHHC5	6.47	0.0466
51	MRPL2		Hsapiens–hPDI–MRPL2	6.2	0.0472
52	MCTP2		Hsapiens–hPDI–MCTP2	5.23	0.0487
53	STAT2::STAT1		Hsapiens–JASPAR_2014–STAT2::STAT1–MA0517.1	4.03	0.0496
54	PTCD1		Hsapiens–hPDI–PTCD1	3.95	0.0507
55	AGGF1		Hsapiens–hPDI–AGGF1	9.08	0.0513
56	EXOSC3		Hsapiens–hPDI–EXOSC3	4.8	0.0536
57	ZSCAN9		Hsapiens–hPDI–ZNF193	3.95	0.054
58	BARX1		Hsapiens–hPDI–BARX1	5.53	0.0548
59	GTPBP6		Hsapiens–hPDI–GTPBP6	4.04	0.0585
60	DUSP26		Hsapiens–hPDI–DUSP26	3.79	0.0589

Figure 18. Position Weight Matrices for all candidate transcription factors ($p < 0.05$), rs10518693, Ancestral

PWMs from PWMenrich for all transcription factors with $p < 0.05$ for binding to the queried sequence.

Figure 18 (Continued)

Position Weight Matrices for all candidate transcription factors (p < 0.05)

rs10518693
Ancestral

TGGTAAATGAAGCCTCTCTAAGAAT

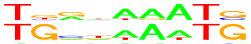
Rank	Target	PWM	Motif ID	Raw score	P–value
1	RBM22		Hsapiens–hPDI–RBM22	260	0.000213
2	MRPS25		Hsapiens–hPDI–MRPS25	160	0.000222
3	PPP1R10		Hsapiens–hPDI–PPP1R10	55.7	0.000581
4	RFX4		Hsapiens–hPDI–RFX4	217	0.000968
5	ZNF830		Hsapiens–hPDI–CCDC16	101	0.00204
6	TRMT1		Hsapiens–hPDI–TRMT1	126	0.00211
7	DTL		Hsapiens–hPDI–DTL	40.1	0.00397
8	MYOD1		Hsapiens–hPDI–MYOD1	27.3	0.00465
9	TCEAL2		Hsapiens–hPDI–TCEAL2	31	0.00585
10	DLX6		Hsapiens–hPDI–DLX6	21.3	0.00599
11	BARHL2		Hsapiens–jolma2013–BARHL2–4	12.4	0.00869
12	NFIL3		Hsapiens–hPDI–NFIL3	12.5	0.0126
13	BARHL2		Hsapiens–jolma2013–BARHL2	9.43	0.0137
14	HHEX		Hsapiens–hPDI–HHEX	14.8	0.014
15	RFC2		Hsapiens–hPDI–RFC2	12.7	0.0149
16	CNOT6		Hsapiens–hPDI–CNOT6	13	0.0154
17	RIOK2		Hsapiens–hPDI–RIOK2	13.9	0.0166
18	CBFB		Hsapiens–hPDI–CBFB	9.95	0.0179
19	BARHL2		Hsapiens–jolma2013–BARHL2–5	7.78	0.0184
20	MAGEF1		Hsapiens–hPDI–MAGEF1	7.7	0.0187
21	E2F2		Hsapiens–jolma2013–E2F2–3	0.0168	0.0188
22	CSTF2		Hsapiens–hPDI–CSTF2	39.3	0.019
23	BARHL2		Hsapiens–jolma2013–BARHL2–2	7.85	0.0195
24	SMAD2		Hsapiens–hPDI–SMAD2	9.01	0.0201
25	FOXD3		Hsapiens–jolma2013–FOXD3–2	11.4	0.0204
26	RNF138		Hsapiens–hPDI–RNF138	6.6	0.0224
27	TSNAX		Hsapiens–hPDI–TSNAX	14.8	0.0236
28	DUSP22		Hsapiens–hPDI–DUSP22	8.83	0.0239
29	SOX13		Hsapiens–hPDI–SOX13	8.73	0.0249
30	ZBTB43		Hsapiens–hPDI–ZBTB43	6.57	0.025
31	ESX1		Hsapiens–hPDI–ESX1	4.74	0.026
32	LARP1		Hsapiens–hPDI–LARP1	9.42	0.0285
33	GLTPD1		Hsapiens–hPDI–MGC10334	6.18	0.0291
34	NOTO		Hsapiens–jolma2013–NOTO	5.14	0.0329
35	CBX3		Hsapiens–hPDI–LOC653972	3.72	0.0351
36	HNRNPA0		Hsapiens–hPDI–HNRPA0	7.15	0.0365
37	HSF4		Hsapiens–jolma2013–HSF4	1.27	0.0394
38	MECP2		Hsapiens–hPDI–MECP2	4.53	0.0397
39	LMX1B		Hsapiens–jolma2013–LMX1B–2	3.99	0.0454
40	MEX3C		Hsapiens–hPDI–RKHD2	7.64	0.0461
41	ZNF3		Hsapiens–hPDI–ZNF3	6.38	0.0463
42	TFE3		Hsapiens–hPDI–TFE3	6.42	0.0464
43	TRIP10		Hsapiens–hPDI–TRIP10	4.81	0.0476
44	HSF1		Hsapiens–jolma2013–HSF1–2	0.041	0.0484
45	HOXD8		Hsapiens–jolma2013–HOXD8	6	0.0491
46	METTL21B		Hsapiens–hPDI–FAM119B	7.17	0.0498
47	STAT1		Hsapiens–JASPAR_2014–STAT1–MA0137.3	1.36	0.0508
48	ZFP3		Hsapiens–hPDI–ZFP3	4.11	0.0525
49	E2F1		Hsapiens–jolma2013–E2F1–4	0.154	0.0551
50	UTP18		Hsapiens–hPDI–UTP18	3.32	0.0573
51	FOXD2		Hsapiens–jolma2013–FOXD2–2	4.09	0.0601
52	NKX3–1		Hsapiens–JASPAR_CORE–NKX3–1–MA0124.1	2.23	0.0601
53	HSF1		Hsapiens–jolma2013–HSF1	0.9	0.064
54	SCAND2P		Hsapiens–hPDI–SCAND2	4.29	0.0644
55	EMX1		Hsapiens–jolma2013–EMX1	2.35	0.0682
56	ZRSR2		Hsapiens–hPDI–ZRSR2	4.19	0.0707
57	PLAGL1		Hsapiens–hPDI–PLAGL1	2.88	0.0709
58	NCALD		Hsapiens–hPDI–NCALD	1.43	0.0738
59	HSF2		Hsapiens–jolma2013–HSF2	0.0799	0.076
60	KIF22		Hsapiens–hPDI–KIF22	2.88	0.0762

Figure 19. Position Weight Matrices for all candidate transcription factors (p < 0.05), rs10518693, Derived

PWMs from PWMenrich for all transcription factors with p < 0.05 for binding to the queried sequence.

Figure 19 (Continued)

Position Weight Matrices for all candidate transcription factors (p < 0.05)

rs10518693
Derived

TGGTAAATGAAGTCTCTCTAAGAAT

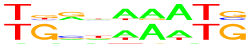
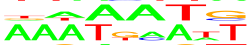
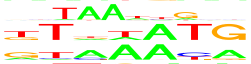
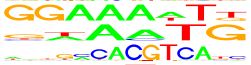
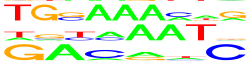
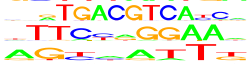
Rank	Target	PWM	Motif ID	Raw score	P–value
1	RBM22		Hsapiens–hPDI–RBM22	260	0.000213
2	MRPS25		Hsapiens–hPDI–MRPS25	160	0.000223
3	RFX4		Hsapiens–hPDI–RFX4	217	0.000968
4	ZNF830		Hsapiens–hPDI–CCDC16	101	0.00204
5	TRMT1		Hsapiens–hPDI–TRMT1	126	0.00211
6	DTL		Hsapiens–hPDI–DTL	40.1	0.00397
7	MYOD1		Hsapiens–hPDI–MYOD1	27.3	0.00465
8	ING3		Hsapiens–hPDI–ING3	8.51	0.00564
9	TCEAL2		Hsapiens–hPDI–TCEAL2	31	0.00585
10	DLX6		Hsapiens–hPDI–DLX6	21.4	0.00596
11	GLYCTK		Hsapiens–hPDI–GLYCTK	43.4	0.00663
12	ETS1		Hsapiens–hPDI–ETS1	15.9	0.00722
13	BARHL2		Hsapiens–jolma2013–BARHL2–4	12.4	0.00869
14	NFIL3		Hsapiens–hPDI–NFIL3	12.6	0.0125
15	BARHL2		Hsapiens–jolma2013–BARHL2	9.43	0.0137
16	HHEX		Hsapiens–hPDI–HHEX	14.9	0.0139
17	RFC2		Hsapiens–hPDI–RFC2	12.7	0.0149
18	CNOT6		Hsapiens–hPDI–CNOT6	13	0.0154
19	RIOK2		Hsapiens–hPDI–RIOK2	14	0.0164
20	CBFB		Hsapiens–hPDI–CBFB	9.89	0.0181
21	BARHL2		Hsapiens–jolma2013–BARHL2–5	7.78	0.0184
22	MAGEF1		Hsapiens–hPDI–MAGEF1	7.69	0.0187
23	CSTF2		Hsapiens–hPDI–CSTF2	39.4	0.019
24	BARHL2		Hsapiens–jolma2013–BARHL2–2	7.86	0.0195
25	SMAD2		Hsapiens–hPDI–SMAD2	9.01	0.0201
26	FOXD3		Hsapiens–jolma2013–FOXD3–2	11.4	0.0204
27	RNF138		Hsapiens–hPDI–RNF138	6.6	0.0223
28	DUSP22		Hsapiens–hPDI–DUSP22	8.85	0.0238
29	TSNAX		Hsapiens–hPDI–TSNAX	14.4	0.0243
30	ZBTB43		Hsapiens–hPDI–ZBTB43	6.61	0.0248
31	WDR83		Hsapiens–hPDI–MORG1	8.46	0.0265
32	FOSL1		Hsapiens–hPDI–FOSL1	8.09	0.0283
33	LARP1		Hsapiens–hPDI–LARP1	9.42	0.0284
34	MEF2BNB–MEF2B		Hsapiens–hPDI–MEF2B	12.2	0.0297
35	ESX1		Hsapiens–hPDI–ESX1	4.46	0.0299
36	EDN1		Hsapiens–hPDI–EDN1	9.48	0.0299
37	NOTO		Hsapiens–jolma2013–NOTO	5.14	0.0329
38	CREB3		Hsapiens–jolma2013–CREB3–2	0.206	0.0333
39	HNRNPA0		Hsapiens–hPDI–HNRPA0	7.28	0.0357
40	MECP2		Hsapiens–hPDI–MECP2	4.53	0.0396
41	XBP1		Hsapiens–jolma2013–XBP1–2	0.1	0.0416
42	LMX1B		Hsapiens–jolma2013–LMX1B–2	3.99	0.0455
43	MEX3C		Hsapiens–hPDI–RKHD2	7.65	0.0461
44	ZNF3		Hsapiens–hPDI–ZNF3	6.41	0.0461
45	TFE3		Hsapiens–hPDI–TFE3	6.42	0.0464
46	TRIP10		Hsapiens–hPDI–TRIP10	4.83	0.0473
47	BAX		Hsapiens–hPDI–BAX	2.82	0.0476
48	METTL21B		Hsapiens–hPDI–FAM119B	7.28	0.0488
49	HOXD8		Hsapiens–jolma2013–HOXD8	6.02	0.049
50	MAGOH		Hsapiens–hPDI–MAGOH	3.26	0.0494
51	CREB3		Hsapiens–jolma2013–CREB3	0.0374	0.0502
52	STAT1		Hsapiens–JASPAR_2014–STAT1–MA0137.3	1.36	0.0508
53	ZFP3		Hsapiens–hPDI–ZFP3	4.12	0.0524
54	PPP1R10		Hsapiens–hPDI–PPP1R10	3.31	0.0532
55	HSF4		Hsapiens–jolma2013–HSF4	0.778	0.0543
56	FOXC2		Hsapiens–jolma2013–FOXC2	2.9	0.0552
57	UTP18		Hsapiens–hPDI–UTP18	3.31	0.0574
58	NKX3–1		Hsapiens–JASPAR_CORE–NKX3–1–MA0124.1	2.26	0.0595
59	XBP1		Hsapiens–jolma2013–XBP1	0.0937	0.0601
60	FOXD2		Hsapiens–jolma2013–FOXD2–2	4.09	0.0601

Table 11. Strongest candidates for binding rs66791338

Ancestral, Insertion binding candidates in red
Derived, Deletion binding candidates in green

		Enrichment			Differential Expression	
Gene Name	**Raw Score**	**-log10(p)**	**Differential Score**	**Differential Rank**	**Congruent Count**	**Incongruent Count**
RARG	98.75	2.81	4.69	4	0	2
RARA	1.56	2.24	4.39	5	0	3
TCF7L2	1354.25	3.88	8.03	1	2	7
LEF1	580.52	3.78	7.68	2	2	5
SPATS2	6.01	1.52	2.75	21	2	5
RFC3	14.60	1.90	2.40	29	4	13
STAT2	4.03	1.30	1.59	58	0	4
ELK1	3.85	1.36	0.92	104	0	3
SMUG1	29.94	3.04	4.5240	1	3	0
SFT2D1	8.67	1.52	2.8177	7	5	0
PAX5	3.99	2.14	4.4484	2	2	0
TFAM	19.04	1.89	3.05	6	7	2
AVEN	7.64	1.44	1.9697	18	3	0
ETFB	4.92	1.34	0.3624	116	11	0

Table 12. Strongest candidates for binding rs10518693

Ancestral binding candidates in red
Derived binding candidates in green

			Enrichment		Differential Expression	
Gene Name	**Raw Score**	**-log10(p)**	**Differential Score**	**Differential Rank**	**Congruent Count**	**Incongruent Count**
PPP1R10	55.73	3.24	4.52	1	1	0
E2F2	0.02	1.73	2.37	4	4	6
SOX13	8.73	1.60	2.81	3	1	4
GLTPD1	6.18	1.54	2.95	2	0	0
CBX3	3.72	1.46	2.28	5	4	1
ING3	8.51	2.25	3.77	1	5	4
GLYCTK	43.37	2.18	2.70	4	2	0
ETS1	15.89	2.14	3.44	2	6	9
WDR83	8.46	1.58	1.98	10	2	0
FOSL1	8.09	1.55	2.04	8	0	1
MEF2B	12.23	1.53	2.31	6	0	0
EDN1	9.48	1.52	2.38	5	1	7
CREB3	0.21	1.48	2.89	3	5	2
XBP1	0.10	1.38	1.65	19	1	1

Figure 20. Position Weight Matrices for best candidate transcription factors
Best candidate transcription factors by differential binding analysis for (A) binding to the rs66791338 deletion sequence over the insertion sequence; (B) binding to the rs66791338 insertion sequence over the deletion sequence.

Figure 20 (Continued)

Position Weight Matrices for best candidate transcription factors

(a)

rs66791338
Deletion

ATAAGGACATGGAAGAGGGGTTG

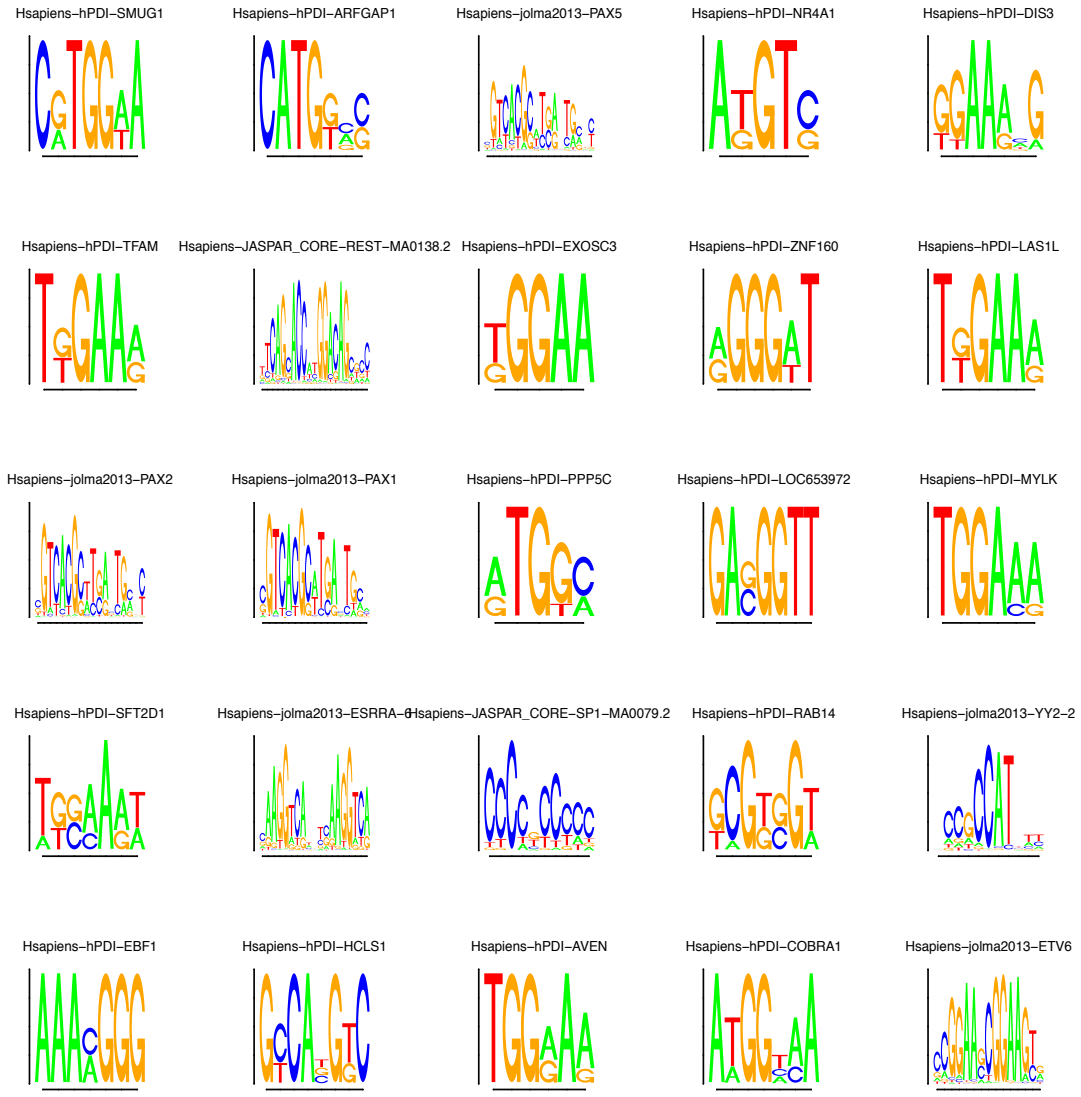


Figure 20 (Continued)

Position Weight Matrices for best candidate transcription factors

(b)

rs66791338
Insertion

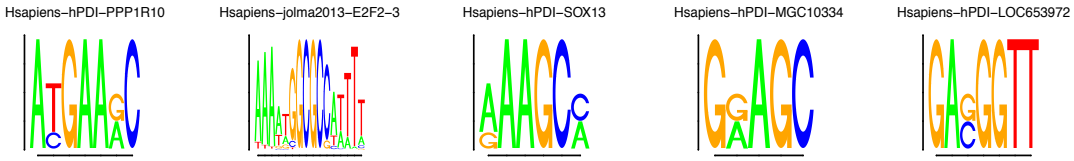
ATAAGGACATGAAAGGGAAGAGGGGTTG

Position Weight Matrices for best candidate transcription factors

(a)

rs10518693
Ancestral

TGGTAAATGAAGCCTCTCTAAGAAT



(b)

rs10518693
Derived

TGGTAAATGAAGTCTCTCTAAGAAT

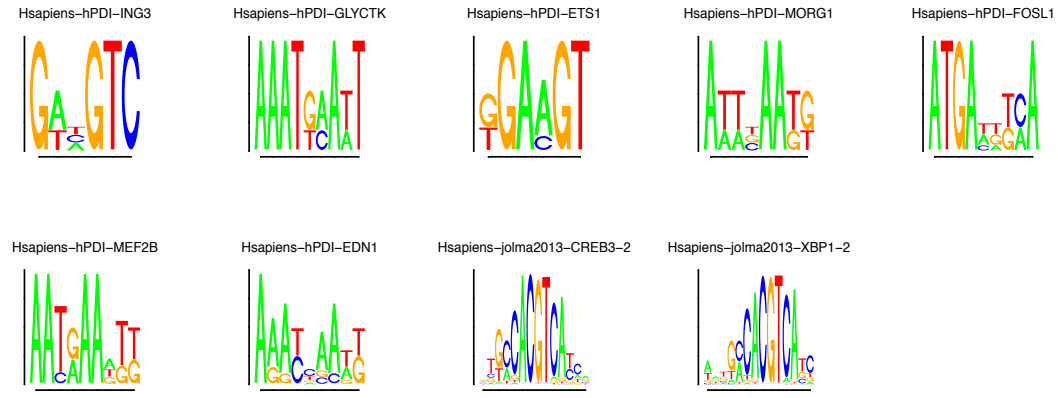


Figure 21. Best candidate transcription factors by differential binding analysis for (A) binding to the rs10518693 ancestral sequence over the derived sequence; (B) binding to the rs10516893 derived sequence over the ancestral sequence.

Figure 22. Population haplotypes for 150-kb region

Haplotypes for all variants with MAF > 0.05 for 1000 genomes phase 3 populations, beginning with East Asians: (A) CDX—Chinese Dai in Xishuangbanna, China (B) CHB—Han Chinese in Bejing, China (C) Souther Han Chinese (D) JPT—Japanese in Tokyo, Japan (E) KHV—Kinh in Ho Chi Minh City, Vietnam. Then, South and Southeast Asians: (F) BEB—Bengali in Bangladesh (G) GIH—Gujarati Indian in Houston, TX (H) ITU—Indian Telugu in the UK (I) PJL—Punjabi in Lahore, Pakistan (J) STU—Sri Lankan Tamil in the UK. Then, Europeans: (K) CEU—Utah residents with Northern and Western European ancestry (L) GBR—British in England and Scotland (M) FIN—Finnish in Finland (N) IBS—Iberian populations in Spain (O) TSI—Toscani in Italy. Then, admixed populations in the Americas: (P) CLM—Colombian in Medellin, Colombia (Q) MXL—Mexican in Los Angeles, California (R) PEL—Peruvian in Lima, Peru (S) PUR—Puerto Rican in Puerto Rico (T) ACB—African Caribbean in Barbados (U) ASW—African Ancestry in Southwest US. And, finally African populations: (V) ESN—Esan in Nigeria (W) GWD—Gambian in Western Division, the Gambia (X) LWK—Luhya in Webuye, Kenya (Y) MSL—Mende in Sierra Leone (Z) YRI—Yoruba in Ibadan, Nigeria.

Figure 22 (Continued)

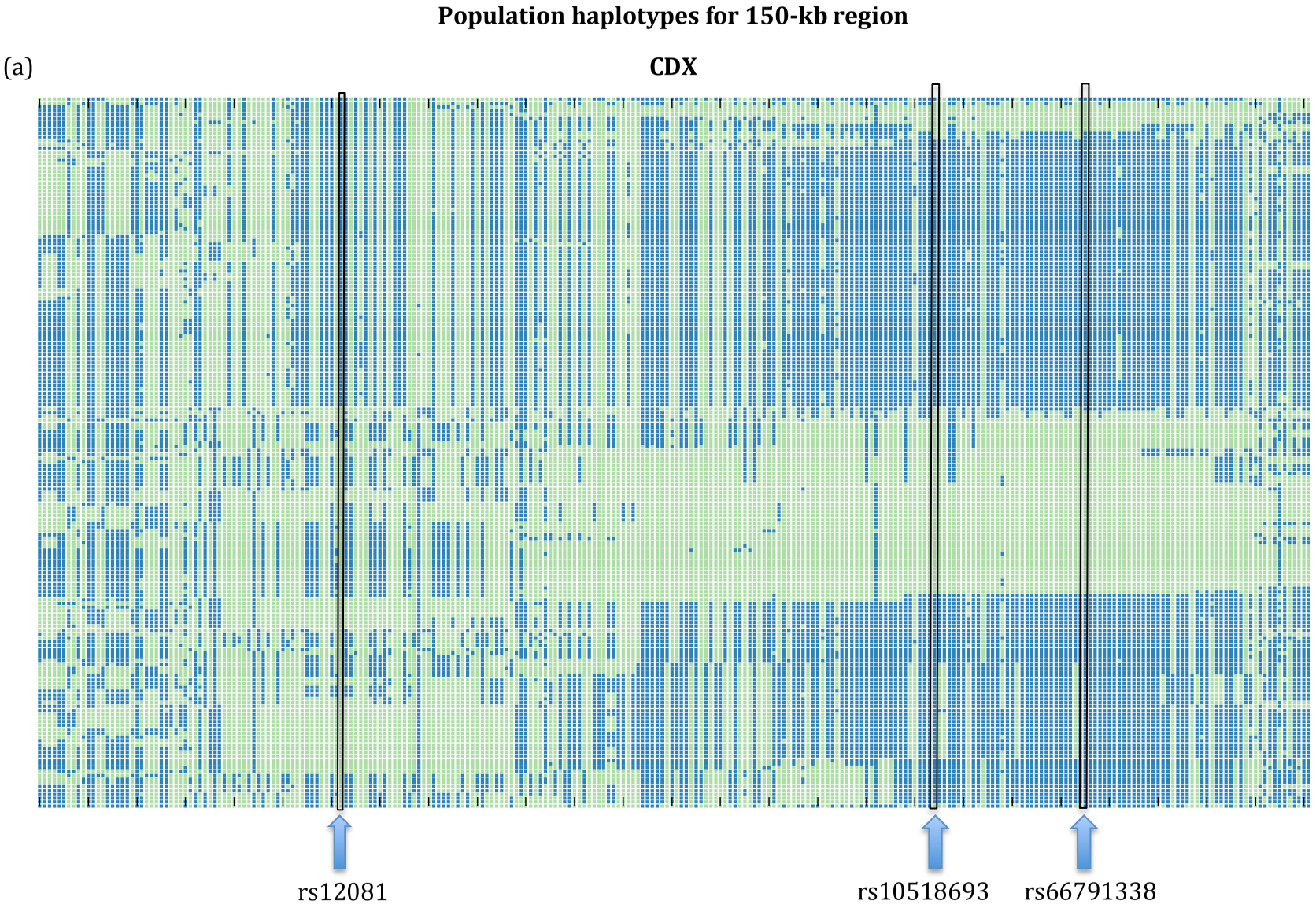


Figure 22 (Continued)

Population haplotypes for 150-kb region

(b)

CHB

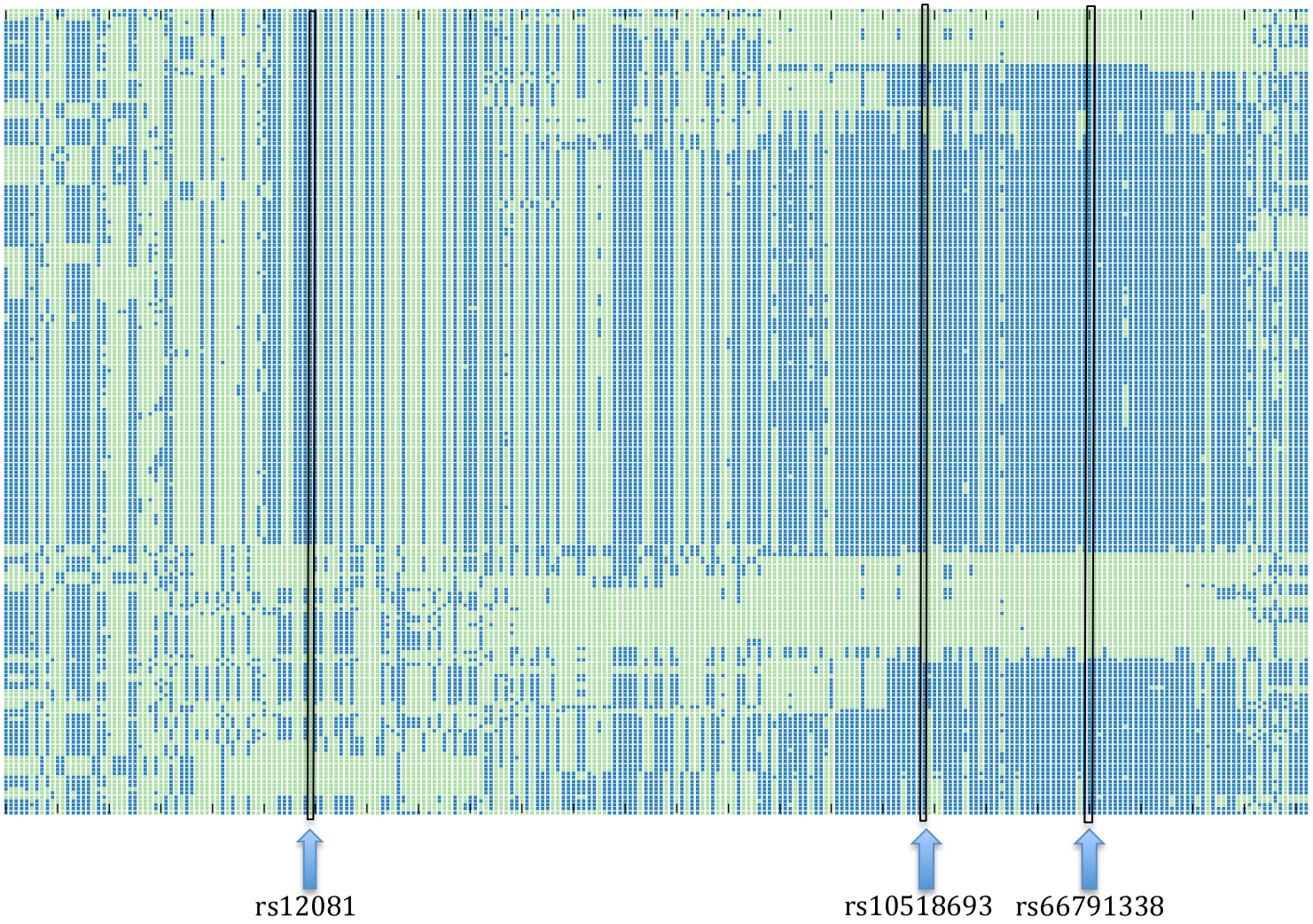


Figure 22 (Continued)

Population haplotypes for 150-kb region

(c)

CHS

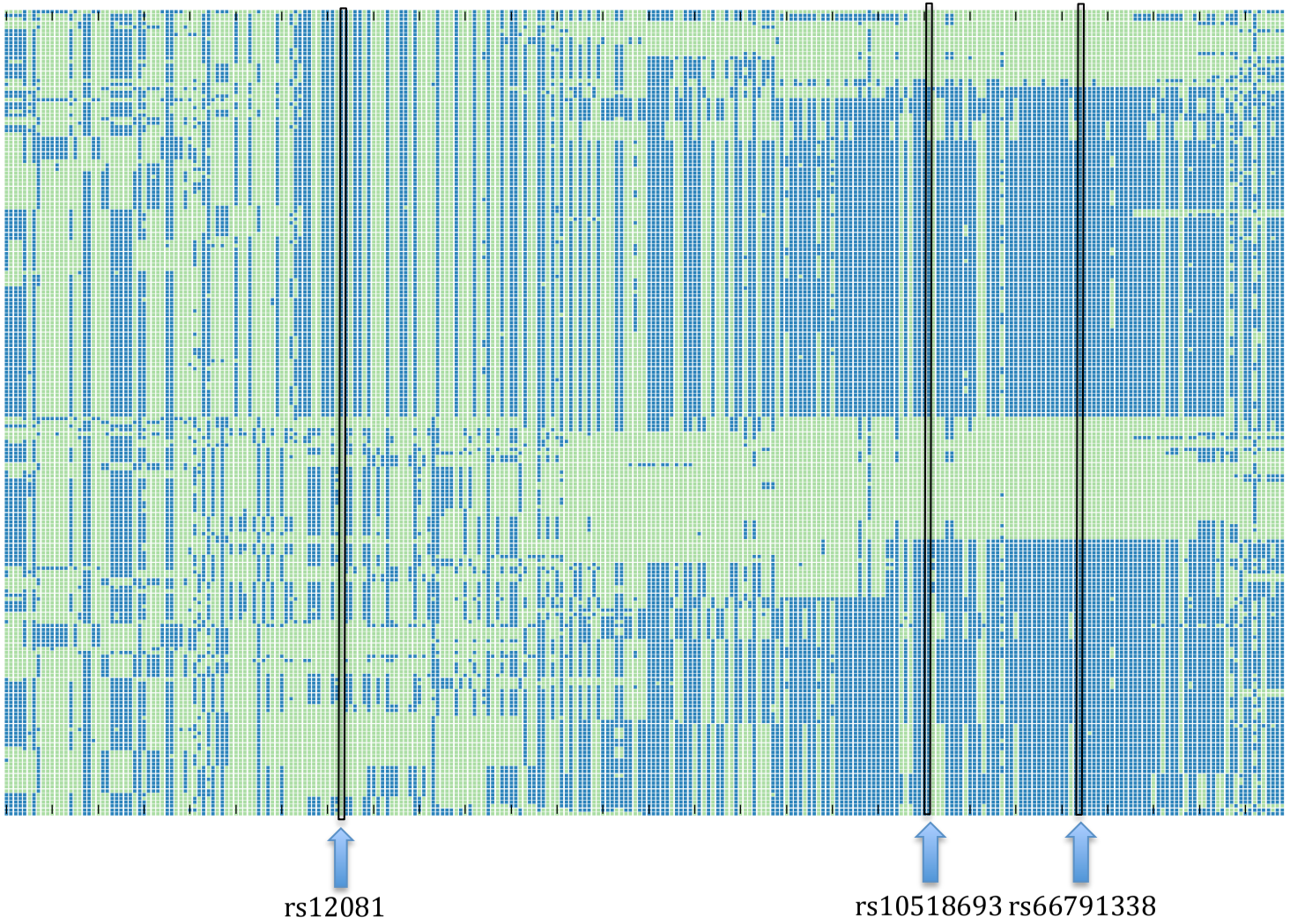


Figure 22 (Continued)

Figure 22 (Continued)

Population haplotypes for 150-kb region

(e) **KHV**

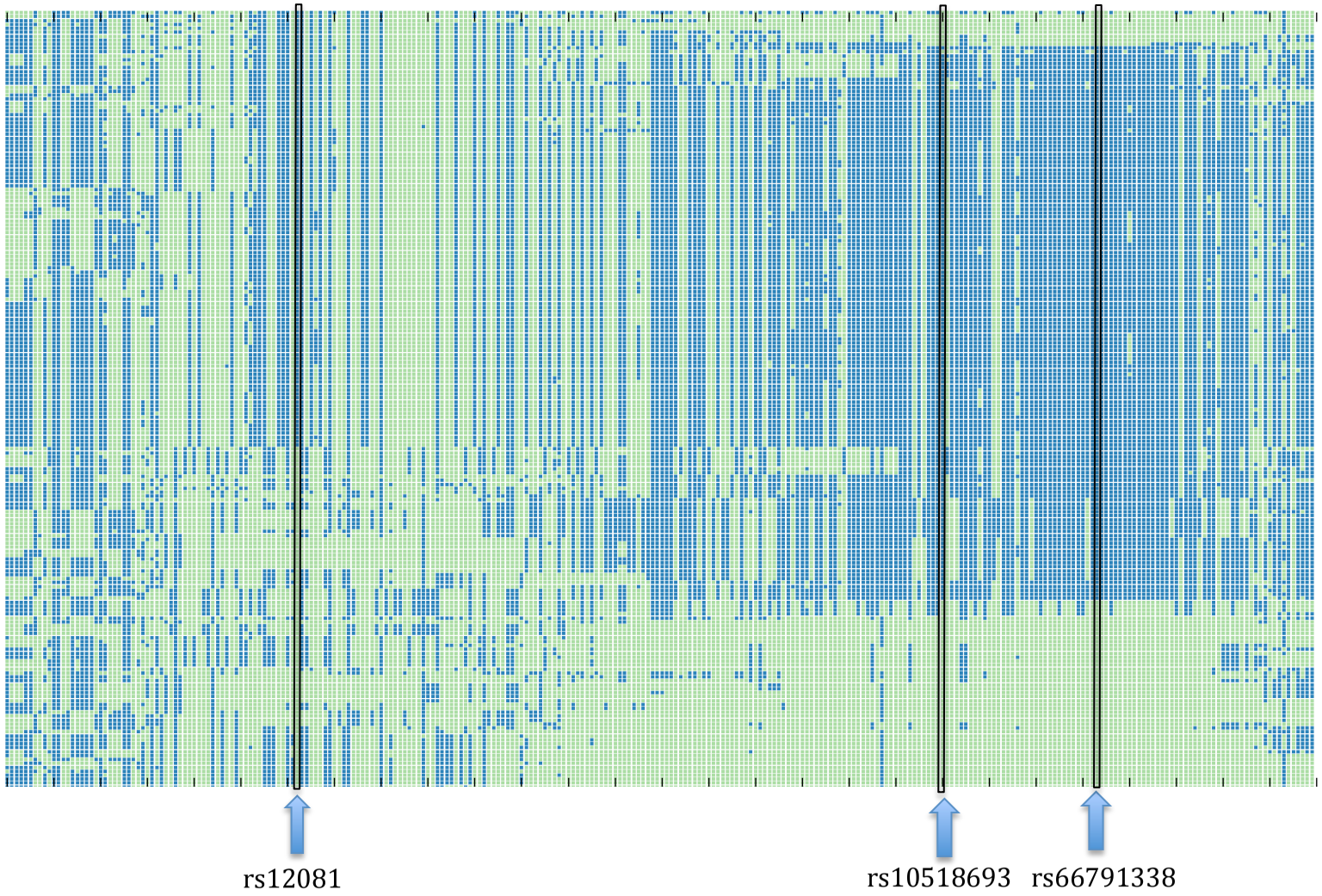


Figure 22 (Continued)

Population haplotypes for 150-kb region

(f) **BEB**

Figure 22 (Continued)

Population haplotypes for 150-kb region

(g)

GIH

Figure 22 (Continued)

Population haplotypes for 150-kb region

(h) **ITU**

Figure 22 (Continued)

Population haplotypes for 150-kb region

(i) **PJL**

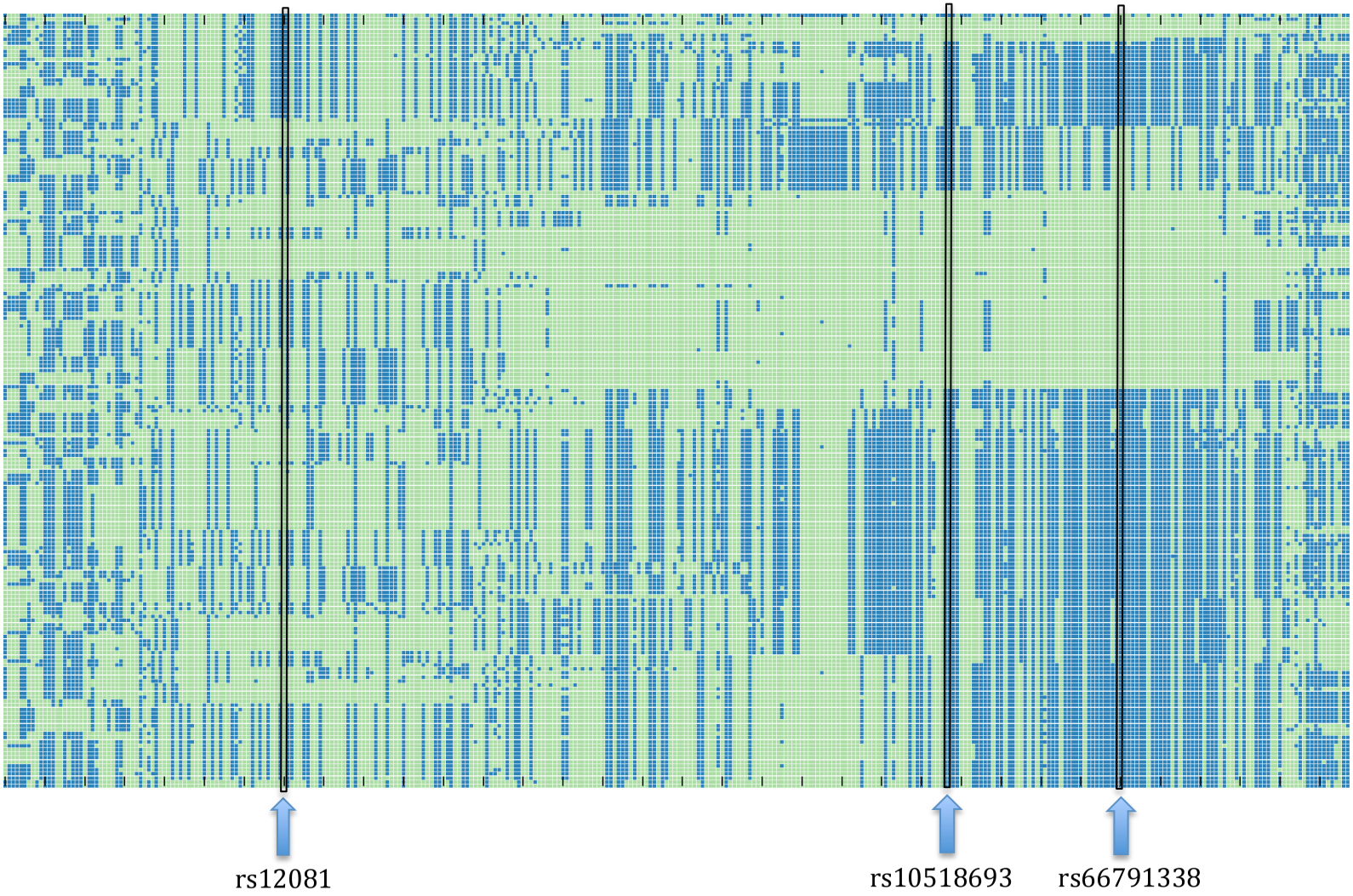


Figure 22 (Continued)

Population haplotypes for 150-kb region

(j) **STU**

Figure 22 (Continued)

Population haplotypes for 150-kb region

(k)

CEU

Figure 22 (Continued)

Population haplotypes for 150-kb region

(l) **GBR**

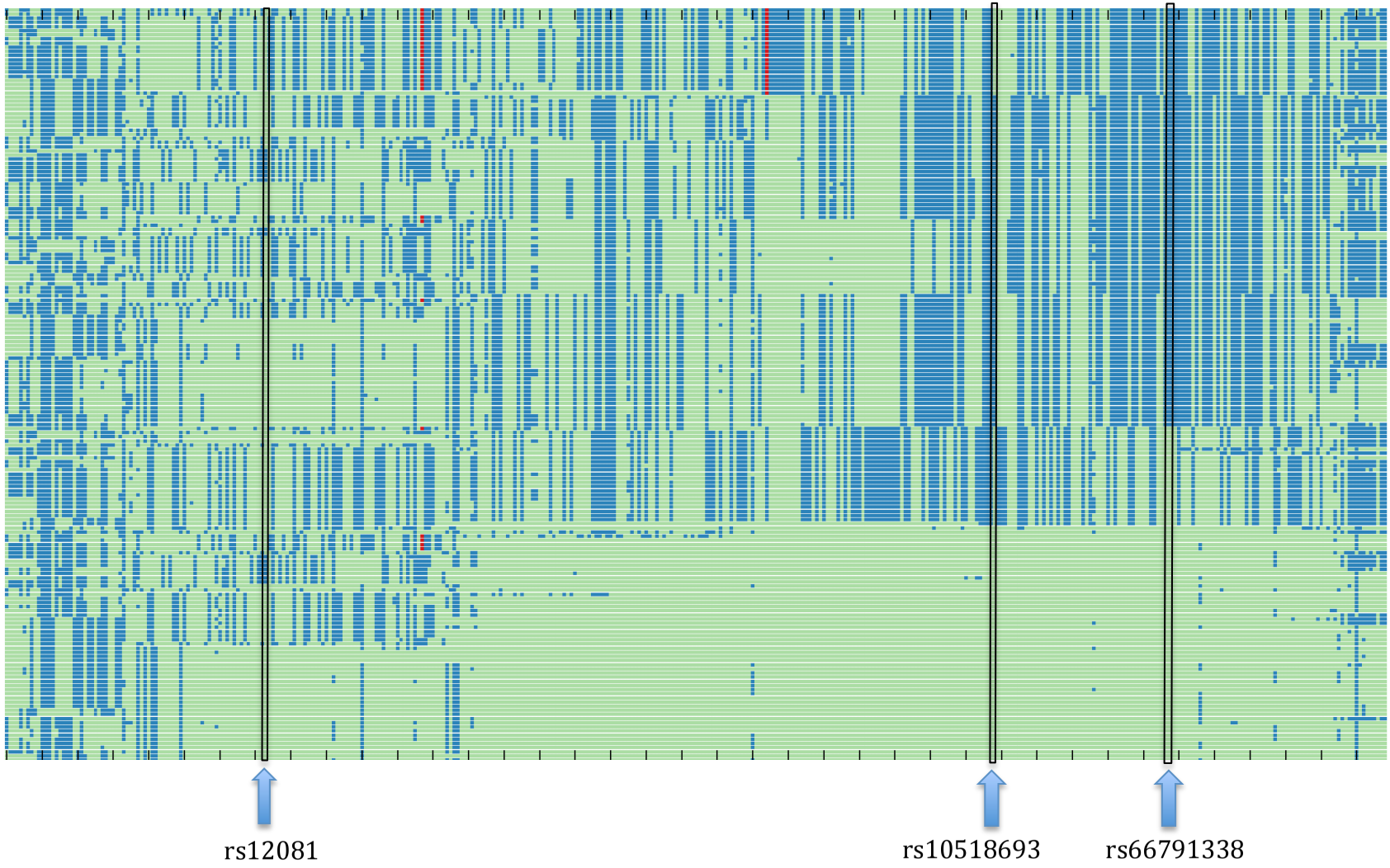


Figure 22 (Continued)

Population haplotypes for 150-kb region

(m) **FIN**

Figure 22 (Continued)

Population haplotypes for 150-kb region

(n) **IBS**

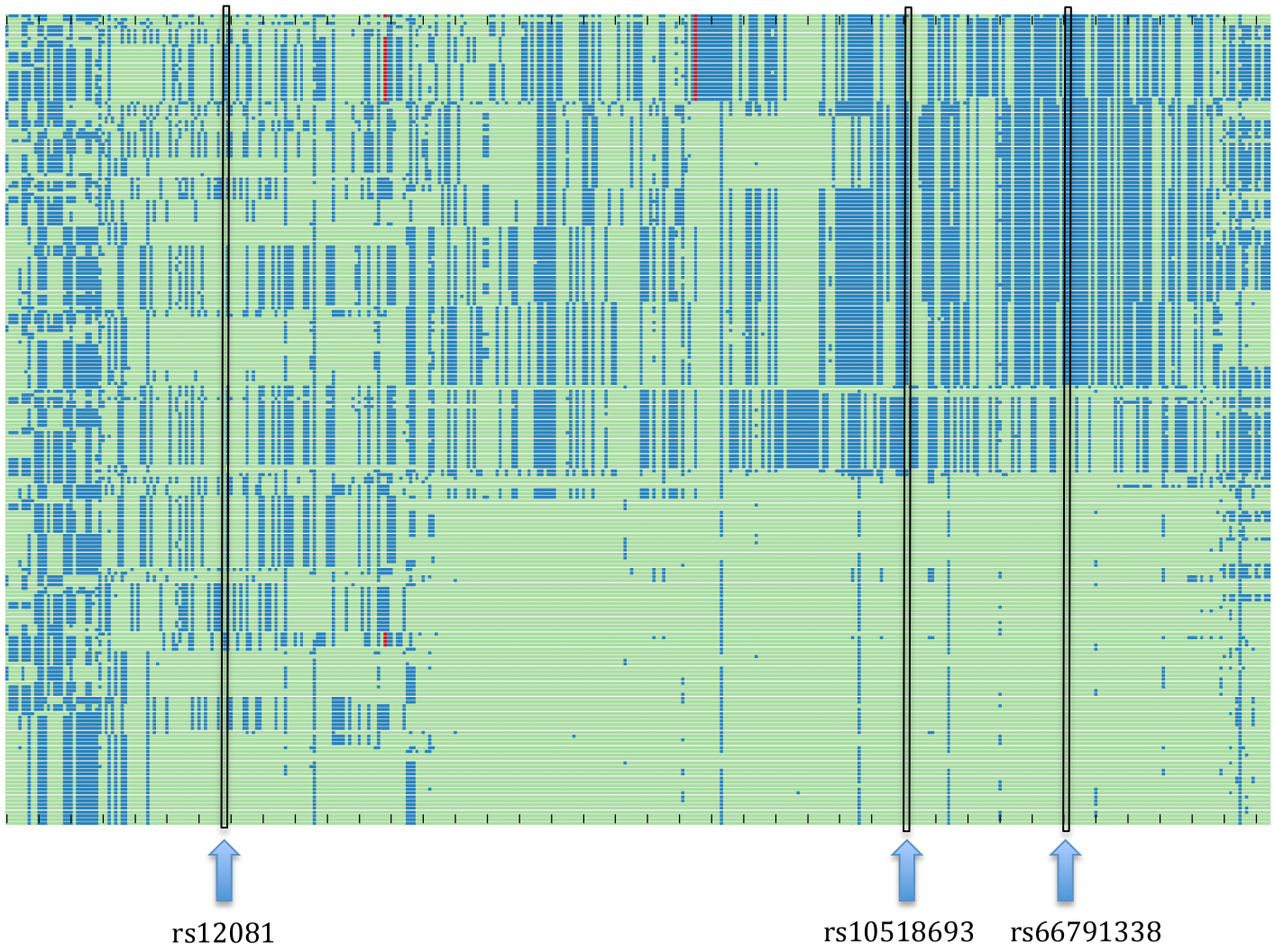


Figure 22 (Continued)

Population haplotypes for 150-kb region

(o) **TSI**

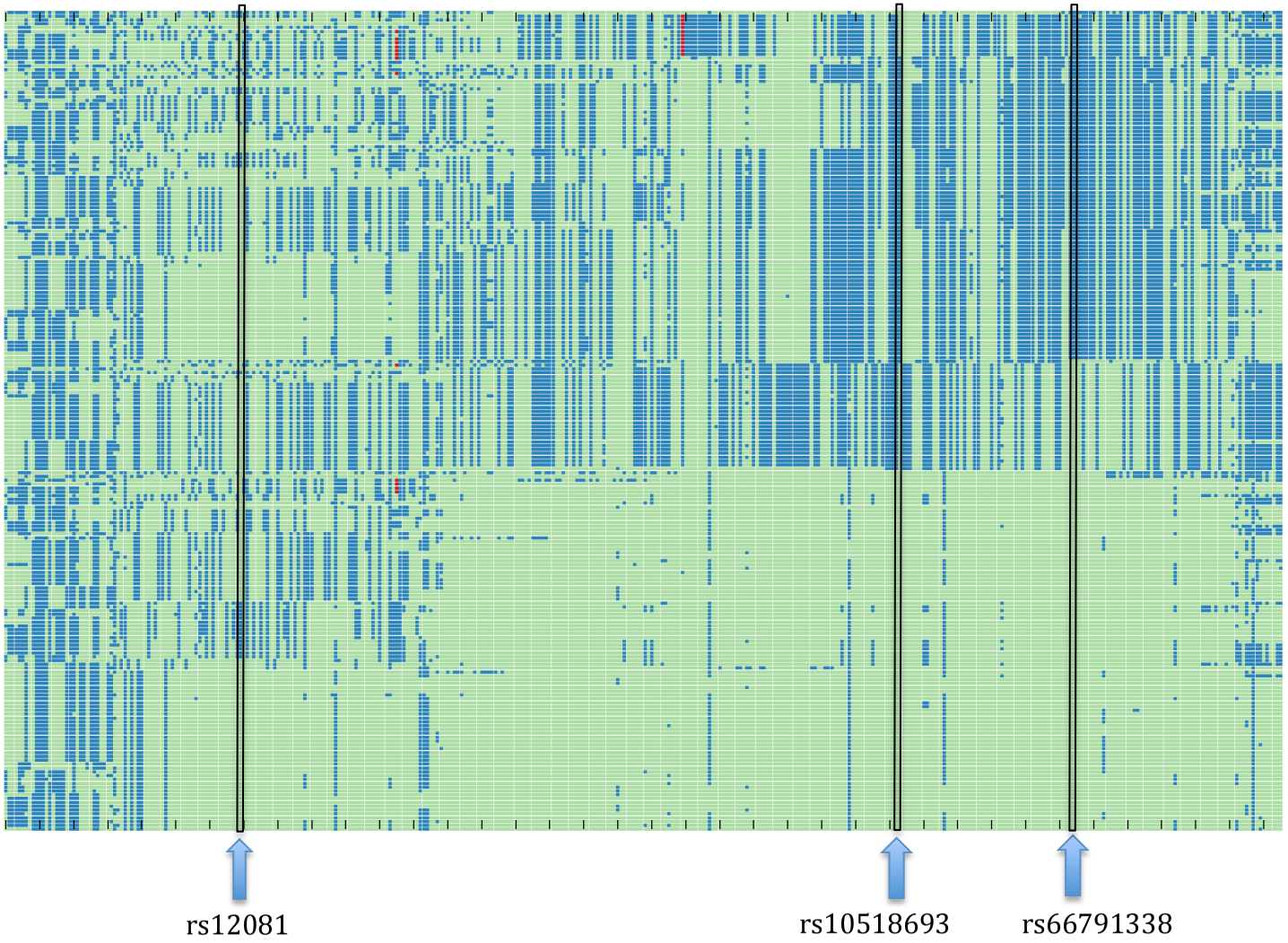


Figure 22 (Continued)

Population haplotypes for 150-kb region

(p) **CLM**

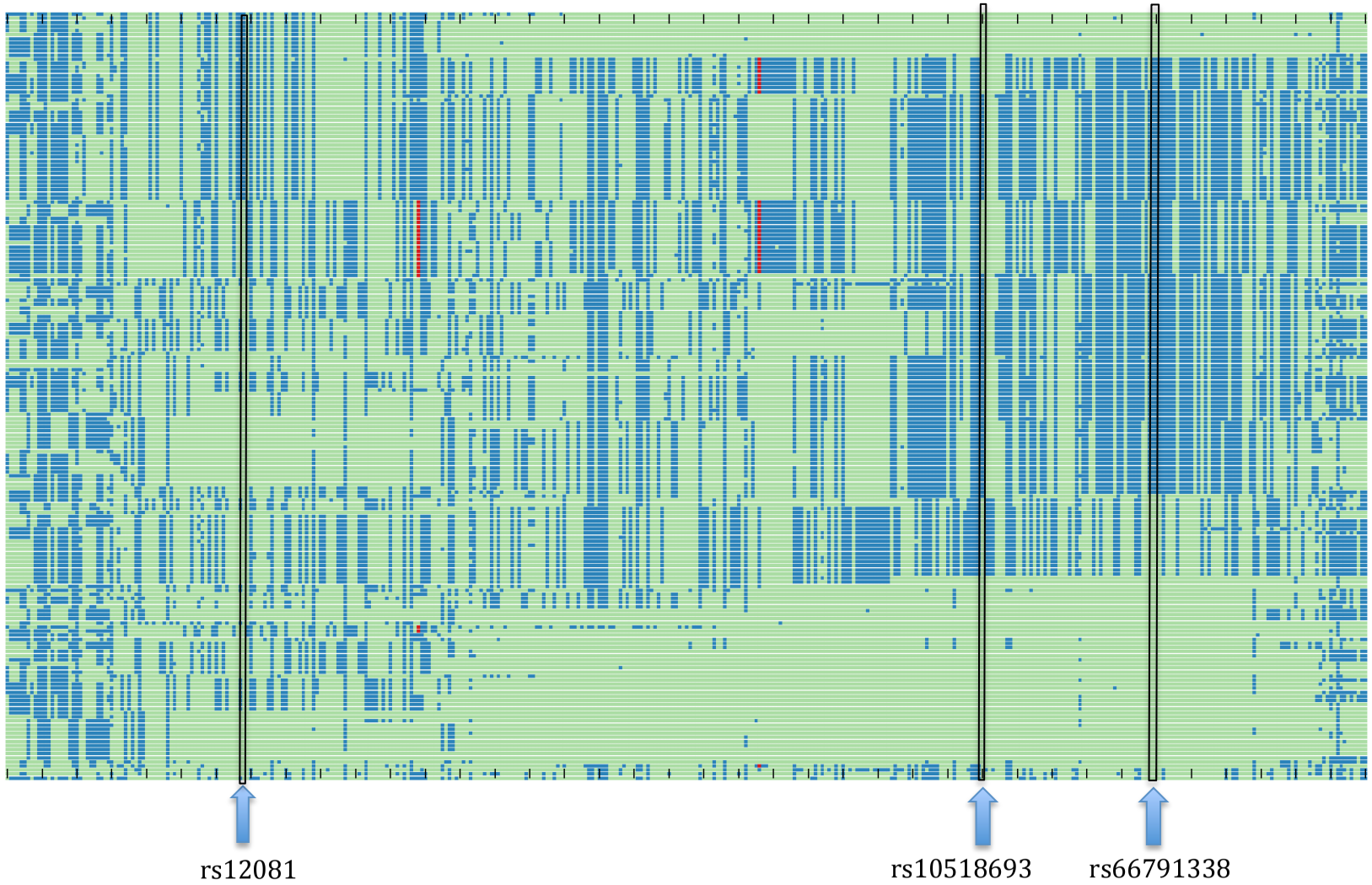


Figure 22 (Continued)

Population haplotypes for 150-kb region

(q) **MXL**

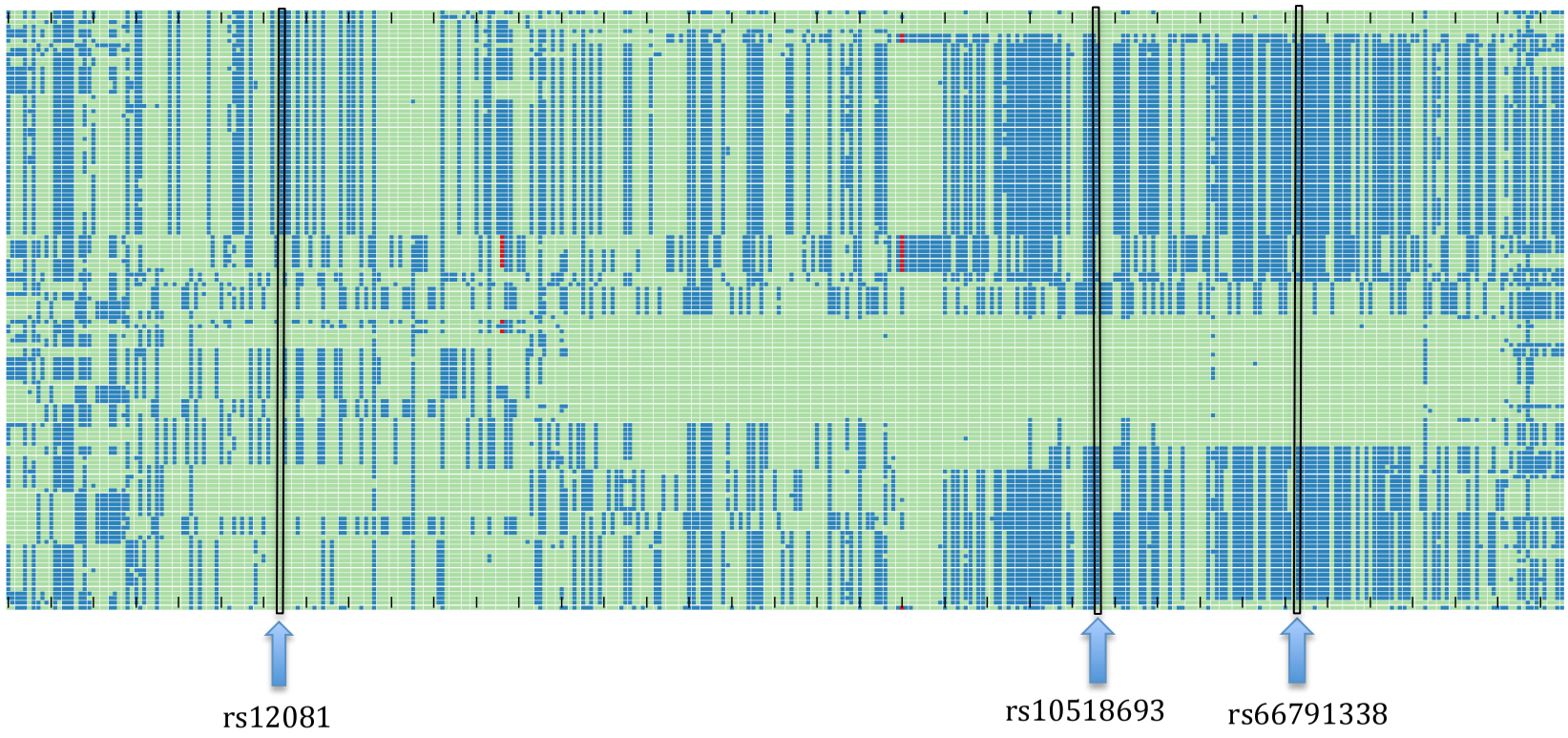


Figure 22 (Continued)

Population haplotypes for 150-kb region

(r) **PEL**

Figure 22 (Continued)

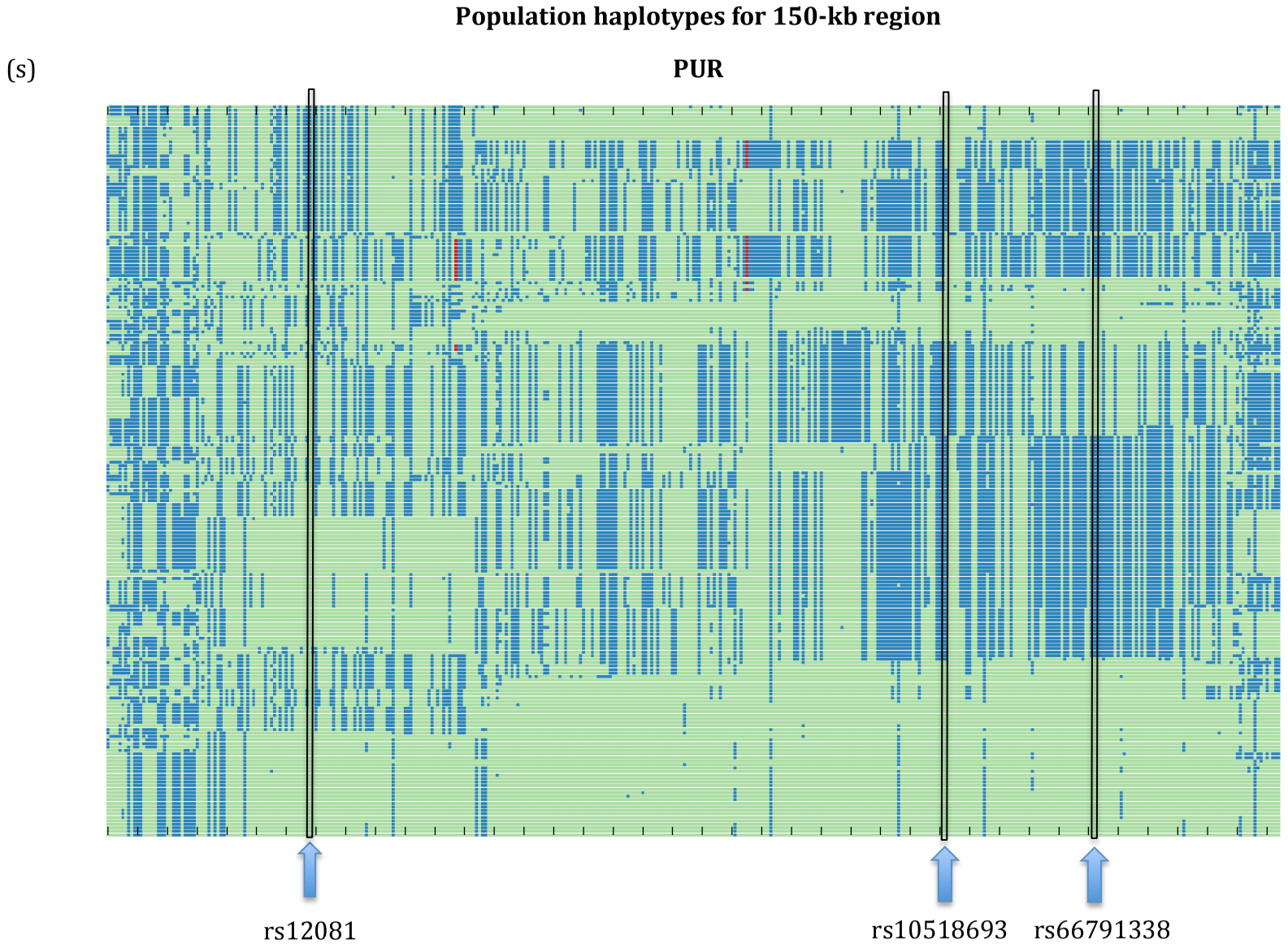


Figure 22 (Continued)

Population haplotypes for 150-kb region

(t) **ACB**

Figure 22 (Continued)

Population haplotypes for 150-kb region

(u) **ASW**

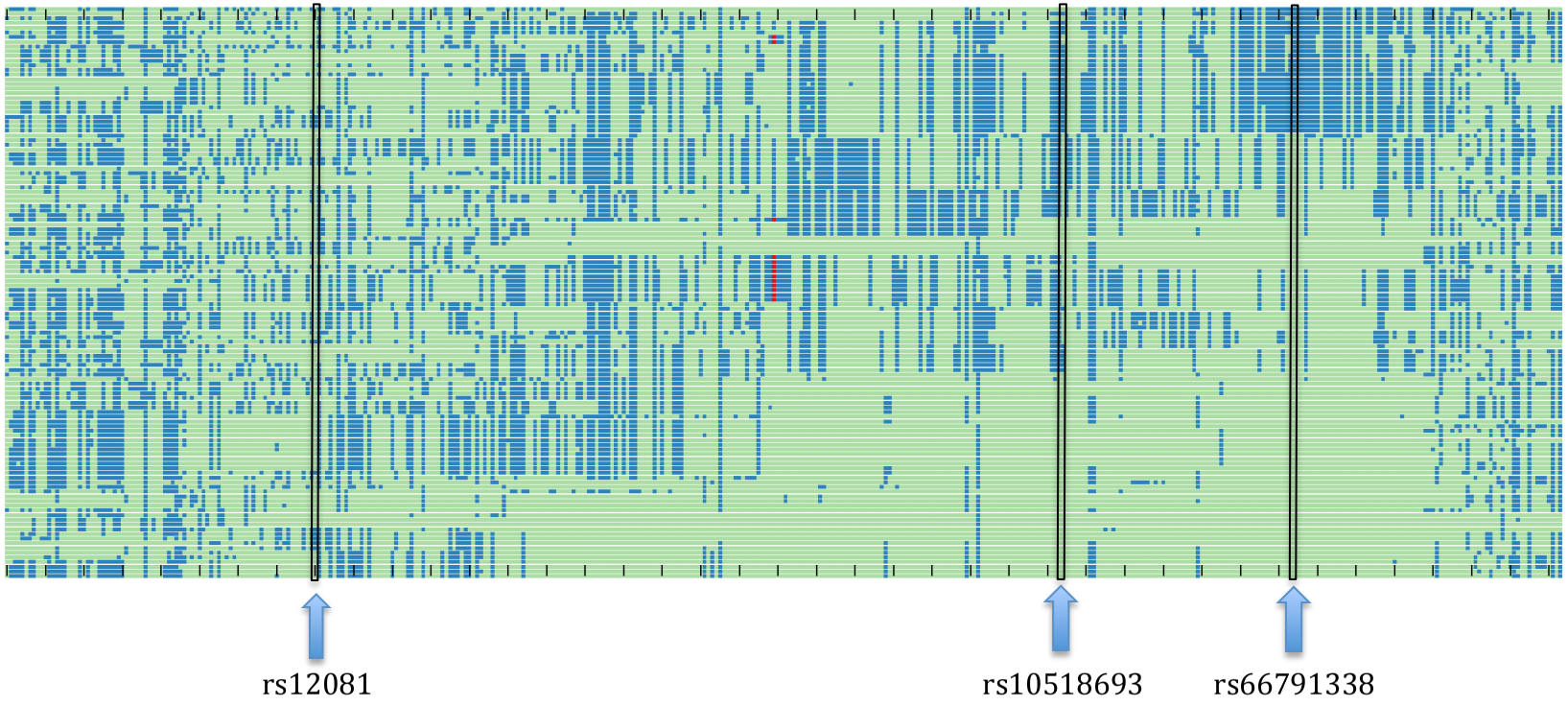


Figure 22 (Continued)

Population haplotypes for 150-kb region

(v) **ESN**

Figure 22 (Continued)

Population haplotypes for 150-kb region

(w) **GWD**

Figure 22 (Continued)

Population haplotypes for 150-kb region

(x) **LWK**

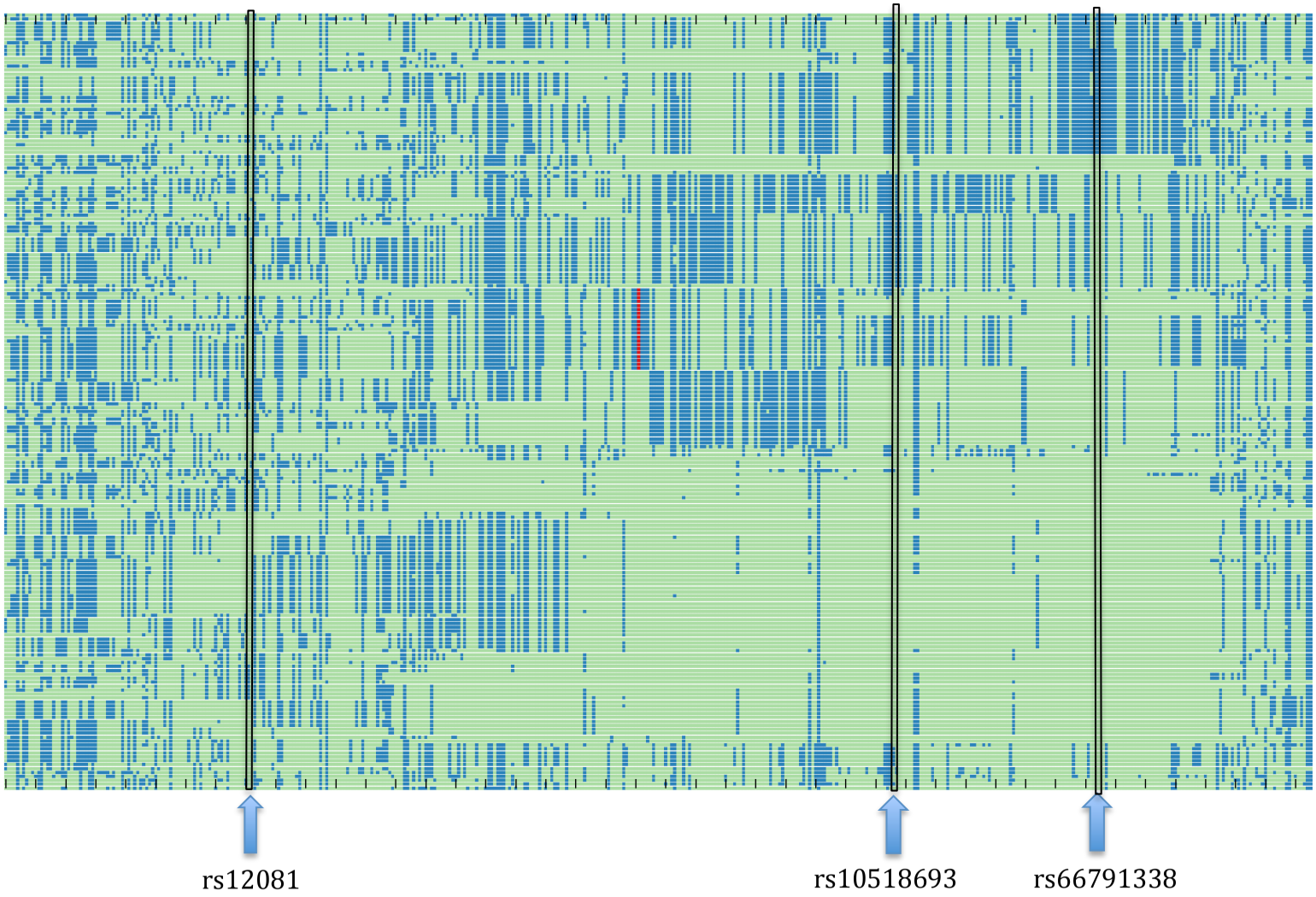


Figure 22 (Continued)

Population haplotypes for 150-kb region

(y) **MSL**

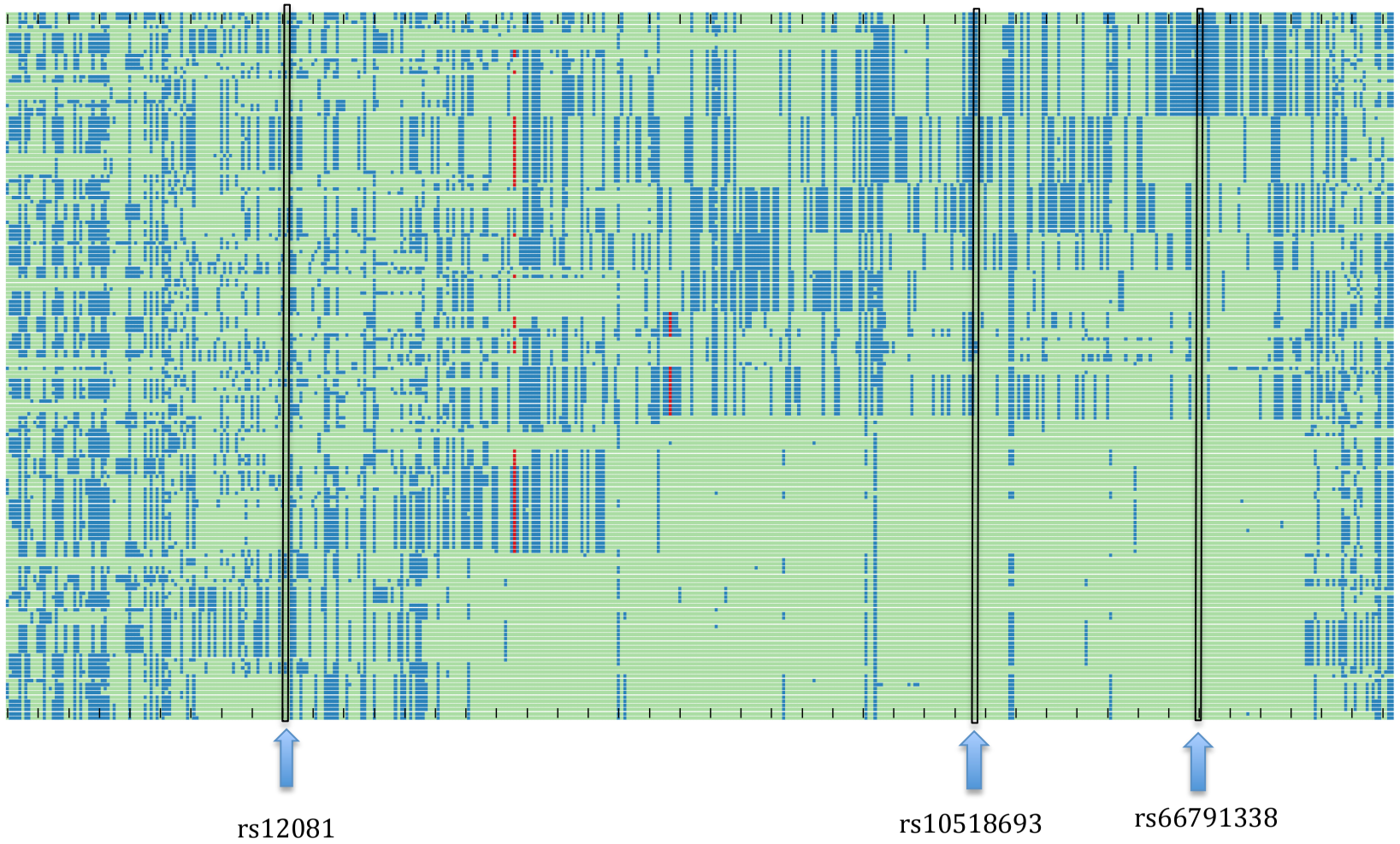


Figure 22 (Continued)

visible from a bifurcation plot centered on the highest-ranking CMS SNP, rs12081, for the region (Figure 23). Derived alleles of the functional variants, rs66791338 and rs10518693 tag the long haplotype, though for each of these variants a fraction of derived alleles also occur outside of the identified long haplotype. Analysis of linkage disequilibrium in Haploview also reveals a local block of linkage disequilibrium in East Asians that encompasses the two functional variants (Figure 24).

Frequencies of the functional regulatory variants, rs10518693 and rs66791338, were plotted for 1000 genomes phase 3 populations on a world map to show the variation in their co-occurrence on the same haplotypes in diverse populations (Figure 25). This reveals that in East Asians, and also South Asians, where both derived alleles are at highest frequency, they also co-occur with the highest frequency.

Finally, the relationship between the CMS signature of selection in the region and the regulatory functional variants was examined (Table 13). The best proxy for rs66791338 in the region is rs11637756, with similar frequencies and r^2 close to 1.0 in diverse populations. Its rank in CMS is 53. The CMS rank of rs10518693 is 91. The CMS score for rs11633883, which is 30 bp from rs66791338 and may modulate its effect the 23rd highest for the region. Frequency differentiation and high-frequency derived alleles are the strongest signatures indicating selection at these loci.

Allele dating using haplotype analysis gives two age estimates for the derived functional alleles from the haplotypes to the left and right of the variants. The age estimates for the derived alleles using African haplotypes are 214,000 and 191,000 years for rs10518693 and 79,000 and 166,000 years for rs66791338. Since the derived alleles are present in all populations worldwide, the age estimates are in line with their having arisen

Bifurcator plot for 150-kb region in JPT

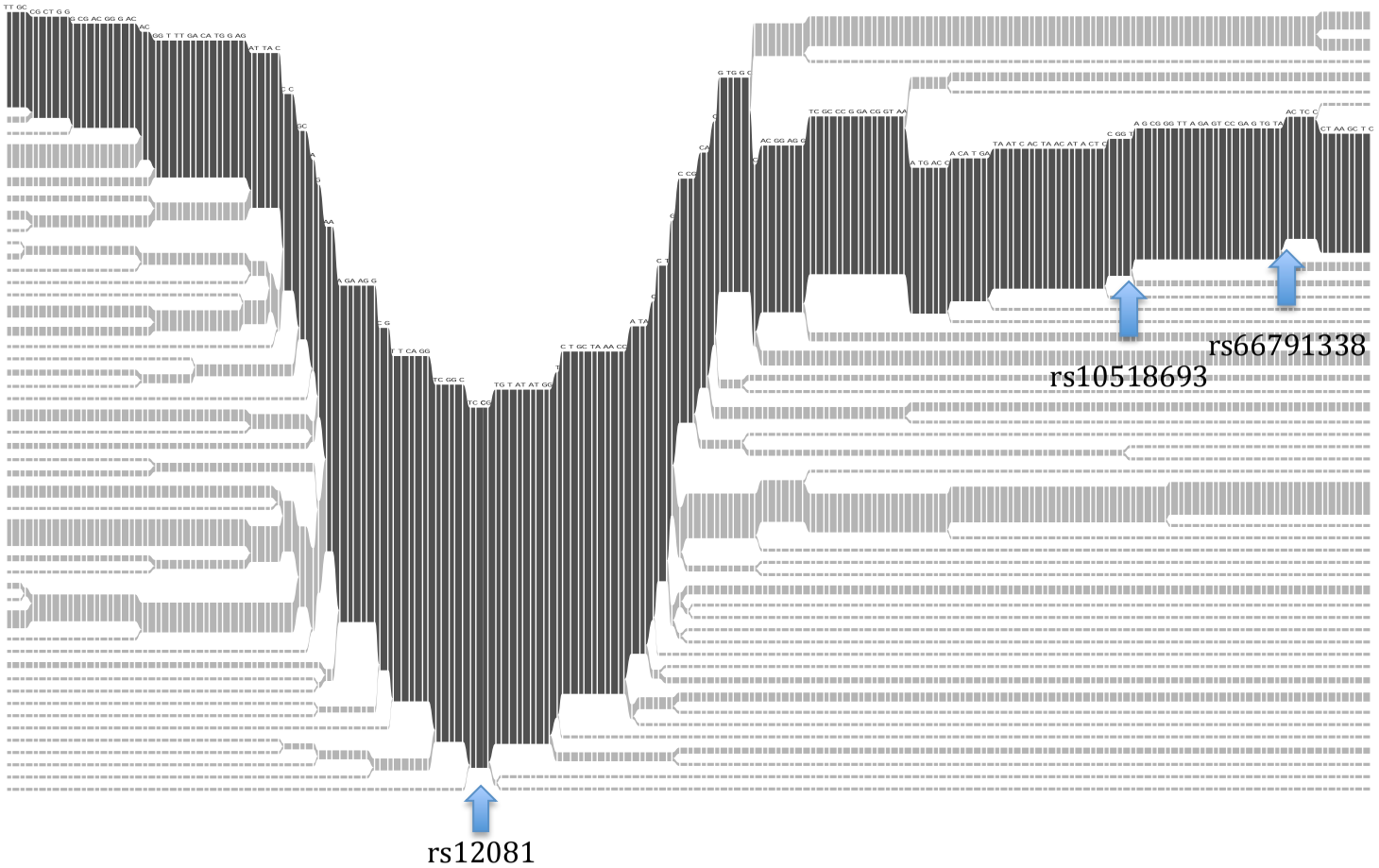


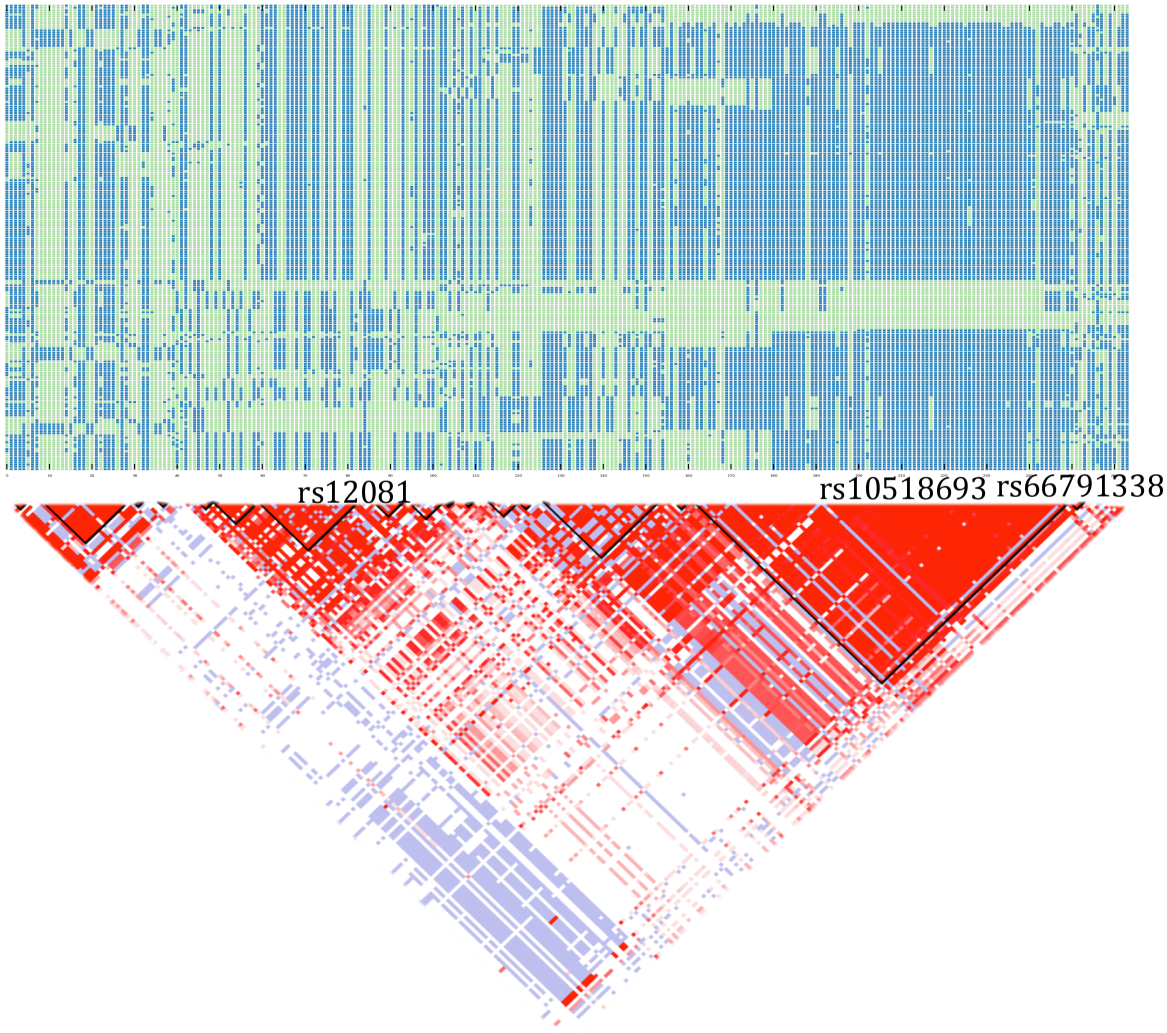
Figure 23. A bifurcator plot produced by the Bifurcator program (Ben Fry) centered around rs12081 the highest ranking SNP for CMS in the region, which shows the derived variants for rs10518693 and rs66791338 falling on the long haplotype. Shown in the JPT 1000 genomes, phase 3 population.

Figure 24. Block of linkage disequilibrium surrounds functional variants
(A) Haploview linkage disequilibrium (LD) plot shown for JPT haplotypes in the 150-kb region reveals block of LD surrounding the locations of the functional variants rs10518693 and rs66791338. (B) Expanded view of the block of linkage disequilibrium surrounding the functional variants (marked with arrows) on East Asian haplotypes. Less linkage disequilibrium is evident in haploview plots for the same region in South Asians, GIH shown (C), Europeans, GBR shown (D), and Africans, YRI shown (E).

Figure 24 (Continued)

Block of linkage disequilibrium surrounds functional variants

(a) JPT haplotypes for 150-kb region and haploview LD plot



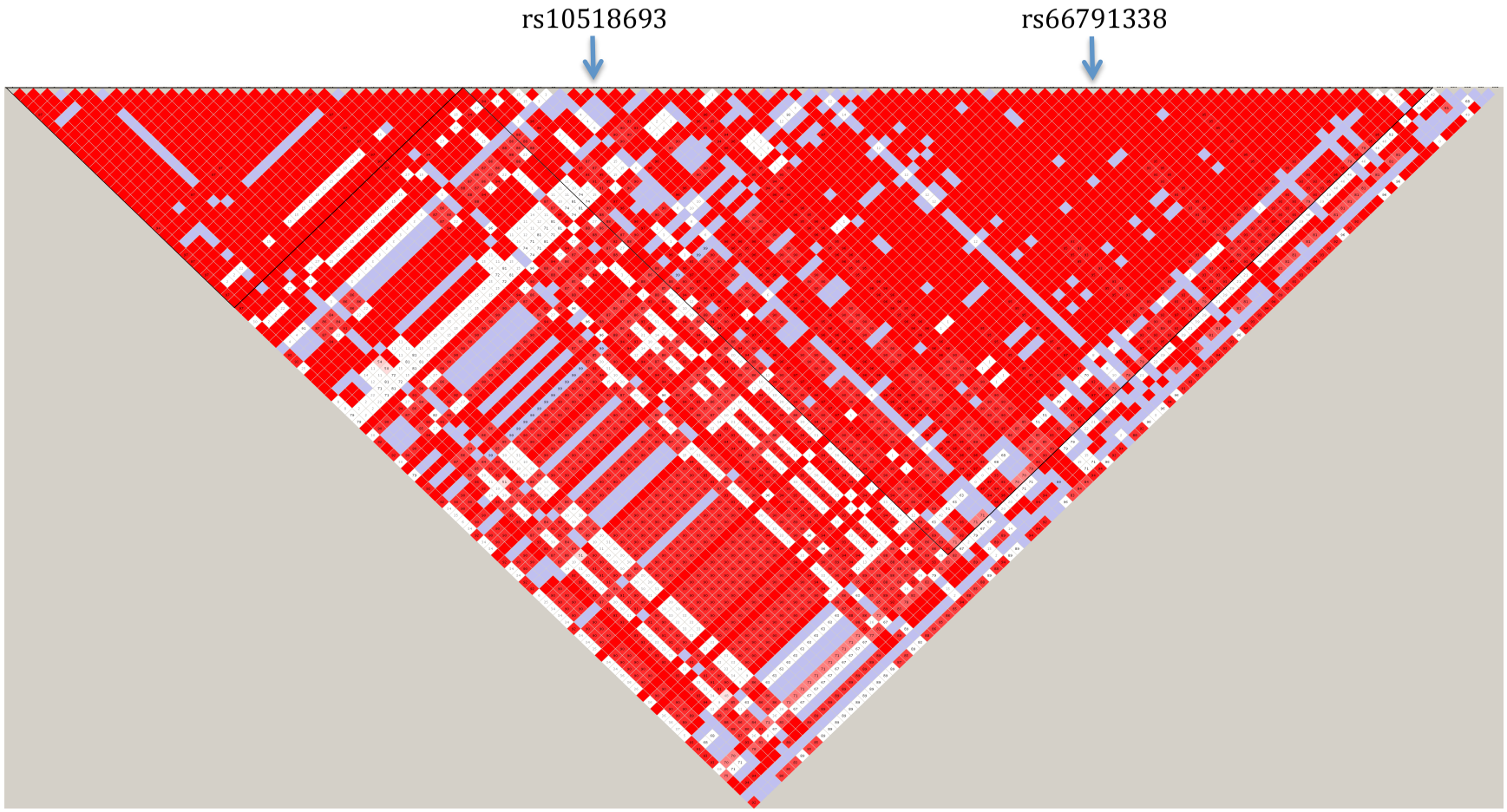
MAF > 0.05

Figure 24 (Continued)

(b) Block of solid LD surrounds functional variants in East Asians (JPT)

Figure 24 (Continued)

(c) Less LD surrounds functional variants in South Asians (GIH)



MAF > 0.05

Figure 24 (Continued)

(d) Less LD surrounds functional variants in Europeans (GBR)

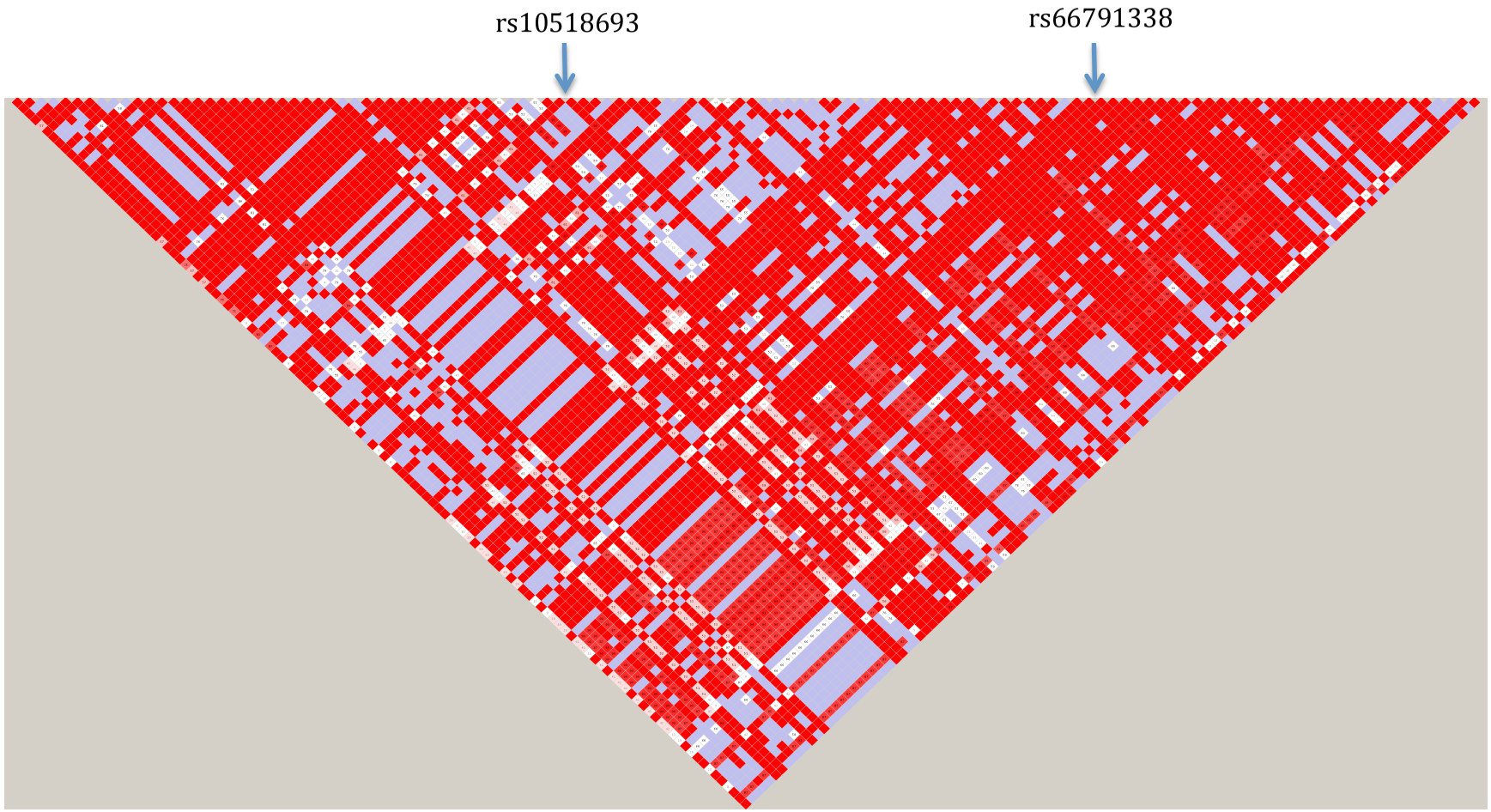


Figure 24 (Continued)

(e) Less LD surrounds functional variants in Africans (YRI)

MAF > 0.05

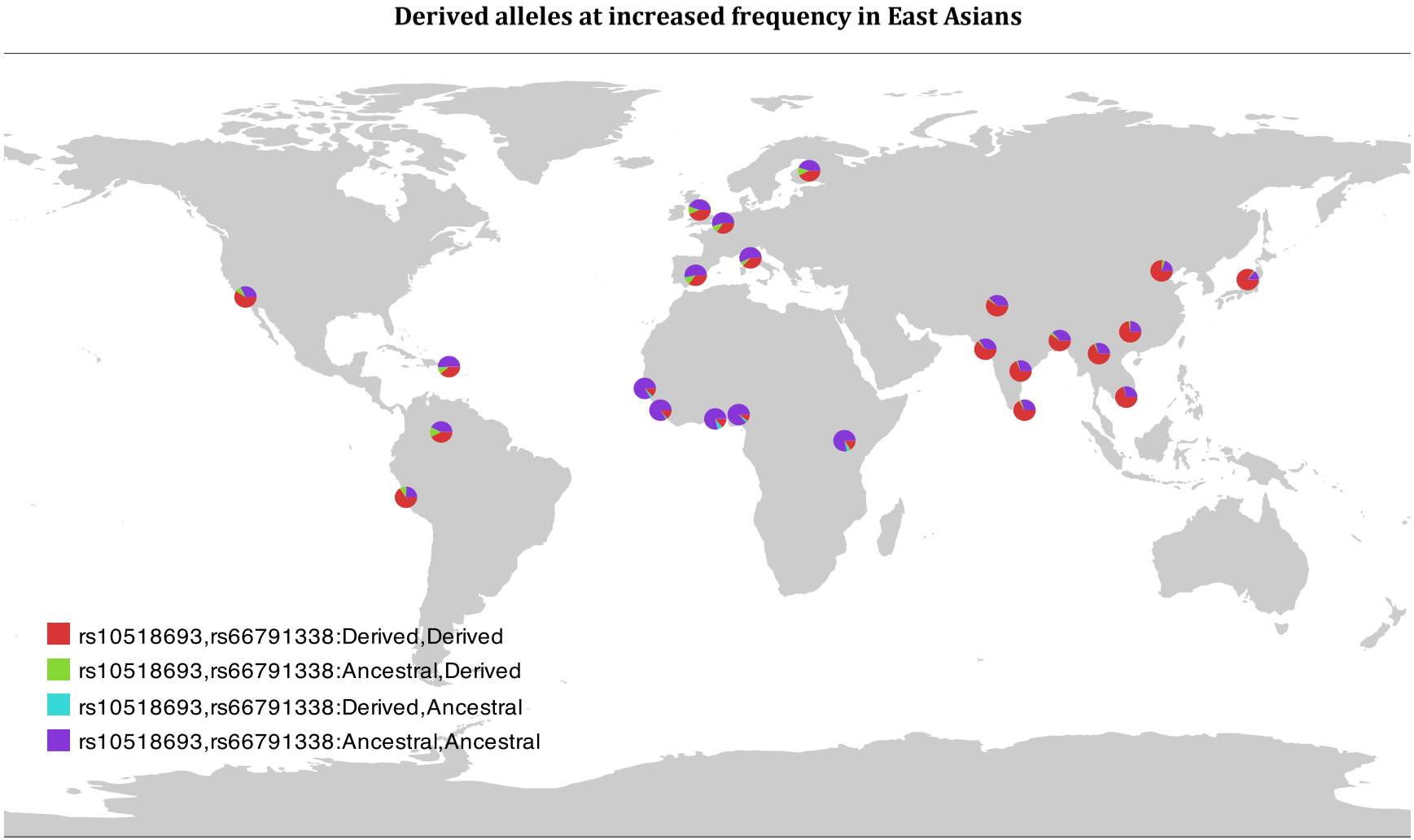


Figure 25. Haplotypes for rs10518693 and rs66791338 plotted in 1000 genomes phase 3 individuals around the world. Derived, derived haplotypes are in red. Derived, ancestral haplotypes are in blue and green. Ancestral, ancestral haplotypes are in purple.

Table 13. Functional variants and association with *IVD* expression, isovaleryl-CoA, and the CMS selection signal

Variant	Position (hg19)	r^2 (GBR) rs10518693	r^2 (GBR) rs66791338	eQTL -log10(p)	eQTL rs10518693 conditioned -log10(p)	GWAS Propionyl-CoA: Isovaleryl-CoA -log10(p)
rs10518693	40700022	-	0.635	24.31	-	41.75
rs66791338	40714431	0.635	-	7.08	6.67	-
rs11637756 (proxy)	40716654	0.621	0.978	7.10	5.87	-
rs7207 (proxy)	40713306	0.635	1	7.09	7.20	13.85

Variant	CMS Rank	ΔDAF Rank	Fst Rank	*XP-EHH* Rank	*iHS* Rank	ΔiHH Rank
rs10518693	91	46	55	231	251	372
rs66791338	-	-	-	-	-	-
rs11637756 (proxy)	53	45	50	97	217	336
rs7207 (proxy)	-	-	-	-	-	-

prior to human migrations out of Africa. This indicates that the selective sweep in East Asians altered the frequency of standing regulatory variation in the human population. Additionally, haplotype-dating of the derived haplotype in East Asians estimates the derived alleles to be younger (54,000 and 105,000 years for rs10518693 and 60,000 and 105,000 years for rs66791338), which is consistent with selection on haplotypes containing the derived functional alleles. This evidence combined with presence of long-range haplotypes matching the East Asian type present in American admixed populations reveals that this haplotype must have arisen and experienced selection prior to the emigration of people to the Americas 12,000 years ago.

Discussion

Luciferase assays identify two independent regulatory variants for *IVD* expression: rs10518693 lying in the third intron of *IVD* and rs66791338 in the 3' region downstream of *IVD*. Furthermore, a third variant, rs11633883 30bp upstream of rs66791338, modulates the behavior of this variant when tested in Hek293 and HepG2 cell lines, though this result does not replicate in the LCLs from GM12878. For rs10518693 and rs66791338, the derived alleles increase expression, and for rs11633883 the derived version enhances expression when paired with the derived allele of rs66791338 in the HepG2 and Hek293 experiments, and has no impact in the LCL experiments. The increase in expression for the derived alleles is consistent in direction with *IVD* eQTL association signals for these loci in the Geuvadis dataset (Lappalainen et al., 2013), as well as with the GWAS association with these loci (rs10518693) and linked loci (rs7207, $r^2 = 1$ with rs66791338 in GBR) for isovalerylcarnitine levels—IVD's substrate (Shin et al., 2014). Furthermore, when the *IVD*

eQTL peak of association is conditioned on the top hit, rs10518693 (Table 14), rs66791338 loses very little power in the strength of its association—its p-value changes from 8.28E-8 to 2.12E-7, despite having $r^2 = 0.635$ with rs10518693 in the British population (GBR, 1000 genomes phase 3) (Figure 26). This is consistent with the discovery of rs66791338 as an independent genetic regulatory of *IVD* expression. On the other hand, rs11633883 is almost a perfect proxy for rs10518693 ($r^2 = 0.977$ in GBR). So, all significance of association for this variant is lost when conditioned on the rs10518693 association: Its p-value of association drops from 6.97E-22 to 0.87. As a result the validity of rs11633883 as a regulator of *IVD* expression cannot be evaluated in this context.

The additional data on gene expression association from the gTEX dataset indicates the potential for tissue specificity in the actions of the identified regulatory variants. rs10581963 appears to be active in esophageal muscularis (the thin lining of muscle coating the esophagus), while rs66791338 appears to be active in skeletal muscle. The importance of leucine metabolism to esophageal muscle is unclear, but the association of rs66791338 in skeletal muscle potentially implicates leucine stimulation of the mTOR-pathway in myocytes as an important downstream phenotype of this genetic regulator.

In addition, both variants also associate with expression of *DISP2* and *RP11-64K12.4*, genes of unknown function in adipose tissue and skin. Therefore, additional impacts on downstream phenotypes through other genes cannot be excluded, though these associations with *DISP2* and *RP11-64K12.4* do not replicate in the Geuvadis dataset. rs66791338 is significantly associated with another gene, *KNSTRN*, a kinetochore-localized astrin, *SPAG5* binding protein, ($p = 2.25E-8$) in the Geuvadis dataset. On the other hand, the

Table 14. *IVD* eQTLs and conditional analysis on top eQTL association

Variant	*IVD* eQTL -log10(p)	-log10(p) Conditioned on rs10518693
rs10518693	24.31	-
rs17733008	12.64	9.77
rs139273189	11.58	9.53
rs8033938	12.27	9.40
rs7164132	11.99	9.38
rs17671194	11.77	9.34
rs8026523	12.26	9.29
rs11557072	11.62	9.28
rs4575496	8.67	9.20
rs76277863	11.80	9.19
rs55788133	11.80	9.19
rs75159543	11.80	9.19
rs34088794	11.80	9.19
rs17672041	11.94	9.16
rs8033303	10.44	9.14
rs7169404	11.83	9.12
rs7169262	11.83	9.12
rs79044944	11.83	9.12
rs75278386	11.83	9.12
rs113189667	11.83	9.12
rs78604110	11.83	9.12
chr15:40685021:D	11.83	9.12
rs113638727	11.83	9.12
rs112366697	11.83	9.12
rs111540938	11.83	9.12
rs76994419	11.31	9.10
rs3803358	11.53	9.03
rs7164321	11.88	9.00
rs8037207	11.72	8.96
rs8040755	11.46	8.96
rs11858714	11.42	8.94
rs7166991	10.54	8.62
rs79626713	11.43	8.58
rs62017982	8.28	8.52
rs11541642	10.50	8.46
rs8027487	10.83	8.21
rs55982218	10.91	8.17
rs112571916	9.38	7.98
rs7207	7.09	7.20
rs11070271	7.40	7.11

Table 14 (Continued).

rs11630850	7.40	7.11
rs1001528	7.40	7.11
rs2304645	7.27	7.10
rs7172000	11.32	7.08
chr15:40712348:D	7.74	7.07
rs12901440	7.08	6.80
rs35700143	7.66	6.75
rs66791338	7.08	6.67
rs76315331	7.47	6.66
chr15:40693274:I	7.69	6.57
rs2034650	6.93	6.52
rs79408247	8.42	6.48
rs8040128	6.93	6.40
rs11630878	7.93	6.38
rs78494549	7.99	6.15
rs77839142	7.03	6.11
rs17671250	9.10	6.05
rs56232597	7.86	6.02
rs11637756	7.10	5.87
rs80160446	7.67	5.84
rs2075625	8.09	5.75
rs8040086	7.08	5.74
rs62017977	9.12	5.66
rs76448349	7.72	5.64
rs56289107	7.72	5.64
rs74341421	7.59	5.57
rs78291781	7.59	5.57
rs11540714	7.59	5.57
rs2008462	7.59	5.57
rs77276368	7.59	5.57
rs79246418	7.59	5.57
rs59424629	7.30	5.42
rs78153629	7.99	5.40
rs77531164	7.99	5.40
rs113640277	9.74	5.34
rs10851395	7.37	5.24
rs61661087	7.32	5.17
rs74435374	6.67	5.16
chr15:40606228:D	6.17	5.03
rs55945670	7.70	4.89
rs7165012	19.04	4.70
rs62017972	8.24	3.93

Table 14 (Continued).

rs4244577	7.93	3.82
rs7165541	10.65	3.60
rs4514650	12.22	3.29
rs12443160	11.49	3.29
rs62017987	11.45	3.11
rs58640104	10.90	3.09
rs2412513	10.90	3.09
rs4924458	10.89	3.09
rs61149329	15.21	3.06
rs12904187	5.78	2.92
rs57389306	5.87	2.89
rs896798	13.46	2.82
rs4244575	5.27	2.79
rs1979189	13.82	2.63
chr15:41056071:D	5.63	2.31
rs12368	14.82	2.17
chr15:40728412:D	13.84	1.93
rs62017969	6.29	1.62
rs7176722	5.48	1.57
rs4244578	14.78	1.54
rs35556292	7.29	1.49
rs17672016	16.17	1.48
rs2075624	16.27	1.37
chr15:40695291:I	15.91	1.32
rs1453184	6.06	1.31
rs12593066	15.91	1.29
chr15:40695293:I	15.76	1.25
rs11636361	6.95	1.22
rs11070269	6.93	1.22
chr15:40692200:I	6.71	1.20
rs59714765	7.29	1.19
rs2289332	6.68	1.19
rs12916629	6.69	1.18
rs12440453	21.19	1.18
rs4924464	15.30	1.18
chr15:40728714:D	8.31	1.18
rs55679082	6.81	1.14
rs62018005	20.82	1.10
rs1898884	6.38	1.02
rs12593725	20.35	0.82
rs7165636	6.29	0.70
rs8034859	19.51	0.69

Table 14 (Continued).

rs6492945	5.99	0.66
rs8032408	5.85	0.46
rs4244579	6.64	0.45
rs12911691	6.20	0.37
rs4923865	7.55	0.35
rs1007177	7.90	0.32
rs11790	8.48	0.32
rs11632012	7.87	0.30
rs17733719	22.26	0.30
rs661488	9.03	0.25
rs1992272	7.54	0.25
rs2289330	9.14	0.24
rs12594728	22.81	0.23
rs4924457	7.31	0.22
rs12914710	7.64	0.21
rs8034177	6.83	0.20
rs4924465	6.12	0.20
rs8033249	21.29	0.18
rs11070268	7.66	0.18
rs11638033	22.61	0.16
rs12914315	8.82	0.15
rs11070267	7.04	0.15
rs2412523	8.48	0.13
rs1898883	7.03	0.12
rs8034416	7.90	0.11
rs12902310	22.24	0.11
rs12898710	6.29	0.11
rs11070270	7.87	0.08
rs11070272	8.80	0.08
rs8034217	8.64	0.08
rs11633883	21.16	0.06
rs2289329	22.33	0.06
rs2289331	9.42	0.06
rs1984793	9.30	0.04
rs600791	8.72	0.03
rs9635324	22.23	0.02
chr15:40705878:D	7.45	0.00

rs66791338 and other variants maintain association with *IVD* expression after conditioning on rs10518693, the top eQTL

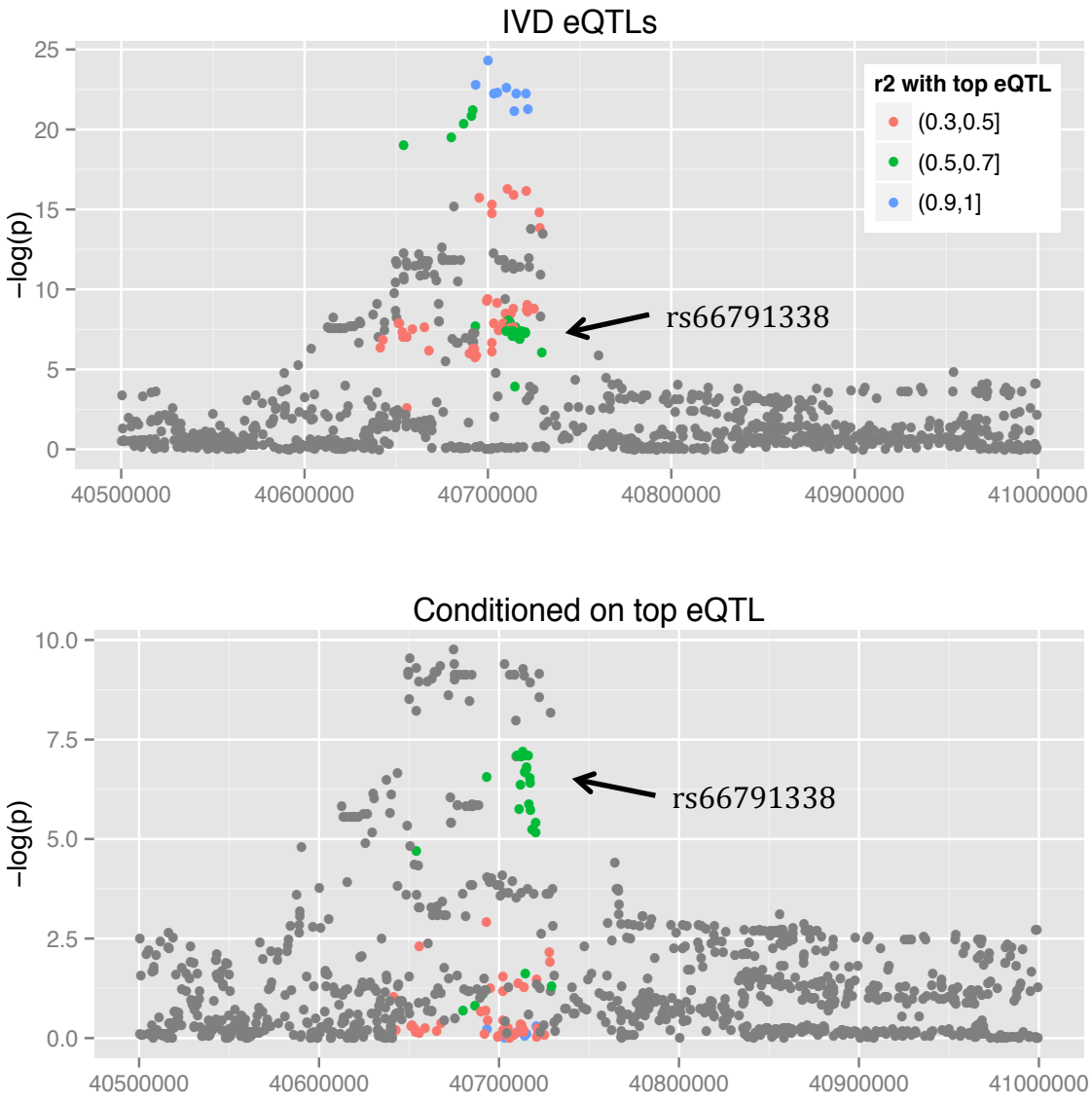


Figure 26. The eQTLs for *IVD* expression are colored according to correlation with rs10518693, the top eQTL for *IVD* expression. The top plot shows the association with *IVD* expression with –log(p). The bottom plot shows the association with *IVD* expression after conditioning on the association of rs10518693. The variants colored in green, which include rs66791338, maintain most of their power of association with *IVD* expression, despite having r^2 with rs10518693 of around 0.6.

power for detecting associations in the gTEX dataset is limited, so *IVD* expression may be impacted in adipocytes, skin, and other untested cell types, such as hepatocytes, as well.

Large region sizes of contextual genomic DNA surrounding the regulatory variants are critical for detecting the functional differences of alleles at each of these loci. For rs66791338, increasing the surrounding genomic context from 150 bp to 2.5 kb magnifies the impact of the derived allele from 1.36-fold increase to 2.61-fold increase, with the most critical change seeming to occur when the region upstream of the variant is increased from 75bp to 400bp. ChIP-seq data reveals strong evidence for the transcriptional insulator CCCTC-binding factor (CTCF) binding 100bp upstream of rs66791338 in the ENCODE dataset (Figure 27) (Dunham et al., 2012). CTCF works by creating looping structures in the 3D structure of chromatin (Phillips & Corces, 2009), which can bring repressor or enhancer elements in closer contact with transcription start sites. This suggests that CTCF binding may amplify whatever transcriptional regulation is occurring through interacting with a transcription factor that binds differentially to the rs66791338 allele.

For rs10518693, testing smaller 150-bp insert sizes yields no evidence of differential regulatory behavior of at this locus, neither through luciferase assays in GM12878 LCLs (Figure 15) nor through a high-throughput test of regulatory function (Tewhey, RS; unpublished data). However, increasing the genomic context from 150 bp to 1 kb reveals the regulatory function of this locus (Figure 15). At these loci, genetic regulation of *IVD* requires ample genomic context to function. Although most transcription factors bind to very localized binding motifs, the affinity for these motifs may be modulated by other factors binding in complex. This appears to be the case for the identified regulatory sequence in the *IVD* region. This may also generally hold true for this region as the

CTCF binding may amplify expression signal for larger regions

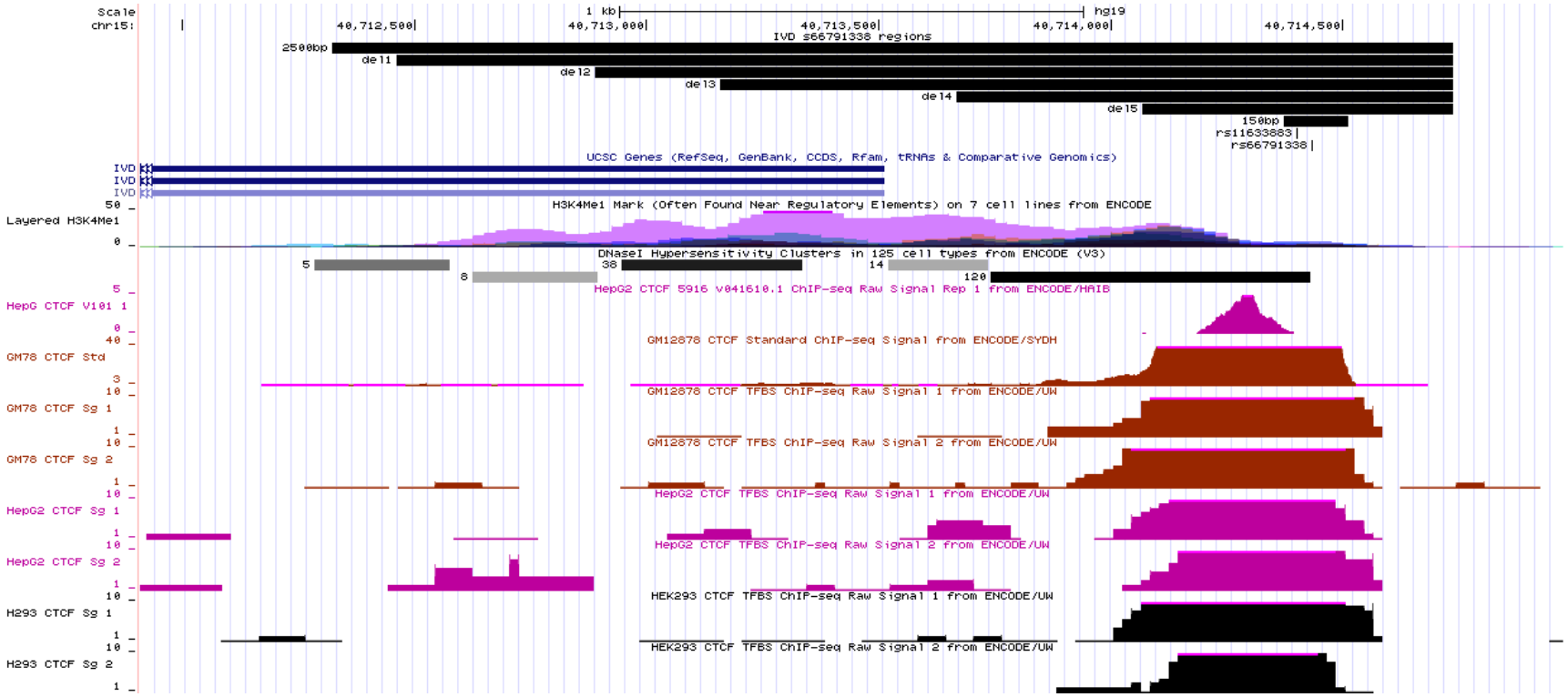


Figure 27. Tested region sizes surrounding rs66791338 shown in the UCSC genome browser along with the *IVD* gene, histone methylation marks, DNase I hypersensitivity clusters, and ChIP-seq peaks for CTCF binding.

unpublished data of the high-throughput assay, which used 150-bp regions across all tested sites, gives no evidence for regulatory variation in this region (Tewhey, RS), despite eQTL and GWAS studies showing strong peaks of association. The eQTL association conditional analysis also indicates that still other undiscovered regulatory loci of *IVD* remain to be identified (Table 14).

PWMEnrich and expression analysis suggest a few candidates for transcription factors that may link these sequences to their impact on gene expression (Table 11-12; Figures 20-21). For rs66791338, *SMUG1* (single-strand-selective monofunctional uracil-DNA glycosylase 1) and *SFT2D1* (SFT2 domain containing 1) are potential enhancers binding to the derived deletion allele, and *RARα* and *RARγ* (retinoic acid receptors α and γ) are potential silencers binding to the ancestral insertion allele. While *SMUG1*, which functions in DNA base excision repair (Jobert et al., 2013), and *SFT2D1*, which is of unknown function, but is homologous with genes involved in the Golgi complex according to GeneCards, have no known function as transcription factors, *RARα* and *RARγ* are known transcription factors, and members of the nuclear hormone receptor family (Q. Tang et al., 2011). While *RARα* and *RARγ* are known to be activators, a recent analysis of their known binding sites across the genome gave evidence that they also act as repressors (Q. Tang et al., 2011). Furthermore, a ChIP-seq experiment done on mouse liver tissue shows a signal of *RARα* binding somewhere in a 900-bp region spanning the site of rs66791338 (He, Tsuei, & Wan, 2014), which in mice is TGAAAG, most similar to the human ancestral version. This evidence suggests *RARα* may be the strongest candidate for binding to rs66791338-ancestral allele, potentially in a complex with CTCF upstream. No published literature has demonstrated these two transcription factors as acting together in a complex

before. For rs10518693, several of the candidate genes for regulation are known transcription factors, including *SOX13* (SRY-box 13), which has silencing activity (Melichar et al., 2007), and is a potential repressor binding to the ancestral version, and *MEF2B* (MADs box transcription enhancer factor 2, polypeptide B), which is a known activator in myocytes (Rodriguez et al., 2014), and is a potential enhancer binding to the derived version.

Examining haplotype structure, allele age, and selective signature for the identified genetic regulators of *IVD* and the surrounding region suggests several features of the natural history of these loci. First, both rs66791338 and rs10518693 are variants that arose prior to modern human migrations out of Africa. Derived alleles are at greater than 0.1 frequency in all populations available in the 1000 genomes phase 3 dataset. Age of the derived alleles is estimated to be 214,000 and 191,000 years for rs10518693 and 79,000 and 166,000 years for rs66791338 using the methods described in (Stephens et al., 1998). Interestingly, the Neanderthal and Denisovan genetic data currently available show ancestral alleles at both positions.

Second, visualizing, sorting, and examining linkage disequilibrium between variants in the East Asian population demonstrates that the derived alleles for rs10518693 and rs66791338 tag the long-range selected haplotype of greater than 150-kb length in this population. This long-range haplotype is at greater than 0.4 frequency in all East Asian populations, including JPT, CDX, KHV, CHS, and CHB (Figure 22a-e). However, an additional 30% of haplotypes in East Asian populations also exhibit both derived alleles in a smaller chunk of solid linkage disequilibrium of about 50kb (Figure 24a,b). Europeans and Africans do not exhibit the long-range haplotype for this region, and though rs10518693 derived

allele does occur with high-fidelity on the derived haplotype of rs66791338 in Europeans, that region does not exhibit the same high degree of linkage disequilibrium that it has in East Asians (Figure 24d).

Third, examining the CMS score at these functional variants, and breaking the scores into their component statistics, shows that the high degree of population differentiation at these loci are primarily responsible for their high rating of selection as opposed to the haplotype-based methods. These data are consistent either with selection on these functional standing variants in the East Asian populations, or with selection on other, newly derived variants that occurred on the background of the identified regulatory variants for *IVD*. The increase in co-occurrence of the derived alleles in Asian populations, and the presence of a secondary block of linkage disequilibrium surrounding some of the haplotypes in the East Asian populations contributes evidence that selection acted on these functional variants specifically. However, selection on linked loci cannot be excluded.

Conclusion

This investigation into the function and selective history of regulatory variants associated with the expression of *IVD* finds that multiple loci, frequently occurring together on the same haplotypes, govern the phenotype of *IVD* expression and the levels of its substrate, isovaleryl-CoA. The derived alleles that increase expression of *IVD*, and decrease relative levels of isovaleryl-CoA, increase in frequency in concert in the East Asian population. While these variants are old and present in diverse populations all over the world, selection on a long haplotype in the East Asian population dramatically altered allele frequency of these functional loci. While neutrality at these loci, accompanied by selection on linked variation, cannot be excluded, increasing numbers of functional variants that

exhibit the same direction of effect, co-occurring with increased frequency on the same selected haplotype contribute to evidence of selection on a complex trait governed by multiple alleles in the same region.

In addition, leucine, the amino acid metabolized by *IVD*, is a critical component of diet and metabolism in humans, making it a potential target of selection. As described above, it makes up the largest component of human protein consumption, and regulates phenotypes such as muscle anabolism, fatty acid oxidation, insulin production, and glucose uptake. Furthermore, its function shifts at critical stages in human life history, such as pregnancy, in which it may permanently shape fetal capacity for glucose uptake and prime the mTOR-pathway. Pregnancy, infancy, and the prenatal period are windows of time when the actions of selection are most potent. Therefore, genetic variation that alters leucine catabolism is highly susceptible to selection if it impacts metabolic signaling pathways such as blood glucose uptake or muscle anabolism at these life stages.

Also, leucine catabolism could be a focal point for selection as a source of energy during either a historical period, or life-history period, of energy constraint. Leucine is prevalent in the East Asian diet given their reliance on soy for protein, one of the most concentrated sources of leucine. In addition, researchers hypothesize that East Asians may have experience selection for small stature in response to energetic constraint in genes involved in skeletal growth, such as *GDF5* (Wu, Li, Jin, Li, & Zhang, 2012). If so, selection for increased efficiency in leucine catabolism could provide a compensatory adaptation for more rapidly or effectively extracting glucose, ketone bodies, or cholesterol from leucine in the diet.

Metabolic phenotypes tend to have complex genetic underpinnings, but are critical to human health and also vary across human populations with exposures to different diets, climates, life styles, and migratory histories. This study detects functional variants impacting metabolism in a region of the genome under selection in East Asian populations. In the process of this discovery, this study models how genomic signatures of selection in conjunction with genomic annotation of functional data can be leveraged in a candidate gene approach to study novel hypotheses of adaptation in diverse human populations.

Final Comments

This is unpublished research that I conducted in the lab of Pardis Sabeti. To identify candidate loci for functional follow-up, I intersected regions identified by CMS with eQTLs from the Geuvadis dataset and genes involved in the metabolism. I designed test luciferase constructs and cloned most of the constructs tested. Michael Boyle also cloned some luciferase regions for testing. I conducted transfections and luciferase assays in HepG2s, Hek293s, and LCLs. Ryan Tewhey conducted the conditional analysis of IVD eQTLs as part of his larger re-analysis of all eQTLs in the Geuvadis dataset. Michael Boyle conducted the motif enrichment analysis of rs66791338 and analyzed expression of transcription factor binding candidates across tissues. I conducted motif enrichment analysis of rs10518693 and analyzed expression of transcription factor binding candidates across tissues. I analyzed expression of transcription factor binding candidates for rs66791338 and rs10518693 across different experimental conditions in the EMBL database. Stephen Shaffner dated the ages of derived alleles using a haplotype-based method. I analyzed

haplotypes and mapped haplotype frequencies across phase 3 1000 genomes populations. I reviewed the literature, analyzed results, synthesized conclusions, and wrote the chapter.

CHAPTER 5

CONCLUSION

CONCLUDING REMARKS

Human metabolism has been under recent selective pressures in diverse populations in response to differences in climate, diet, and other factors, but pinpointing local adaptations at the level of genetic variation is challenging for several reasons. 1) Metabolic phenotypes are complex and have multiple genetic and environmental contributors. 2) Genome sequencing data for diverse human populations has been scarce. 3) Phenotype and genotype-to-phenotype associations for many populations have not been systematically studied. 4) Different genetic signatures of selection have strong power to detect selection only of certain types and certain time frames. 5) Rapid human cultural change and large-scale migrations have distanced many human populations from the climate and dietary factors that shaped their evolution. 6) Functional assays of genetic variation have limitations and require intensive research.

These challenges to finding human metabolic adaptations have been mitigated in several ways. 1) More high-powered association studies of metabolic phenotypes are being conducted in more diverse populations, though this remains a limiting factor in many areas of research. 2) The 1000 genomes project is now providing full sequenced genomes for large numbers of individuals from diverse populations (Abecasis et al., 2012). 3) The CMS test for selection combines different statistical signatures such as population differentiation (Grossman et al., 2013), which may improve its power to detect selection on standing variation in addition to newly derived variants of large effect size. 4) More high-

throughput functional assays are being conducted on genomic datasets (Consortium et al., 2015).

Despite these advances, many problems remain. Focusing more medical research on diverse human populations could still be very beneficial. For example, a recent large scale association study of type II diabetes in Mexicans found a variant of large effect size explaining up to 20% of the increase in disease prevalence in this population (A. L. Williams et al., 2013). This genetic variant is too low frequency in European populations to be detected by association studies in those populations. Even with better phenotype and genotype-to-phenotype data from diverse populations, data on diet, environment, and lifestyle for these populations prior to broad cultural shifts and migrations may remain elusive to some extent.

In this difficult landscape, this dissertation provides a model of how new adaptive hypotheses can be generated and supports several insights. 1) Knowledge of the constraints experienced in critical life stages like pregnancy can help predict human metabolic adaptations. 2) Multiple variants on the same haplotype may impact the same phenotype. 3) Selection on multiple standing variants on the same haplotype may be difficult to distinguish from selection on a single newly derived variant on that haplotype. 4) Functional assays may require ample genomic context to exhibit regulatory activity. 5) Current understanding of human phenotypic variation is incomplete, and overlooks important differences in how humans interact with diet and energy metabolism.

Furthermore, this dissertation provides evidence that many more adaptations in diverse human populations remain to be explored, and new genomic research enables this pursuit. This research on *IVD* regulatory variants in Asian populations is just one of many

studies that may now be undertaken. As a final example, another strong candidate gene, *ERLIN1*, for local metabolic adaptation in human populations is described below to bolster the sense that human metabolic adaptations are prevalent, and gaining a better understanding of them will lend critical insight into human history and health.

ERLIN1 and Selection on Cholesterol Metabolism in the Yoruba

eQTLs for a gene called *ERLIN1*, the endoplasmic reticulum lipid raft-associated protein 1 (function described below), lie in a CMS region under selection in the Yoruba (Grossman et al., 2013). A high-throughput functional assay to test regulatory function of candidate eQTLs confirmed three eQTLs in the peak of association for *ERLIN1*, and another variant that failed to reach genome-wide significance ($p=2.6E-3$) (Lappalainen et al., 2013), to exhibit strong enhancer activity with significantly higher expression driven by the ancestral alleles compared to the derived alleles at these loci (Figure 28) (Tewhey, unpublished data). Of these four SNPs, rs7089292 is the seventh highest scoring variant for CMS in the region, and it is in perfect linkage disequilibrium with the third highest SNP and three others in the top ten variants for CMS in the Yoruba population. The other three regulatory SNPs were not tested by CMS, but are in linkage disequilibrium with rs7089292, and all four regulatory variants are at high frequency (~0.3) only in the Yoruba and a few other African populations. The Neanderthal and Denisovan genetic data currently available show ancestral alleles at each of these positions. In addition, the top two scoring SNPs for CMS in the region are located in a gene intron and intergenic regions, and exhibit no marks of regulatory activity, making them less likely to be functional genetic candidates driving the selection in this region.

ERLIN1 eQTLs have high probability of selection in YRI, and four eQTLs validate as regulatory variants

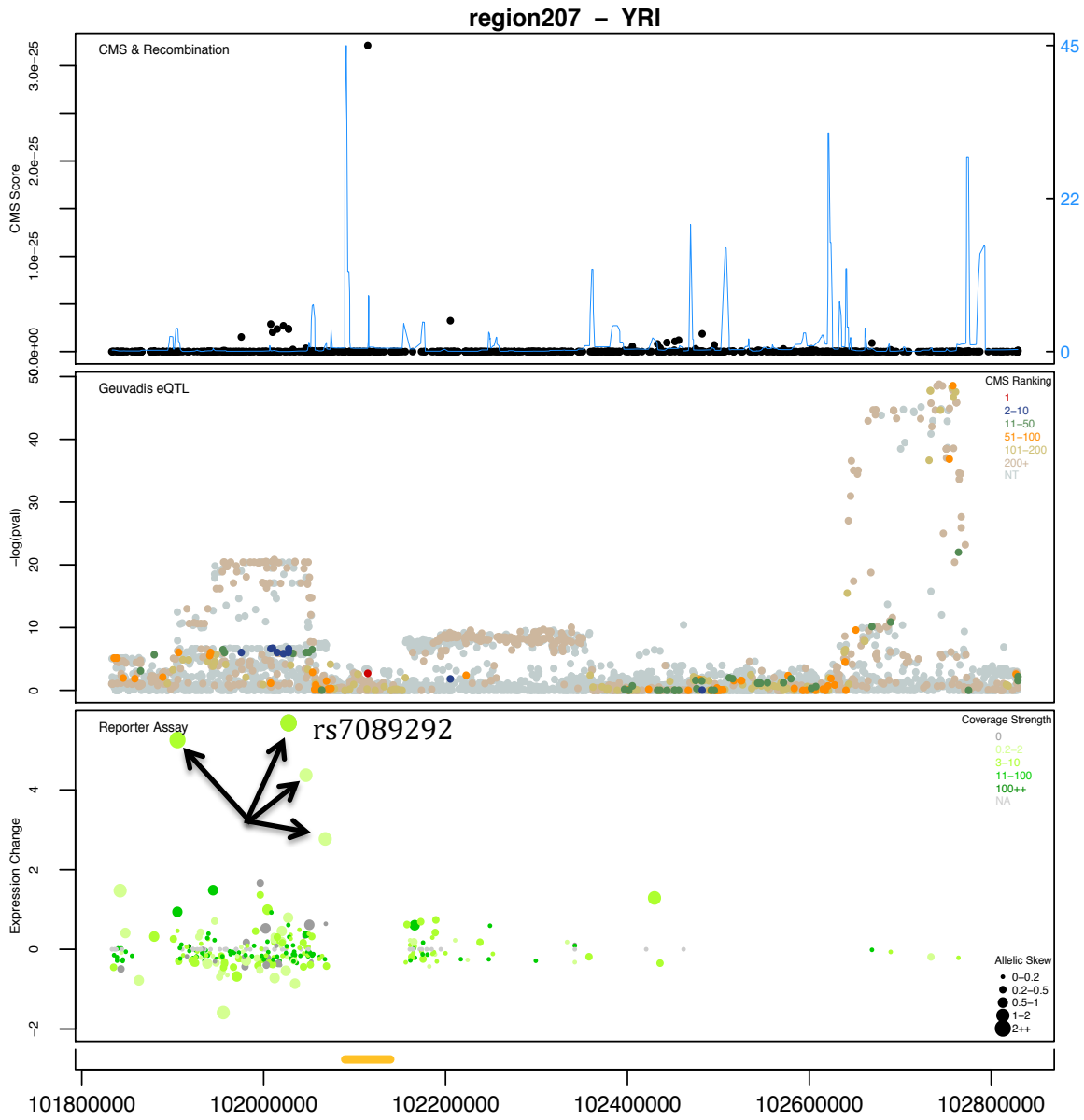


Figure 28. The top plot shows the CMS scores for a region on chromosome 10 in YRI with recombination in the background. The middle plot shows eQTLs for all genes in the plotted region, colored with where they fall in the peak of CMS scores for the region. The bottom plot shows the results of a high-throughput assay to test regulatory function of eQTLs. Larger circles indicate a change in expression between derived and ancestral alleles, while the y-axis indicates an overall change in expression compared to a neutral distribution of variants and regions. (Tewhey, unpublished data)

The function of *ERLIN1* and the consequences of nearby genetic variation make it a strong candidate for local metabolic adaptation in humans. ERLIN1, most highly expressed in the liver (Figure 29), is a transmembrane protein in the endoplasmic reticulum (ER) where it is part of a protein-binding complex that regulates the synthesis of HMG-CoA reductase, which catalyzes the synthesis of cholesterol. Specifically, ERLIN1, together with ERLIN2, bind to a complex of proteins in the ER, including SREBPs (sterol regulatory element-binding proteins) and inhibit them from travelling to the nucleus and stimulating the synthesis of HMG-CoA reductase to make cholesterol (Huber, Vesely, Datta, & Gerace, 2013).

In line with ERLIN1 control of cholesterol synthesis, European eQTLs for *ERLIN1*, including a missense mutation in the gene, associate with non-alcoholic fatty liver disease (NAFLD) in Europeans (Feitosa et al., 2013). NAFLD is the accumulation of triglycerides in the liver, and increased cholesterol levels contribute to this condition. One of the variants, rs12784396, significant in Europeans for fatty liver and *ERLIN1* expression, is within 50bp of rs7089292, the Yoruba eQTL and confirmed regulatory variant for *ERLIN1*. These SNPs are located in the 5'UTR and promoter, respectively, of another gene, *CWF19L1*, but only associate with *ERLIN1* expression. Variation in the *ERLIN1* region also associates with elevated liver levels of alanine aminotransferase (ALT), which accompanies NAFLD, in Europeans and Indians (Feitosa et al., 2013; Yuan et al., 2008).

Unfortunately, only one GWAS of NAFLD has been conducted in an African population. This GWAS included 1,032 African-Americans and used a panel of 9,229 non-synonymous coding variants (Romeo et al., 2008). It detected only one significant hit in the gene *PNPLA3*. One other GWAS of triglyceride response to anti-hypertensive drugs in

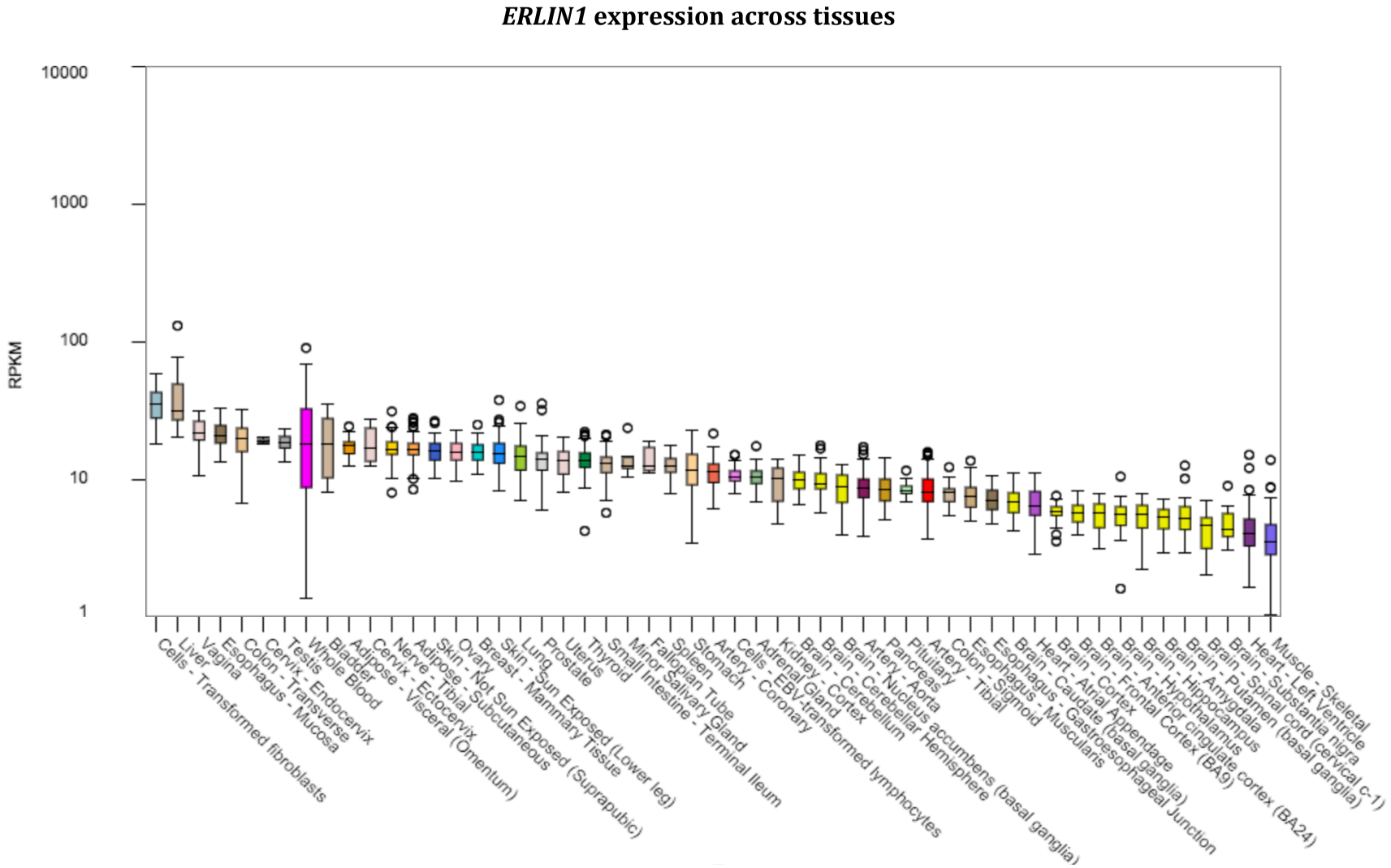


Figure 29. *ERLIN1* expression as measured by RNA-seq in diverse tissue types in the gTEX database, plotted by log transformed values of RPKM, Reads Per Kilobase of transcript per Million mapped reads.

African-Americans detected the variant, rs9420790, in the *ERLIN1* region (Del-Aguila et al., 2014), but this variant is not in strong linkage disequilibrium with any of the Yoruba regulatory variants or eQTLs for *ERLIN1*.

Despite this limitation, the Yoruba regulatory variants for *ERLIN1* are likely candidates for metabolic adaptation, given that they fall near the top of selected variants in CMS for the region. The fact that *ERLIN1* eQTLs in Europeans also associate with NAFLD makes it very likely that *ERLIN1* regulatory variants in the Yoruba also impact NAFLD and related phenotypes. The case for this is especially strong given ERLIN1's role in limiting cholesterol production in the liver. Furthermore, NAFLD and related phenotypes vary across ethnicities with African-Americans having lower incidence (24%) than Hispanics (45%) and European-Americans (33%) in the Dallas Heart Study (Browning et al., 2004). Therefore, some African populations may have metabolic adaptations that impact their genetic risk of NAFLD. While NAFLD is often part of metabolic syndrome, hepatitis virus B and C also cause it. Other infectious diseases, such as malaria, have been the strongest drivers of selection in humans, so this could be an overlap of metabolic and infectious disease adaptation. This example represents fruitful potential research both for human health in diverse populations and also for human evolutionary history.

REFERENCES

Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. a., Durbin, R. M., Handsaker, R. E., . . . McVean, G. a. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature, 491*, 56-65. doi: 10.1038/nature11632

Ahlborg, G., Felig, P., Hagenfeldt, L., Hendler, R., & Wahren, J. (1974). Substrate Turnover during Prolonged Exercise in Man. *Journal of Clinical Investigation, 53*, 1080-1090. doi: 10.1172/JCI107645

Akey, J. M. (2009). Constructing genomic maps of positive selection in humans: Where do we go from here? *Genome research, 19*, 711-722. doi: 10.1101/gr.086652.108

Akey, J. M., Eberle, M. A., Rieder, M. J., Carlson, C. S., Shriver, M. D., Nickerson, D. A., & Kruglyak, L. (2004). Population history and natural selection shape patterns of genetic variation in 132 genes. *PLoS biology, 2*, 1591-1599. doi: 10.1371/journal.pbio.0020286

Alejandro, E. U., Gregg, B., Wallen, T., Kumusoglu, D., Meister, D., Chen, A., . . . Bernal-Mizrachi, E. (2014). Maternal diet-induced microRNAs and mTOR underlie β cell dysfunction in offspring. *The Journal of clinical investigation, 124*, 1-16. doi: 10.1172/JCI74237

Alkorta-Aranburu, G., Beall, C. M., Witonsky, D. B., Gebremedhin, A., Pritchard, J. K., & Di Rienzo, A. (2012). The Genetic Architecture of Adaptations to High Altitude in Ethiopia. *PLoS Genetics, 8*. doi: 10.1371/journal.pgen.1003110

Almgren, P., Lehtovirta, M., Isomaa, B., Sarelin, L., Taskinen, M. R., Lyssenko, V., . . . Groop, L. (2011). Heritability and familiality of type 2 diabetes and related quantitative traits in the Botnia Study. *Diabetologia, 54*, 2811-2819. doi: 10.1007/s00125-011-2267-5

Alonso, S., Izagirre, N., Smith-Zubiaga, I., Gardeazabal, J., Díaz-Ramón, J. L., Díaz-Pérez, J. L., . . . de la Rúa, C. (2008). Complex signatures of selection for the melanogenic loci TYR, TYRP1 and DCT in humans. *BMC evolutionary biology, 8*, 74. doi: 10.1186/1471-2148-8-74

Anthony, J. C., Yoshizawa, F., Anthony, T. G., Vary, T. C., Jefferson, L. S., & Kimball, S. R. (2000). Leucine stimulates translation initiation in skeletal muscle of postabsorptive rats via a rapamycin-sensitive pathway. *The Journal of nutrition, 130*, 2413-2419.

Asmann, Y. W., Necela, B. M., Kalari, K. R., Hossain, A., Baker, T. R., Carr, J. M., . . . Thompson, E. A. (2012). Detection of redundant fusion transcripts as biomarkers or disease-specific therapeutic targets in breast cancer. *Cancer Research, 72*, 1921-1928. doi: 10.1158/0008-5472.CAN-11-3142

Atkinson, F. S., Foster-Powell, K., & Brand-Miller, J. C. (2008). International tables of glycemic index and glycemic load values: 2008. *Diabetes care, 31*, 2281-2283. doi: 10.2337/dc08-1239

Barbosa-Morais, N. L., Irimia, M., Pan, Q., Xiong, H. Y., Gueroussov, S., Lee, L. J., . . . Blencowe, B. J. (2012). The evolutionary landscape of alternative splicing in vertebrate species. *Science (New York, NY), 338*, 1587-1593. doi: 10.1126/science.1230612

Barbour, L. a., McCurdy, C. E., Hernandez, T. L., Kirwan, J. P., Catalano, P. M., & Friedman, J. E. (2007). Cellular mechanisms for insulin resistance in normal pregnancy and gestational diabetes. *Diabetes care, 30 Suppl 2*, S112-119. doi: 10.2337/dc07-s202

Beall, C. M. (2007). Two routes to functional adaptation: Tibetan and Andean high-altitude natives. *Proceedings of the National Academy of Sciences of the United States of America, 104 Suppl* 8655-8660. doi: 10.1073/pnas.0701985104

Beall, C. M., Cavalleri, G. L., Deng, L., Elston, R. C., Gao, Y., Knight, J., . . . Zheng, Y. T. (2010a). Natural selection on EPAS1 (HIF2{alpha}) associated with low hemoglobin concentration in Tibetan highlanders. *Proceedings of the National Academy of Sciences of the United States of America, 107*, 11459-11464. doi: 10.1073/pnas.1002443107

Beall, C. M., Cavalleri, G. L., Deng, L., Elston, R. C., Gao, Y., Knight, J., . . . Zheng, Y. T. (2010b). Natural selection on EPAS1 (HIF2α) associated with low hemoglobin concentration in Tibetan highlanders. *Proceedings of the National Academy of Sciences of the United States of America, 107*, 11459-11464.

Beall, C. M., Strohl, K. P., Blangero, J., Williams-Blangero, S., Almasy, L. a., Decker, M. J., . . . Gonzales, C. (1997). Ventilation and hypoxic ventilatory response of Tibetan and Aymara high altitude natives. *American journal of physical anthropology, 104*, 427-447. doi: 10.1002/(SICI)1096-8644(199712)104:4<427::AID-AJPA1>3.0.CO;2-P

Becker, N. S. a., Verdu, P., Froment, A., Le Bomin, S., Pagezy, H., Bahuchet, S., & Heyer, E. (2011). Indirect evidence for the genetic determination of short stature in African Pygmies. *American Journal of Physical Anthropology, 145*, 390-401. doi: 10.1002/ajpa.21512

Bellamy, L., Casas, J.-P., Hingorani, A. D., & Williams, D. (2009). Type 2 diabetes mellitus after gestational diabetes: a systematic review and meta-analysis. *Lancet, 373*, 1773-1779. doi: 10.1016/S0140-6736(09)60731-5

Bigham, A., Bauchet, M., Pinto, D., Mao, X., Akey, J. M., Mei, R., . . . Shriver, M. D. (2010a). Identifying Signatures of Natural Selection in Tibetan and Andean Populations Using Dense Genome Scan Data. *PLoS Genetics, 6*, 1-14.

Bigham, A., Bauchet, M., Pinto, D., Mao, X., Akey, J. M., Mei, R., . . . Shriver, M. D. (2010b). Identifying Signatures of Natural Selection in Tibetan and Andean Populations Using

Dense Genome Scan Data. *PLoS Genetics, 6*, e1001116. doi: 10.1371/journal.pgen.1001116

Blake, C. a., Brown, L. M., Duncan, M. W., Hunsucker, S. W., & Helmke, S. M. (2005). Estrogen regulation of the rat anterior pituitary gland proteome. *Experimental biology and medicine (Maywood, N.J.), 230*, 800-807.

Bloom-Feshbach, K., Simonsen, L., Viboud, C., Mølbak, K., Miller, M. a., Gottfredsson, M., & Andreasen, V. (2011). Natality decline and miscarriages associated with the 1918 influenza pandemic: the Scandinavian and United States experiences. *The Journal of infectious diseases, 204*, 1157-1164. doi: 10.1093/infdis/jir510

Blum, H. (1961). Does the Melanin Pigment of Human Skin Have Adaptive Value?: An Essay in Human Ecology and the Evolution of Race. *Quarterly Review of Biology, 36*, 50-63.

Brown, E. A. (2012). Genetic explorations of recent human metabolic adaptations: hypotheses and evidence. *Biological reviews of the Cambridge Philosophical Society, 87*, 838-855. doi: 10.1111/j.1469-185X.2012.00227.x

Brown, E. A., Ruvolo, M., & Sabeti, P. C. (2013). Many ways to die, one way to arrive: how selection acts through pregnancy. *Trends in Genetics*, 1-8. doi: 10.1016/j.tig.2013.03.001

Brown, I. J., Tzoulaki, I., Candeias, V., & Elliott, P. (2009). Salt intakes around the world: implications for public health. *International journal of epidemiology, 38*, 791-813. doi: 10.1093/ije/dyp139

Browning, J. D., Szczepaniak, L. S., Dobbins, R., Nuremberg, P., Horton, J. D., Cohen, J. C., . . . Hobbs, H. H. (2004). Prevalence of hepatic steatosis in an urban population in the United States: Impact of ethnicity. *Hepatology, 40*, 1387-1395. doi: 10.1002/hep.20466

Brunvand, L., Quigstad, E., Urdal, P., & Haug, E. (1996). Vitamin D deficiency and fetal growth. *Early human development, 45*, 27-33.

Butte, N. F. (2000). Carbohydrate and lipid metabolism in pregnancy: normal compared with gestational diabetes mellitus. *The American journal of clinical nutrition, 71*, 1256S-1261S.

Campino, S., Kwiatkowski, D., & Dessein, A. (2006). Mendelian and complex genetics of susceptibility and resistance to parasitic infections. *Seminars in immunology, 18*, 411-422. doi: 10.1016/j.smim.2006.07.011

Cardona, A., Pagani, L., Antao, T., Lawson, D. J., Eichstaedt, C. a., Yngvadottir, B., . . . Kivisild, T. (2014). Genome-wide analysis of cold adaptation in indigenous Siberian populations. *PLoS ONE, 9*. doi: 10.1371/journal.pone.0098076

Carmody, R. N., Weintraub, G. S., & Wrangham, R. W. (2011). Energetic consequences of thermal and nonthermal food processing. *PNAS*, 1-5. doi: 10.1073/pnas.1112128108

Carmody, R. N., & Wrangham, R. W. (2009). Cooking and the Human Commitment to a High-quality Diet. *Cold Spring Harbor symposia on quantitative biology*, 1-8.

Carrera, J. (2007). Recommendations and Guidelines for Perinatal Medicine.

Carroll, S. B. (2003). Genetics and the making of Homo sapiens. *Nature, 422*, 849-857.

Caughey, A. B., Cheng, Y. W., Stotland, N. E., Washington, a. E., & Escobar, G. J. (2010). Maternal and paternal race/ethnicity are both associated with gestational diabetes. *American journal of obstetrics and gynecology, 202*, 616.e611-615. doi: 10.1016/j.ajog.2010.01.082

Chen, Q. H., Ge, R. L., Wang, X. Z., Chen, H. X., Wu, T. Y., Kobayashi, T., & Yoshimura, K. (1997). Exercise performance of Tibetan and Han adolescents at altitudes of 3,417 and 4,300 m. *Journal of applied physiology (Bethesda, Md. : 1985), 83*, 661-667.

Chen, X., Wang, H., Zhou, G., Zhang, X., Dong, X., Zhi, L., . . . He, F. (2009). Molecular Population Genetics of Human CYP3A Locus: Signatures of Positive Selection and Implications for Evolutionary Environmental Medicine. *Environmental Health Perspectives, 117*, 1541-1548. doi: 10.1289/ehp.0800528

Cheng, C.-Y., Reich, D., Coresh, J., Boerwinkle, E., Patterson, N., Li, M., . . . Kao, W. H. L. (2010). Admixture mapping of obesity-related traits in African Americans: the Atherosclerosis Risk in Communities (ARIC) Study. *Obesity (Silver Spring, Md.), 18*, 563-572. doi: 10.1038/oby.2009.282

Chiu, M., Austin, P. C., Manuel, D. G., Shah, B. R., & Tu, J. V. (2011). Deriving Ethnic-Specific BMI Cutoff Points for Assessing Diabetes Risk. *Diabetes care, 34*, 1741-1748. doi: 10.2337/dc10-2300

Cnattingius, S., Reilly, M., Pawitan, Y., & Lichtenstein, P. (2004). Maternal and fetal genetic factors account for most of familial aggregation of preeclampsia: a population-based Swedish cohort study. *American journal of medical genetics. Part A, 130A*, 365-371. doi: 10.1002/ajmg.a.30257

Collins, J. E., Umpleby, a. M., Boroujerdi, M. a., Leonard, J. V., & Sonksen, P. H. (1987). Effect of insulin on leucine kinetics in maple syrup urine disease. *Pediatric research, 21*, 10-13. doi: 10.1203/00006450-198701000-00004

Consortium, R. E., Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., . . . Kellis, M. (2015). Integrative analysis of 111 reference human epigenomes. *Nature, 518*, 317-330. doi: 10.1038/nature14248

Coop, G., Witonsky, D., Di Rienzo, A., & Pritchard, J. K. (2010). Using environmental correlations to identify Loci underlying local adaptation. *Genetics, 185*, 1411-1423. doi: 10.1534/genetics.110.114819

Corbett, S. J., McMichael, A. J., & Prentice, A. M. (2008). Type 2 diabetes, cardiovascular disease, and the evolutionary paradox of the polycystic ovary syndrome: a fertility first hypothesis. *American journal of human biology : the official journal of the Human Biology Council, 21*, 587-598. doi: 10.1002/ajhb.20937

Cordain, L., Miller, J. B., Eaton, S. B., Mann, N., Holt, S. H., & Speth, J. D. (2000). Plant-animal subsistence ratios and macronutrient energy estimations in worldwide hunter-gatherer diets. *The American journal of clinical nutrition, 71*, 682-692.

Courtiol, A., Pettay, J. E., Jokela, M., Rotkirch, A., & Lummaa, V. (2012). Natural and sexual selection in a monogamous historical human population. *Proceedings of the National Academy of Sciences of the United States of America, 109*, 8044-8049. doi: 10.1073/pnas.1118174109

Cui, B., Zhu, X., Xu, M., Guo, T., Zhu, D., Chen, G., . . . Ning, G. (2011). A genome-wide association study confirms previously reported loci for type 2 diabetes in Han Chinese. *PloS one, 6*, 1-6. doi: 10.1371/journal.pone.0022353

Currat, M., Trabuchet, G., Rees, D., Perrin, P., Harding, R. M., Clegg, J. B., . . . Excoffier, L. (2002). Molecular analysis of the beta-globin gene cluster in the Niokholo Mandenka population reveals a recent origin of the beta(S) Senegal mutation. *American journal of human genetics, 70*, 207-223. doi: 10.1086/338304

De La Vega, F. M., Bustamante, C., & Leal, S. M. (2011). *Genome-Wide Association Mapping and Rare Alleles: From population genomics to personalized medicine.* Paper presented at the Pacific Symposium on Biocomputing, Foster City, CA.

De Silva, N. M. G., & Frayling, T. M. (2010). Novel biological insights emerging from genetic studies of type 2 diabetes and related metabolic traits. *Current Opinion in Lipidology, 21*, 44-50.

Del-Aguila, J. L., Beitelshees, a. L., Cooper-Dehoff, R. M., Chapman, a. B., Gums, J. G., Bailey, K., . . . Boerwinkle, E. (2014). Genome-wide association analyses suggest NELL1 influences adverse metabolic response to HCTZ in African Americans. *The pharmacogenomics journal, 14*, 35-40. doi: 10.1038/tpj.2013.3

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., . . . Guigó, R. (2012). The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Research, 22*, 1775-1789. doi: 10.1101/gr.132159.111

Diamond, J. (2003). The double puzzle of diabetes. *Nature, 423*, 599-602. doi: 10.1038/423599a

Dickinson, S., Colagiuri, S., Faramus, E., Petocz, P., & Brand-Miller, J. C. (2002). Postprandial Hyperglycemia and Insulin Sensitivity Differ among Lean Young Adults of Different Ethnicities. *The Journal of Nutrition*, 2574-2579.

Ding, K., & Kullo, I. J. (2008). Molecular population genetics of PCSK9: a signature of recent positive selection. *Pharmacogenetics and genomics, 18*, 169-179. doi: 10.1097/FPC.0b013e3282f44d99.Molecular

Djebali, S., Davis, C. a., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., . . . Gingeras, T. R. (2012). Landscape of transcription in human cells. *Nature, 489*, 101-108. doi: 10.1038/nature11233

Duan, Y., Li, F., Liu, H., Li, Y., Liu, Y., Kong, X., . . . Yin, Y. (2015). Nutritional and regulatory roles of leucine in muscle growth and fat reduction. *Frontiers in Bioscience*, 796-813.

Duley, L. (1992). Maternal mortality associated with hypertensive disorders of pregnancy in Africa, Asia, Latin America and the Caribbean. *British journal of obstetrics and gynaecology, 99*, 547-553.

An integrated encyclopedia of DNA elements in the human genome, 489 57-74 (2012).

Dunsworth, H. M., Warrener, A. G., Deacon, T., Ellison, P. T., & Pontzer, H. (2012). Metabolic hypothesis for human altriciality. *Proceedings of the National Academy of Sciences of the United States of America, 109*, 15212-15216. doi: 10.1073/pnas.1205282109

Dupuis, J., Langenberg, C., Prokopenko, I., Saxena, R., Soranzo, N., Jackson, A. U., . . . Barroso, I. (2010). New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nature Genetics, 42*, 105-116.

Eisenberg, D. T. a., Kuzawa, C. W., & Hayes, M. G. (2010). Worldwide allele frequencies of the human apolipoprotein E gene: climate, local adaptations, and evolutionary history. *American journal of physical anthropology, 143*, 100-111. doi: 10.1002/ajpa.21298

Ellison, P. T. (1990). Human Ovarian Function and Reproductive Ecology: New Hypotheses. *American Anthropologist, 92*, 933-952. doi: 10.1525/aa.1990.92.4.02a00050

Enattah, N. S., Jensen, T. G. K., Nielsen, M., Lewinski, R., Kuokkanen, M., Rasinpera, H., . . . Peltonen, L. (2008). Independent Introduction of Two Lactase-Persistence Alleles into Human Populations Reflects Different History of Adaptation to Milk Culture. *Journal of Human Genetics, 82*, 57-72. doi: 10.1016/j.ajhg.2007.09.012.

Fagerberg, L., Hallstrom, B. M., Oksvold, P., Kampf, C., Djureinovic, D., Odeberg, J., . . . Uhlen, M. (2013). Analysis of the Human Tissue-specific Expression by Genome-wide Integration of Transcriptomics and Antibody-based Proteomics. *Molecular & Cellular Proteomics, 13*, 397-406. doi: 10.1074/mcp.M113.035600

Fagundes, N. J. R., Salzano, F. M., Batzer, M. a., Deininger, P. L., & Bonatto, S. L. (2005). Worldwide genetic variation at the 3'-UTR region of the LDLR gene: possible influence of natural selection. *Annals of human genetics, 69*, 389-400. doi: 10.1046/j.1529-8817.2005.00163.x

FAO, I., ISRIC-World Soil Information, Institute of Soil Science, Chinese Academy of Sciences (ISSCAS), Joint Research Centre of the European Commission (JRC). (2012). Harmonized World Soil Database v 1.2.

Feinstein, J. A., & O'Brien, K. (2003). Acute metabolic decompensation in an adult patient with isovaleric acidemia. *The Southern medical journal, 96*, 500-503. doi: 10.1097/01.SMJ.0000051141.03668.1D

Feitosa, M. F., Wojczynski, M. K., North, K. E., Zhang, Q., Province, M. a., Carr, J. J., & Borecki, I. B. (2013). The ERLIN1-CHUK-CWF19L1 gene cluster influences liver fat deposition and hepatic inflammation in the NHLBI Family Heart Study. *Atherosclerosis, 228*, 175-180. doi: 10.1016/j.atherosclerosis.2013.01.038

Fernández, J. R., Shriver, M. D., Beasley, T. M., Rafla-Demetrious, N., Parra, E., Albu, J., . . . Allison, D. B. (2003). Association of African genetic admixture with resting metabolic rate and obesity among women. *Obesity research, 11*, 904-911. doi: 10.1038/oby.2003.124

Fleming, A., & Copp, A. J. (1998). Embryonic folate metabolism and mouse neural tube defects. *Science (New York, N.Y.), 280*, 2107-2109. doi: 10.1126/science.280.5372.2107

Fogelman, Y., Rakover, Y., & Luboshitzky, R. (1995). High prevalence of vitamin D deficiency among Ethiopian women immigrants to Israel: exacerbation during pregnancy and lactation. *Israel journal of medical sciences, 31*, 221-224.

Forster, P., & Matsumura, S. (2005). Did early humans go north or south? *Science (New York, N.Y.), 308*, 965-966. doi: 10.1126/science.1113261

Fraser, H. B. (2013). Gene expression drives local adaptation in humans. *Genome research, 23*, 1089-1096. doi: 10.1101/gr.152710.112

Friedman, J. M., & Halaas, J. L. (1998). Leptin and the regulation of body weight in mammals. *Nature, 395*, 763-770. doi: 10.1038/27376

Frisch, R. E. (1984). Body fat, puberty and fertility. *Biological reviews of the Cambridge Philosophical Society, 59*, 161-188.

Ge, R. L., Chen, Q. H., Wang, L. H., Gen, D., Yang, P., Kubo, K., . . . others. (1994). Higher exercise performance and lower VO2 max in Tibetan than Han residents at 4,700 m altitude. *Journal of Applied Physiology, 77*, 684-691.

Goldin, B. R., Adlercreutz, H., Gorbach, S. L., Warram, J. H., Dwyer, J. T., Swenson, L., & Woods, M. N. (1982). Estrogen excretion patterns and plasma levels in vegetarian and omnivorous women. *The New England journal of medicine, 307*, 1542-1547. doi: 10.1056/NEJM198212163072502

Gong, J., Savitz, D. a., Stein, C. R., & Engel, S. M. (2012). Maternal ethnicity and pre-eclampsia in New York City, 1995-2003. *Paediatric and perinatal epidemiology, 26*, 45-52. doi: 10.1111/j.1365-3016.2011.01222.x

Goodman, M. (1999). The genomic record of Humankind's evolutionary roots. *American journal of human genetics, 64*, 31.

Graunt, J. (1662). Natural and political observations mentioned in a following index, and made upon the bills of mortality.

Grossman, S. R., Andersen, K. G., Shlyakhter, I., Tabrizi, S., Winnicki, S., Yen, A., . . . Sabeti, P. C. (2013). Identifying Recent Adaptations in Large-Scale Genomic Data. *Cell, 152*, 703-713.

Grossman, S. R., Shylakhter, I., Karlsson, E. K., Byrne, E. H., Morales, S., Frieden, G., . . . Sabeti, P. C. (2010). A Composite of Multiple Signals Distinguishes Causal Variants in Regions of Positive Selection. *Science, 327*, 883-886.

Groves, B. M., Droma, T., Sutton, J. R., McCullough, R. G., McCullough, R. E., Zhuang, J., . . . Moore, L. G. (1993). Minimal hypoxic pulmonary hypertension in normal Tibetans at 3,658 m. *Journal of applied physiology (Bethesda, Md. : 1985), 74*, 312-318.

Guo, K., Yu, Y.-H., Hou, J., & Zhang, Y. (2010). Chronic leucine supplementation improves glycemic control in etiologically distinct mouse models of obesity and diabetes mellitus. *Nutrition & metabolism, 7*, 57. doi: 10.1186/1743-7075-7-57

Guzmán, M., & Blázquez, C. (2004). Ketone body synthesis in the brain: Possible neuroprotective effects. *Prostaglandins Leukotrienes and Essential Fatty Acids, 70*, 287-292. doi: 10.1016/j.plefa.2003.05.001

Gylfe, E. (1976). Comparison of the effects of leucines, non-metabolizable leucine analogues and other insulin secretagogues on the activity of glutamate dehydrogenase. *Acta diabetologica latina, 13*, 20-24. doi: 10.1007/BF02591577

Hales, C. N., & Barker, D. J. (2001). The thrifty phenotype hypothesis. *British medical bulletin, 60*, 5-20.

Hancock, A. M., Alkorta-Aranburu, G., Witonsky, D. B., & Di Rienzo, A. (2010). Adaptations to new environments in humans: the role of subtle allele frequency shifts. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 365*, 2459-2468. doi: 10.1098/rstb.2010.0032

Hancock, A. M., Clark, V. J., Qian, Y., & Di Rienzo, A. (2010a). Population Genetic Analysis of the Uncoupling Proteins Supports a Role for UCP3 in Human Cold Resistance. *Molecular biology and evolution, 28*, 601-614. doi: 10.1093/molbev/msq228

Hancock, A. M., Clark, V. J., Qian, Y., & Di Rienzo, A. (2010b). Population Genetic Analysis of the Uncoupling Proteins Supports a Role for UCP3 in Human Cold Resistance. *Molecular Biology and Evolution*, 1-47.

Hancock, A. M., Witonsky, D. B., Alkorta-Aranburu, G., Beall, C. M., Gebremedhin, A., Sukernik, R., . . . Di Rienzo, A. (2011a). Adaptations to Climate-Mediated Selective Pressures in Humans. *PLoS Genetics, 7*, e1001375. doi: 10.1371/journal.pgen.1001375

Hancock, A. M., Witonsky, D. B., Alkorta-Aranburu, G., Beall, C. M., Gebremedhin, A., Sukernik, R., . . . Di Rienzo, A. (2011b). Adaptations to Climate-Mediated Selective Pressures in Humans. *PLoS Genetics, 7*, 1-16.

Hancock, A. M., Witonsky, D. B., Ehler, E., Alkorta-Aranburu, G., Beall, C., Gebremedhin, A., . . . Di Rienzo, A. (2010). Human adaptations to diet, subsistence, and ecoregion are due to subtle shifts in allele frequency. *Proceedings of the National Academy of Sciences of the United States of America, 107*, 8924-8930. doi: 10.1073/pnas.0914625107

Hancock, A. M., Witonsky, D. B., Gordon, A. S., Eshel, G., Pritchard, J. K., Coop, G., & Di Rienzo, A. (2008). Adaptations to climate in candidate genes for common metabolic disorders. *PLoS Genetics, 4*, 1-13.

Harger, J. H., Ernest, J. M., Thurnau, G. R., Moawad, A., Momirova, V., Landon, M. B., . . . Van Dorsten, P. (2002). Risk factors and outcome of varicella-zoster virus pneumonia in pregnant women. *The Journal of infectious diseases, 185*, 422-427. doi: 10.1086/338832

Harris, J. (1919). Influenza occurring in pregnant women. *Journal of the American Medical Association*.

Hayes, M. G., Urbanek, M., Hivert, M.-F., Armstrong, L. L., Morrison, J., Guo, C., . . . Lowe, W. L. (2013). Identification of HKDC1 and BACE2 as Genes Influencing Glycemic Traits During Pregnancy Through Genome-Wide Association Studies. *Diabetes*. doi: 10.2337/db12-1692

He, Y., Tsuei, J., & Wan, Y.-j. Y. (2014). Biological functional annotation of retinoic acid alpha and beta in mouse liver based on genome-wide binding. *American Journal of Physiology. Gastrointestinal liver physiology, 307*, 205-218. doi: 10.1152/ajpgi.00105.2014

Helgason, A., Pálsson, S., Thorleifsson, G., Grant, S. F. a., Emilsson, V., Gunnarsdottir, S., . . . Stefánsson, K. (2007). Refining the impact of TCF7L2 gene variants on type 2 diabetes and adaptive evolution. *Nature genetics, 39*, 218-225. doi: 10.1038/ng1960

Henderson, J. B., Dunnigan, M. G., McIntosh, W. B., Abdul-Motaal, a. a., Gettinby, G., & Glekin, B. M. (1987). The importance of limited exposure to ultraviolet radiation and dietary factors in the aetiology of Asian rickets: a risk-factor model. *The Quarterly journal of medicine, 63*, 413-425.

Henquin, J. C., Gembal, M., Detimary, P., Gao, Z. Y., Warnotte, C., & Gilon, P. (1994). Multisite control of insulin release by glucose. *Diabete & metabolisme, 20*, 132-137.

Henry, C. J. K., Lightowler, H. J., Newens, K., Sudha, V., Radhika, G., Sathya, R. M., & Mohan, V. (2008). Glycaemic index of common foods tested in the UK and India. *The British journal of nutrition, 99*, 840-845. doi: 10.1017/S0007114507831801

Hermida, R. C., Ayala, D. E., Mojon, a., Fernandez, J. R., Alonso, I., Silva, I., . . . Iglesias, M. (2000). Blood Pressure Patterns in Normal Pregnancy, Gestational Hypertension, and Preeclampsia. *Hypertension, 36*, 149-158. doi: 10.1161/01.HYP.36.2.149

Hermisson, J., & Pennings, P. S. (2005). Soft sweeps: molecular population genetics of adaptation from standing genetic variation. *Genetics, 169*, 2335-2352. doi: 10.1534/genetics.104.036947

Hernandez, R. D., Kelley, J. L., Elyashiv, E., Melton, S. C., Auton, a., McVean, G., . . . Przeworski, M. (2011). Classic Selective Sweeps Were Rare in Recent Human Evolution. *Science, 331*, 920-924. doi: 10.1126/science.1198878

Heron, M. (2012). Deaths: leading causes for 2009. *National vital statistics reports : from the Centers for Disease Control and Prevention, National Center for Health Statistics, National Vital Statistics System, 61*, 1-95.

Herrera, E. (2002). Lipid metabolism in pregnancy and its consequences in the fetus and newborn. *Endocrine, 19*, 43-55. doi: 10.1385/ENDO:19:1:43

Heyer, E., Brazier, L., Ségurel, L., Hegay, T., Austerlitz, F., Quintana-Murci, L., . . . Veuille, M. (2011). Lactase persistence in central Asia: phenotype, genotype, and evolution. *Human biology, 83*, 379-392. doi: 10.3378/027.083.0304

Holick, M. F. (1987). Photosynthesis of vitamin D in the skin: effect of environmental and life-style variables. *Federation proceedings, 46*, 1876-1882.

Holick, M. F. (2004). Vitamin D: importance in the prevention of cancers, type 1 diabetes, heart disease, and osteoporosis. *The American journal of clinical nutrition, 79*, 362-371.

Holt, S., Miller, J., & Petocz, P. (1997). An insulin index of foods: the insulin demand generated by 1000-kJ portions of common foods. *The American journal of clinical*, 1264-1276.

Holzinger, A., Röschinger, W., Lagler, F., Mayerhofer, P. U., Lichtner, P., Kattenfeld, T., . . . Roscher, A. A. (2001). Cloning of the human MCCA and MCCB genes and mutations therein reveal the molecular cause of 3-methylcrotonyl-CoA: carboxylase deficiency. *Human molecular genetics, 10*, 1299-1306.

Hoppe, C., Mølgaard, C., Vaag, a., Barkholt, V., & Michaelsen, K. F. (2005). High intakes of milk, but not meat, increase s-insulin and insulin resistance in 8-year-old boys. *European journal of clinical nutrition, 59*, 393-398. doi: 10.1038/sj.ejcn.1602086

Horby, P., Nguyen, N. Y., Dunstan, S. J., & Baillie, J. K. (2012). The role of host genetics in susceptibility to influenza: a systematic review. *PloS one, 7*, e33180. doi: 10.1371/journal.pone.0033180

Hotelling, H., & Hotelling, F. (1931). Causes of birth rate fluctuations. *Journal of the American Statistical Association, 26*, 135-149.

Hoyt, G., Hickey, M. S., & Cordain, L. (2007). Dissociation of the glycaemic and insulinaemic responses to whole and skimmed milk. *British Journal of Nutrition, 93*, 175. doi: 10.1079/BJN20041304

Huber, M. D., Vesely, P. W., Datta, K., & Gerace, L. (2013). Erlins restrict SREBP activation in the ER and regulate cellular cholesterol homeostasis. *The Journal of cell biology, 203*, 427-436. doi: 10.1083/jcb.201305076

Huerta-Sánchez, E., DeGiorgio, M., Pagani, L., Tarekegn, A., Ekong, R., Antao, T., . . . Nielsen, R. (2013). Genetic Signatures Reveal High-Altitude Adaptation in a Set of Ethiopian Populations. *Molecular Biology and Evolution, 30*, 1877-1888. doi: 10.1093/molbev/mst089

Huerta-Sánchez, E., Jin, X., Bianba, Z., Peter, B. M., Vinckenbosch, N., Liang, Y., . . . Nielsen, R. (2014). Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature*. doi: 10.1038/nature13408

Hughes, D. A., Jastroch, M., Stoneking, M., & Klingenspor, M. (2009). Molecular evolution of UCP1 and the evolutionary history of mammalian non-shivering thermogenesis. *BMC evolutionary biology, 9*(4), 1-13.

Hunsberger, M., Rosenberg, K. D., & Donatelle, R. J. (2010). Racial/ethnic disparities in gestational diabetes mellitus: findings from a population-based survey. *Women's health issues : official publication of the Jacobs Institute of Women's Health, 20*, 323-328. doi: 10.1016/j.whi.2010.06.003

Ikeda, Y., & Tanaka, K. (1983). Purification and Characterization of 2-Methyl-branched Chain Acyl Coenzyme A Dehydrogenase, an Enzyme Involved in the Isoleucine and Valine Metabolism, from Rat Liver Mitochondria. *Journal of Biological Chemistry, 258*, 9477-9487.

Ingram, C. J. E., Mulcare, C. a., Itan, Y., Thomas, M. G., & Swallow, D. M. (2009). Lactose digestion and the evolutionary genetics of lactase persistence. *Human genetics, 124*, 579-591. doi: 10.1007/s00439-008-0593-6

INTERSALT. (1988). Intersalt: an international study of electrolyte excretion and blood pressure. Results for 24 hour urinary sodium and potassium excretion. Intersalt Cooperative Research Group. *BMJ (Clinical research ed.), 297*, 319-328.

Itan, Y., Jones, B. L., Ingram, C. J. E., Swallow, D. M., & Thomas, M. G. (2010). A worldwide correlation of lactase persistence phenotype and genotypes. *BMC evolutionary biology, 10*, 36-47. doi: 10.1186/1471-2148-10-36

Izagirre, N., García, I., Junquera, C., de la Rúa, C., & Alonso, S. (2006). A scan for signatures of positive selection in candidate loci for skin pigmentation in humans. *Molecular biology and evolution, 23*, 1697-1706. doi: 10.1093/molbev/msl030

Jablonski, N. G., & Chaplin, G. (2000). The evolution of human skin coloration. *Journal of Human Evolution, 39*, 57-106. doi: 10.1006/jhev.2000.0403

Jablonski, N. G., & Chaplin, G. (2010). Colloquium paper: human skin pigmentation as an adaptation to UV radiation. *Proceedings of the National Academy of Sciences of the United States of America, 107 Suppl* 8962-8968. doi: 10.1073/pnas.0914628107

James, A. H., Bushnell, C. D., Jamison, M. G., & Myers, E. R. (2005). Incidence and risk factors for stroke in pregnancy and the puerperium. *Obstetrics and gynecology, 106*, 509-516. doi: 10.1097/01.AOG.0000172428.78411.b0

Jobert, L., Skjeldam, H. K., Dalhus, B., Galashevskaya, A., Vågbø, C. B., Bjørås, M., & Nilsen, H. (2013). The Human Base Excision Repair Enzyme SMUG1 Directly Interacts with DKC1 and Contributes to RNA Quality Control. *Molecular Cell, 49*, 339-345. doi: 10.1016/j.molcel.2012.11.010

Jolly, M. C., Sebire, N. J., Harris, J. P., Regan, L., & Robinson, S. (2003). Risk factors for macrosomia and its clinical consequences: a study of 350,311 pregnancies. *European Journal of Obstetrics & Gynecology and Reproductive Biology, 111*, 9-14. doi: 10.1016/S0301-2115(03)00154-4

Kajimura, S., Seale, P., Kubota, K., Lunsford, E., Frangioni, J. V., Gygi, S. P., & Spiegelman, B. M. (2009). Initiation of myoblast to brown fat switch by a PRDM16-C/EBP-beta transcriptional complex. *Nature, 460*, 1154-1158. doi: 10.1038/nature08262

Kamberov, Y. G., Wang, S., Tan, J., Gerbault, P., Wark, A., Tan, L., . . . Sabeti, P. C. (2013). Modeling recent human evolution in mice by expression of a selected EDAR variant. *Cell, 152*, 691-702. doi: 10.1016/j.cell.2013.01.016

Kathiresan, S., Willer, C., Peloso, G., Demissie, S., & K. (2008). Common variants at 30 loci contribute to polygenic dyslipidemia. *Nature Genetics, 41*, 56-65.

Katz, M. L., & Bergman, E. N. (1969). Hepatic and portal metabolism of glucose, free fatty acids, and ketone bodies in the sheep. *The American journal of physiology, 216*, 953-960.

Katzmarzyk, P. T., & Leonard, W. R. (1998). Climatic influences on human body size and proportions: ecological adaptations and secular trends. *American journal of physical anthropology, 106*, 483-503. doi: 10.1002/(SICI)1096-8644(199808)106:4<483::AID-AJPA4>3.0.CO;2-K

Kaufmann, P., Mayhew, T. M., & Charnock-Jones, D. S. (2004). Aspects of human fetoplacental vasculogenesis and angiogenesis. II. Changes during normal pregnancy. *Placenta, 25*, 114-126. doi: 10.1016/j.placenta.2003.10.009

Kelley, J. L., Madeoy, J., Calhoun, J. C., Swanson, W. J., & Akey, J. M. (2006). Genomic signatures of positive selection in humans and the limits of outlier approaches. *Genome research, 16*, 980-989. doi: 10.1101/gr.5157306

Kelley, J. L., Turkheimer, K., Haney, M., & Swanson, W. J. (2009). Targeted resequencing of two genes, RAGE and POLL, confirms findings from a genome-wide scan for adaptive evolution and provides evidence for positive selection in additional populations. *Human molecular genetics, 18*, 779-784. doi: 10.1093/hmg/ddn399

Kim, H., Toyofuku, Y., Lynn, F. C., Chak, E., Uchida, T., Mizukami, H., . . . German, M. S. (2010). Serotonin regulates pancreatic beta cell mass during pregnancy. *Nature medicine, 16*, 804-808. doi: 10.1038/nm.2173

Klimentidis, Y. C., Abrams, M., Wang, J., Fernandez, J. R., & Allison, D. B. (2011). Natural selection at genomic regions associated with obesity and type-2 diabetes: East Asians and sub-Saharan Africans exhibit high levels of differentiation at type-2 diabetes regions. *Human Genetics, 129*(4), 407-418.

Kooner, J. S., Saleheen, D., Sim, X., Sehmi, J., Zhang, W., Frossard, P., . . . Chambers, J. C. (2011). Genome-wide association study in individuals of South Asian ancestry identifies six new type 2 diabetes susceptibility loci. *Nature Genetics, 43*, 984-989. doi: 10.1038/ng.921

Kuzawa, C. W. (1998). Adipose tissue in human infancy and childhood: an evolutionary perspective. *American Journal of Physical Anthropology, 107*, 177-209.

Laland, K. N., Odling-Smee, J., & Myles, S. (2010). How culture shaped the human genome: bringing genetics and the human sciences together. *Nature reviews. Genetics, 11*, 137-148. doi: 10.1038/nrg2734

Langer, O., Yogev, Y., Most, O., & Xenakis, E. M. J. (2005). Gestational diabetes: the consequences of not treating. *American journal of obstetrics and gynecology, 192*, 989-997. doi: 10.1016/j.ajog.2004.11.039

Lao, O., de Gruijter, J. M., van Duijn, K., Navarro, a., & Kayser, M. (2007). Signatures of positive selection in genes associated with human skin pigmentation as revealed from analyses of single nucleotide polymorphisms. *Annals of human genetics, 71*, 354-369. doi: 10.1111/j.1469-1809.2006.00341.x

Lappalainen, T., Sammeth, M., Friedländer, M. R., 't Hoen, P. A. C., Monlong, J., Rivas, M. a., . . . Dermitzakis, E. T. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*. doi: 10.1038/nature12531

Layden, B. T., Durai, V., Newman, M. V., Marinelarena, a. M., Ahn, C. W., Feng, G., . . . Lowe, W. L. (2010). Regulation of pancreatic islet gene expression in mouse islets by pregnancy. *Journal of Endocrinology, 207*, 265-279. doi: 10.1677/JOE-10-0298

Lehnert, W., Scharf, J., & Wendel, U. (1985). 3-Methylglutaconic and 3-methylglutaric aciduria in a patient with suspected 3-methylglutaconyl-CoA hydratase deficiency. *European Journal of Pediatrics, 143*, 301-303. doi: 10.1007/BF00442306

Li, F., Yin, Y., Tan, B., Kong, X., & Wu, G. (2011). Leucine nutrition in animals and humans: MTOR signaling and beyond. *Amino Acids, 41*, 1185-1193. doi: 10.1007/s00726-011-0983-2

Lieberman, D. E., Pilbeam, D. R., & Wrangham, R. W. (2009). The Transition from Australopithecus to Homo *Transitions in Prehistory: Essays in Honor of Ofer Bar-Yosef Shea JJ* (pp. 1-22). Oxford: Oxbow Publications.

Liu, J., Rotkirch, A., & Lummaa, V. (2012). Maternal risk of breeding failure remained low throughout the demographic transitions in fertility and age at first reproduction in Finland. *PloS one, 7*, e34898. doi: 10.1371/journal.pone.0034898

Lohmueller, K. E., Albrechtsen, A., Li, Y., Kim, S. Y., Korneliussen, T., Vinckenbosch, N., . . . Nielsen, R. (2011). Natural Selection Affects Multiple Aspects of Genetic Variation at Putatively Neutral Sites across the Human Genome. *PLoS Genetics, 7*, 1-15. doi: 10.1371/journal.pgen.1002326

Lohmueller, K. E., Bustamante, C. D., & Clark, A. G. (2010). Detecting Directional Selection in the Presence of Recent Admixture in African Americans. *Genetics, 187*(3), 823-835.

Lopes-Cardozo, M., Mulder, I., van Vugt, F., Hermans, P. G. C., van den Bergh, S. G., Klazinga, W., & de Vries-Akkerman, E. (1975). Aspects of ketogenesis: Control and mechanism of ketone-body formation in isolated RAT-liver mitochondria. *Molecular and Cellular Biochemistry, 9*, 155-173. doi: 10.1007/BF01751311

López Herráez, D., Bauchet, M., Tang, K., Theunert, C., Pugach, I., Li, J., . . . Stoneking, M. (2009). Genetic variation and recent positive selection in worldwide human populations: evidence from nearly 1 million SNPs. *PLoS ONE, 4*(11), 1-16.

Luca, F., Bubba, G., Basile, M., Brdicka, R., Michalodimitrakis, E., Rickards, O., . . . Novelletto, A. (2008). Multiple advantageous amino acid variants in the NAT2 gene in human populations. *PLoS ONE, 3*(9), 1-12.

Lynch, C. J. (2001). Role of leucine in the regulation of mTOR by amino acids: revelations from structure-activity studies. *The Journal of nutrition, 131*, 861S-865S. doi: 10.1021/bi0603625

Macaulay, V., Hill, C., Achilli, A., Rengo, C., Clarke, D., Meehan, W., . . . others. (2005). Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. *Science, 308*, 1034. doi: 10.1126/science.1109792

Makarova, O., Kamberov, E., & Margolis, B. (2000). Generation of deletion and point mutations with one primer in a single cloning step. *BioTechniques, 29*, 970-972.

Manolio, T. a., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. a., Hunter, D. J., . . . Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature, 461*, 747-753. doi: 10.1038/nature08494

Marlowe, F. W. (2005). Hunter-gatherers and human evolution. *Evolutionary Anthropology: Issues, News, and Reviews, 14*, 54-67. doi: 10.1002/evan.20046

McKeigue, P., Shah, B., & Marmot, M. (1991). Relation of central obesity and insulin resistance with high diabetes prevalence and cardiovascular risk in South Asians. *The Lancet*, 382-386.

Mejía, O. M., Prchal, J. T., León-Velarde, F., Hurtado, A., & Stockton, D. W. (2005). Genetic association analysis of chronic mountain sickness in an Andean high-altitude population. *Haematologica, 90*, 13-19.

Melichar, H. J., Narayan, K., Der, S. D., Hiraoka, Y., Gardiol, N., Jeannet, G., . . . Kang, J. (2007). Regulation of gd Versus ab Transcription Factor SOX13. *Science, 315*, 230-233.

Mendizabal, I., Marigorta, U. M., Lao, O., & Comas, D. (2012). Adaptive evolution of loci covarying with the human African Pygmy phenotype. *Human Genetics, 131*, 1305-1317. doi: 10.1007/s00439-012-1157-3

Metzger, B. E., Buchanan, T. a., Coustan, D. R., de Leiva, A., Dunger, D. B., Hadden, D. R., . . . Zoupas, C. (2007). Summary and recommendations of the Fifth International Workshop-Conference on Gestational Diabetes Mellitus. *Diabetes care, 30 Suppl 2*, S251-260. doi: 10.2337/dc07-s225

Migliano, A. B., Romero, I. G., Metspalu, M., Leavesley, M., Pagani, L., Antao, T., . . . Kivisild, T. (2013). Evolution of the pygmy phenotype: evidence of positive selection fro genome-wide scans in African, Asian, and Melanesian pygmies. *Human biology, 85*, 251-284.

Miller, M. (2010). Managing mixed dyslipidemia in special populations. *Preventive cardiology, 13*, 78-83. doi: 10.1111/j.1751-7141.2009.00057.x

Millington, D. S., Roe, C. R., Maltby, D. a., & Inoue, F. (1987). Endogenous catabolism is the major source of toxic metabolites in isovaleric acidemia. *The Journal of pediatrics, 110*, 56-60. doi: 10.1016/S0022-3476(87)80288-3

Miura, K., Okuda, N., Turin, T. C., Takashima, N., Nakagawa, H., Nakamura, K., . . . Ueshima, H. (2010). Dietary Salt Intake and Blood Pressure in a Representative Japanese Population: Baseline Analyses of NIPPON DATA80. *Journal of Epidemiology, 20*, S524-S530. doi: 10.2188/jea.JE20090220

Moodley, J. (2008). Maternal deaths due to hypertensive disorders in pregnancy. *Best practice & research. Clinical obstetrics & gynaecology, 22*, 559-567. doi: 10.1016/j.bpobgyn.2007.11.004

Moore, L. G., Zamudio, S., Zhuang, J., Sun, S., & Droma, T. (2001). Oxygen transport in tibetan women during pregnancy at 3,658 m. *American journal of physical anthropology, 114*, 42-53. doi: 10.1002/1096-8644(200101)114:1<42::AID-AJPA1004>3.0.CO;2-B

Moreau, C., Bhérer, C., Vézina, H., Jomphe, M., Labuda, D., & Excoffier, L. (2011). Deep human genealogies reveal a selective advantage to be on an expanding wave front. *Science (New York, N.Y.), 334*, 1148-1150. doi: 10.1126/science.1212880

Muehlenbachs, A., Fried, M., Lachowitzer, J., Mutabingwa, T. K., & Duffy, P. E. (2008). Natural selection of FLT1 alleles and their association with malaria resistance in utero. *Proceedings of the National Academy of Sciences of the United States of America, 105*, 14488-14491. doi: 10.1073/pnas.0803657105

Mupapa, K., Mukundu, W., Bwaka, M. a., Kipasa, M., De Roo, a., Kuvula, K., . . . Muyembe-Tamfum, J. J. (1999). Ebola hemorrhagic fever and pregnancy. *The Journal of infectious diseases, 179 Suppl* S11-12. doi: 10.1086/514289

Murdock, G. P. (1967). *Ethnographic Atlas*. Pittsburgh: University of Pittsburgh Press.

Myles, S., Bouzekri, N., Haverfield, E., Cherkaoui, M., Dugoujon, J.-M., & Ward, R. (2005). Genetic evidence in support of a shared Eurasian-North African dairying origin. *Human genetics, 117*, 34-42. doi: 10.1007/s00439-005-1266-3

Nagao, M., Parimoo, B., & Tanaka, K. (1993). Developmental, nutritional, and hormonal regulation of tissue-specific expression of the genes encoding various acyl-CoA dehydrogenases and alpha-subunit of electron transfer flavoprotein in rat. *The Journal of biological chemistry, 268*, 24114-24124.

Neel, J. V. (1962). Diabetes mellitus: a "thrifty" genotype rendered detrimental by "progress"? *American Journal of Human Genetics, 14*, 353-362.

Neel, J. V. (1999). The "thrifty genotype" in 1998. *Nutrition reviews, 57*, S2-S9.

Newman, C., Wilson, B., Callaghan, P., & Young, L. (1967). Neonatal death associated with isovalericacidaemia. *The Lancet*, 439-442.

Nielsen, R. (2010). In search of rare human variants. *Nature, 467*, 1050-1051.

Nielsen, R., Hellmann, I., Hubisz, M., Bustamante, C., & Clark, A. G. (2007). Recent and ongoing selection in the human genome. *Nature reviews. Genetics, 8*, 857-868. doi: 10.1038/nrg2187

Nielsen, R., Williamson, S., Kim, Y., Hubisz, M. J., Clark, A. G., & Bustamante, C. (2005). Genomic scans for selective sweeps using SNP data. *Genome research, 15*, 1566-1575. doi: 10.1101/gr.4252305

Norman, A. W. (2008). From vitamin D to hormone D: fundamentals of the vitamin D endocrine system essential for good health. *The American journal of clinical nutrition, 88*, 491S-499S.

Norton, H. L., Kittles, R. a., Parra, E., McKeigue, P., Mao, X., Cheng, K., . . . Shriver, M. D. (2007). Genetic evidence for the convergent evolution of light skin in Europeans and East Asians. *Molecular biology and evolution, 24*, 710-722. doi: 10.1093/molbev/msl203

Biochemical controls of liver cholesterol biosynthesis, 34 2295-2306 (1981).

Panini, S. R., Schnitzer-Polokoff, R., Spencer, T. a., & Sinensky, M. (1989). Sterol-independent regulation of 3-hydroxy-3-methylglutaryl-CoA reductase by mevalonate in Chinese hamster ovary cells. Magnitude and specificity. *Journal of Biological Chemistry, 264*, 11044-11052.

Panter-Brick, C., Lotstein, D. S., & Ellison, P. T. (1993). Seasonality of reproductive function and weight loss in rural Nepali women. *Human reproduction (Oxford, England), 8*, 684-690.

Pazos, M., Sperling, R. S., Moran, T. M., & Kraus, T. a. (2012). The influence of pregnancy on systemic immunity. *Immunologic research*. doi: 10.1007/s12026-012-8303-9

Peng, M.-S., He, J.-D., Zhu, C.-L., Wu, S.-F., Jin, J.-Q., & Zhang, Y.-P. (2012). Lactase persistence may have an independent origin in Tibetan populations from Tibet, China. *Journal of human genetics, 57*, 394-397. doi: 10.1038/jhg.2012.41

Perry, G. H., Dominy, N. J., Claw, K. G., Lee, A. S., Fiegler, H., Redon, R., . . . Stone, A. C. (2007). Diet and the evolution of human amylase gene copy number variation. *Nature genetics, 39*, 1256-1260. doi: 10.1038/ng2123

Perry, G. H., Foll, M., Grenier, J.-C., Patin, E., Nédélec, Y., Pacis, A., . . . Barreiro, L. B. (2014). Adaptive, convergent origins of the pygmy phenotype in African rainforest hunter-gatherers. *Proceedings of the National Academy of Sciences of the United States of America*. doi: 10.1073/pnas.1402875111

Perry, J. R. B., & Frayling, T. M. (2008). New gene variants alter type 2 diabetes risk predominantly through reduced beta-cell function. *Current opinion in clinical nutrition and metabolic care, 11*, 371-377. doi: 10.1097/MCO.0b013e32830349a1

CTCF: Master Weaver of the Genome, 137 1194-1211 (2009).

Pollard, T. M. (2008). The thrifty genotype versus thrifty phenotype debate: efforts to explain between population variation in rates of type 2 diabetes and cardiovascular disease. *Western Diseases: An Evolutionary Perspective* (pp. 50-74). Cambridge: Cambridge University Press.

Pouillot, R., Hoelzer, K., Jackson, K. a., Henao, O. L., & Silk, B. J. (2012). Relative risk of listeriosis in Foodborne Diseases Active Surveillance Network (FoodNet) sites according to age, pregnancy, and ethnicity. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America, 54 Suppl 5*, S405-410. doi: 10.1093/cid/cis269

Price, M. E., Fisher-Hoch, S. P., Craven, R. B., & McCormick, J. B. (1988). A prospective study of maternal and fetal outcome in acute Lassa fever infection during pregnancy. *BMJ (Clinical research ed.), 297*, 584-587.

Pritchard, J. (1965). Changes in the blood volume during pregnancy and delivery. *Anesthesiology*.

Pritchard, J. K., Pickrell, J. K., & Coop, G. (2010). The Genetics of Human Adaptation: Hard Sweeps, Soft Sweeps, and Polygenic Adaptation. *Current Biology, 20*, 208-215. doi: 10.1016/j.cub.2009.11.055

Przeworski, M., Coop, G., & Wall, J. D. (2005). The Signature of Positive Selection on Standing Genetic Variation. *Evolution, 59*(11), 2312-2323. doi: 10.1554/05-273.1

Quillen, E. E., Bauchet, M., Bigham, A. W., Delgado-Burbano, M. E., Faust, F. X., Klimentidis, Y. C., . . . Shriver, M. D. (2012). OPRM1 and EGFR contribute to skin pigmentation differences between Indigenous Americans and Europeans. *Human genetics, 131*, 1073-1080. doi: 10.1007/s00439-011-1135-1

Rachdi, L., Aïello, V., Duvillié, B., & Scharfmann, R. (2012). L-Leucine alters pancreatic β-cell differentiation and function via the mTor signaling pathway. *Diabetes, 61*, 409-417. doi: 10.2337/db11-0765

Rajaratnam, J. K., Marcus, J. R., Flaxman, A. D., Wang, H., Levin-Rector, A., Dwyer, L., . . . Murray, C. J. L. (2010). Neonatal, postneonatal, childhood, and under-5 mortality for

187 countries, 1970-2010: a systematic analysis of progress towards Millennium Development Goal 4. *Lancet, 375*, 1988-2008. doi: 10.1016/S0140-6736(10)60703-9

Raji, a., Seely, E. W., Arky, R. a., & Simonson, D. C. (2001). Body fat distribution and insulin resistance in healthy Asian Indians and Caucasians. *The Journal of clinical endocrinology and metabolism, 86*, 5366-5371.

Reyes, L., Garcia, R., Ruiz, S., Dehghan, M., & López-Jaramillo, P. (2012). Nutritional status among women with pre-eclampsia and healthy pregnant and non-pregnant women in a Latin American country. *The journal of obstetrics and gynaecology research, 38*, 498-504. doi: 10.1111/j.1447-0756.2011.01763.x

Rieck, S., White, P., Schug, J., Fox, A. J., Smirnova, O., Gao, N., . . . Kaestner, K. H. (2009). The transcriptional response of the islet to pregnancy in mice. *Molecular endocrinology (Baltimore, Md.), 23*, 1702-1712. doi: 10.1210/me.2009-0144

Roberts, C., & Walker, W. (2001). Sex-Associated Hormones and Immunity to Protozoan Parasites. *Clinical Microbiology, 14*. doi: 10.1128/CMR.14.3.476

Roberts, D. F. (1953). Body weight, race and climate. *American journal of physical anthropology, 11*, 533-558.

Robinson, D. P., & Klein, S. L. (2012). Pregnancy and pregnancy-associated hormones alter immune responses and disease pathogenesis. *Hormones and behavior*. doi: 10.1016/j.yhbeh.2012.02.023

Rodriguez, a. I., Csanyi, G., Ranayhossaini, D. J., Feck, D. M., Blose, K. J., Assatourian, L., . . . Pagano, P. J. (2014). MEF2B-Nox1 Signaling Is Critical for Stretch-Induced Phenotypic Modulation of Vascular Smooth Muscle Cells. *Arteriosclerosis, Thrombosis, and Vascular Biology, 35*, 430-438. doi: 10.1161/ATVBAHA.114.304936

Romeo, S., Kozlitina, J., Xing, C., Pertsemlidis, A., Cox, D., Pennacchio, L. a., . . . Hobbs, H. H. (2008). Genetic variation in PNPLA3 confers susceptibility to nonalcoholic fatty liver disease. *Nature genetics, 40*, 1461-1465. doi: 10.1038/ng.257

Romeo, S., Pennacchio, L. A., Fu, Y., Boerwinkle, E., Tybjaerg-Hansen, A., Hobbs, H. H., & Cohen, J. C. (2007). Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL. *Nature genetics, 39*, 513-516. doi: 10.1038/ng1984

Ronsmans, C., & Graham, W. J. (2006). Maternal mortality: who, when, where, and why. *Lancet, 368*, 1189-1200. doi: 10.1016/S0140-6736(06)69380-X

Roos, S., Jansson, N., Palmberg, I., Säljö, K., Powell, T. L., & Jansson, T. (2007). Mammalian target of rapamycin in the human placenta regulates leucine transport and is down-

regulated in restricted fetal growth. *The Journal of physiology, 582*, 449-459. doi: 10.1113/jphysiol.2007.129676

Rosenbaum, M., & Leibel, R. L. (1999). The Role of Leptin in Human Physiology. *New England Journal of Medicine, 341*, 913-915. doi: 10.1016/S0735-1097(99)00014-5

Rosenberg, K., & Trevathan, W. (2002). Birth, obstetrics and human evolution. *BJOG : an international journal of obstetrics and gynaecology, 109*, 1199-1206.

Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., . . . Stewart, J. (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature, 449*(7164), 913-918. doi: 10.1038/nature06250

Savitz, D. a., Janevic, T. M., Engel, S. M., Kaufman, J. S., & Herring, a. H. (2008). Ethnicity and gestational diabetes in New York City, 1995-2003. *BJOG : an international journal of obstetrics and gynaecology, 115*, 969-978. doi: 10.1111/j.1471-0528.2008.01763.x

Sazzini, M., Schiavo, G., De Fanti, S., Martelli, P. L., Casadio, R., & Luiselli, D. (2014). Searching for signatures of cold adaptations in modern and archaic humans: hints from the brown adipose tissue genes. *Heredity, 113*, 259-267. doi: 10.1038/hdy.2014.24

Scheinfeldt, L. B., Biswas, S., Madeoy, J., Connelly, C. F., Schadt, E. E., & Akey, J. M. (2009). Population genomic analysis of ALMS1 in humans reveals a surprisingly complex evolutionary history. *Molecular biology and evolution, 26*, 1357-1367. doi: 10.1093/molbev/msp045

Scheinfeldt, L. B., Soi, S., Thompson, S., Ranciaro, A., Woldemeskel, D., Beggs, W., . . . Tishkoff, S. a. (2012). Genetic adaptation to high altitude in the Ethiopian highlands. *Genome biology, 13*, R1. doi: 10.1186/gb-2012-13-1-r1

Seale, P., Kajimura, S., & Spiegelman, B. M. (2009). Transcriptional control of brown adipocyte development and physiological function--of mice and men. *Genes & development, 23*, 788-797. doi: 10.1101/gad.1779209

Seidell, J. C. (2000). Obesity, insulin resistance and diabetes--a worldwide epidemic. *The British journal of nutrition, 83 Suppl 1*, S5-8.

Sermer, M., Naylor, C. D., Farine, D., Kenshole, A. B., Ritchie, J. W., Gare, D. J., . . . Biringer, A. (1998). The Toronto Tri-Hospital Gestational Diabetes Project. A preliminary review. *Diabetes care, 21 Suppl 2*, B33-42.

Shin, S.-Y., Fauman, E. B., Petersen, A.-K., Krumsiek, J., Santos, R., Huang, J., . . . Soranzo, N. (2014). An atlas of genetic influences on human blood metabolites. *Nature genetics, 46*. doi: 10.1038/ng.2982

Sholtis, S. J., & Noonan, J. P. (2010). Gene regulation and the origins of human biological uniqueness. *Trends in genetics : TIG, 26*, 110-118. doi: 10.1016/j.tig.2009.12.009

Shu, X. O., Long, J., Cai, Q., Qi, L., Xiang, Y.-B., Cho, Y. S., . . . Hu, F. B. (2010). Identification of new genetic risk variants for type 2 diabetes. *PLoS genetics, 6*, 1-8. doi: 10.1371/journal.pgen.1001127

Shulman, C. (2003). Importance and prevention of malaria in pregnancy. *Transactions of the Royal Society of Tropical*, 30-35.

Sim, X., Ong, R. T.-H., Suo, C., Tay, W.-T., Liu, J., Ng, D. P.-K., . . . Tai, E.-S. (2011). Transferability of type 2 diabetes implicated loci in multi-ethnic cohorts from Southeast Asia. *PLoS genetics, 7*, 1-12. doi: 10.1371/journal.pgen.1001363

Simonson, T. S., Yang, Y., Huff, C. D., Yun, H., Qin, G., Witherspoon, D. J., . . . Ge, R. (2010a). Genetic Evidence for High-Altitude Adaptation in Tibet. *Science (New York, N.Y.), 329*, 72-75.

Simonson, T. S., Yang, Y., Huff, C. D., Yun, H., Qin, G., Witherspoon, D. J., . . . Ge, R. (2010b). Genetic Evidence for High-Altitude Adaptation in Tibet. *Science (New York, N.Y.), 72*. doi: 10.1126/science.1189406

Sladek, S. M., Magness, R. R., & Conrad, K. P. (1997). Nitric oxide and pregnancy. *The American journal of physiology, 272*, R441-463.

Smith, J. M., & Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genetical Research, 23*(1), 23-25.

Snodgrass, J. J., Sorensen, M. V., Tarskaia, L. A., & Leonard, W. R. (2007). Adaptive dimensions of health research among indigenous Siberians. *American Journal of Human Biology, 19*, 165-180. doi: 10.1002/ajhb

Speliotes, E. K., Willer, C. J., Berndt, S. I., Monda, K. L., Thorleifsson, G., Jackson, A. U., . . . Loos, R. J. F. (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature Genetics, 15*, 1-12.

Steindal, A. H., Tam, T. T. T., Lu, X. Y., Juzeniene, A., & Moan, J. (2008). 5-Methyltetrahydrofolate is photosensitive in the presence of riboflavin. *Photochemical & Photobiological Sciences, 7*, 814. doi: 10.1039/b718907a

Steketee, R. W., Nahlen, B. L., Parise, M. E., & Menendez, C. (2001). The burden of malaria in pregnancy in malaria-endemic areas. *The American journal of tropical medicine and hygiene, 64*, 28-35.

Stephens, J. C., Reich, D. E., Goldstein, D. B., Shin, H. D., Smith, M. W., Carrington, M., . . . Dean, M. (1998). Dating the origin of the CCR5-Delta32 AIDS-resistance allele by the

coalescence of haplotypes. *American journal of human genetics, 62*, 1507-1515. doi: 10.1086/301867

Stöger, R. (2008). The thrifty epigenotype: an acquired and heritable predisposition for obesity and diabetes? *BioEssays: news and reviews in molecular, cellular and developmental biology, 30*(2), 156-166.

Storz, J. F. (2010). Genes for High Altitudes. *Science (New York, N.Y.), 329*, 40-41. doi: 10.1126/science.1192481

Ströhle, A., & Hahn, A. (2011). Diets of modern hunter-gatherers vary substantially in their carbohydrate content depending on ecoenvironments: results from an ethnographic analysis. *Nutrition research (New York, N.Y.), 31*, 429-435. doi: 10.1016/j.nutres.2011.05.003

Suhre, K., Shin, S.-Y., Petersen, A.-K., Mohney, R. P., Meredith, D., Wägele, B., . . . Gieger, C. (2011). Human metabolic individuality in biomedical and pharmaceutical research. *Nature, 477*, 54-60. doi: 10.1038/nature10354

Sun, X., & Zemel, M. B. (2007). Leucine and calcium regulate fat metabolism and energy partitioning in murine adipocytes and muscle cells. *Lipids, 42*, 297-305. doi: 10.1007/s11745-007-3029-5

Suryawan, A., Jeyapalan, A. S., Orellana, R. a., Wilson, F. a., Nguyen, H. V., & Davis, T. a. (2008). Leucine stimulates protein synthesis in skeletal muscle of neonatal pigs by enhancing mTORC1 activation. *American journal of physiology. Endocrinology and metabolism, 295*, E868-E875. doi: 10.1152/ajpendo.90314.2008

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics, 123*(3), 585-595.

Takeuchi, F., Serizawa, M., Yamamoto, K., Fujisawa, T., Nakashima, E., Ohnaka, K., . . . Takayanagi, R. (2009). Confirmation of Multiple Risk Loci and Genetic Impacts by a Genome-Wide Association Study of Type 2 Diabetes in the Japanese Population. *Diabetes, 58*, 1690-1699. doi: 10.2337/db08-1494.

Tanaka, K. (1990). Isovaleric acidemia: personal history, clinical survey and study of the molecular basis. *Progress in Clinical and Biological Research, 321*, 273-290.

Tanaka, K., & Budd, M. (1966). Isovaleric acidemia: a new genetic defect of leucine metabolism. *Proceedings of the …*, 236-242.

Tang, H., Jorgenson, E., Gadde, M., Kardia, S. L. R., Rao, D. C., Zhu, X., . . . Risch, N. (2006). Racial admixture and its impact on BMI and blood pressure in African and Mexican Americans. *Human genetics, 119*, 624-633. doi: 10.1007/s00439-006-0175-4

Tang, Q., Chen, Y., Meyer, C., Geistlinger, T., Lupien, M., Wang, Q., . . . Liu, X. S. (2011). A comprehensive view of nuclear receptor cancer cistromes. *Cancer Research, 71*, 6940-6947. doi: 10.1158/0008-5472.CAN-11-2091

Taubenberger, J. K., & Morens, D. M. (2006). 1918 Influenza: the mother of all pandemics. *Emerging infectious diseases, 12*, 15-22.

Teslovich, T. M., Musunuru, K., Smith, A. V., Edmondson, A. C., Stylianou, I. M., Koseki, M., . . . Kathiresan, S. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. *Nature, 466*, 707-713.

Thompson, E. E., Kuttab-Boulos, H., Witonsky, D. B., Yang, L., Roe, B. A., & Di Rienzo, A. (2004). CYP3A variation and the evolution of salt-sensitivity variants. *American journal of human genetics, 75*, 1059-1069. doi: 10.1086/426406

Tishkoff, S. A., Reed, F. A., Ranciaro, A., Voight, B. F., Babbitt, C. C., Silverman, J. S., . . . Deloukas, P. (2007a). Convergent adaptation of human lactase persistence in Africa and Europe. *Nature genetics, 39*, 31-40. doi: 10.1038/ng1946

Tishkoff, S. A., Reed, F. A., Ranciaro, A., Voight, B. F., Babbitt, C. C., Silverman, J. S., . . . Deloukas, P. (2007b). Convergent adaptation of human lactase persistence in Africa and Europe. Supplementary Table 1. *Nature genetics, 39*.

Tsukiyama-Kohara, K., Poulin, F., Kohara, M., DeMaria, C. T., Cheng, a., Wu, Z., . . . Sonenberg, N. (2001). Adipose tissue reduction in mice lacking the translational inhibitor 4E-BP1. *Nature medicine, 7*, 1128-1132. doi: 10.1038/nm1001-1128

Udpa, N., Ronen, R., Zhou, D., Liang, J., Stobdan, T., Appenzeller, O., . . . Haddad, G. G. (2014). Whole genome sequencing of Ethiopian highlanders reveals conserved hypoxia tolerance genes. *Genome biology, 15*, R36. doi: 10.1186/gb-2014-15-2-r36

Verdu, P., Austerlitz, F., Estoup, A., Vitalis, R., Georges, M., Théry, S., . . . Heyer, E. (2009). Origins and Genetic Diversity of Pygmy Hunter-Gatherers from Western Central Africa. *Current Biology, 19*, 312-318. doi: 10.1016/j.cub.2008.12.049

Vockley, J., & Ensenauer, R. (2006). Isovaleric acidemia: new aspects of genetic and phenotypic heterogeneity. *American journal of medical genetics. Part C, Seminars in medical genetics, 142C*, 95-103. doi: 10.1002/ajmg.c.30089

Voight, B. F., Kudaravalli, S., Wen, X., & Pritchard, J. K. (2006). A map of recent positive selection in the human genome. *PLoS Biology, 4*(3), 446-458.

Wan, J.-P., Wang, H., Li, C.-Z., Zhao, H., You, L., Shi, D.-H., . . . Chen, Z.-J. (2014). The Common Single-Nucleotide Polymorphism rs2681472 Is Associated With Early-Onset Preeclampsia in Northern Han Chinese Women. *Reproductive Sciences, 21*, 1423-1427. doi: 10.1177/1933719114527354

Wan, J.-p., Zhao, H., Li, T., Li, C.-z., Wang, X.-t., & Chen, Z.-j. (2013). The Common Variant rs11646213 Is Associated with Preeclampsia in Han Chinese Women.pdf. *PloS one, 8*, 6-9. doi: 10.1371/journal.pone.0071202

Wang, E. T., Kodama, G., Baldi, P., & Moyzis, R. K. (2006). Global landscape of recent inferred Darwinian selection for Homo sapiens. *Proceedings of the National Academy of Sciences of the United States of America, 103*, 135-140. doi: 10.1073/pnas.0509691102

Wang, Q., Prabhakar, S., Moses, A. M., Chanan, S., Brown, M., Eisen, M. B., . . . Boffelli, D. (2006). Primate-specific evolution of an LDLR enhancer. *Genome biology, 7*, 1-9.

Wang, Y., & Beydoun, M. A. (2007). The obesity epidemic in the United States--gender, age, socioeconomic, racial/ethnic, and geographic characteristics: a systematic review and meta-regression analysis. *Epidemiologic reviews, 29*, 6-28. doi: 10.1093/epirev/mxm007

Wang, Y., Liu, J., Liu, C., Naji, A., & Stoffers, D. a. (2013). MicroRNA-7 regulates the mTOR pathway and proliferation in adult pancreatic b-cells. *Diabetes, 62*, 887-895. doi: 10.2337/db12-0451

Wells, J. C. K. (2006). The evolution of human fatness and susceptibility to obesity: an ethological approach. *Biological reviews of the Cambridge Philosophical Society, 81*, 183-205. doi: 10.1017/S1464793105006974

WHO. (2005). *World Health Report: Make every mother and child count*, Geneva.

WHO. (2010). *Global status report on non-communicable diseases 2010*, Geneva.

Williams, A. L., Jacobs, S. B. R., Moreno-Macías, H., Huerta-Chagoya, A., Churchhouse, C., Márquez-Luna, C., . . . Cortes, M. L. (2013). Sequence variants in SLC16A11 are a common risk factor for type 2 diabetes in Mexico. *Nature*. doi: 10.1038/nature12828

Williams, R. C., Long, J. C., Hanson, R. L., Sievers, M. L., & Knowler, W. C. (2000). Individual Estimates of European Genetic Admixture Associated with Lower Body-Mass Index, Plasma Glucose, and Prevalence of Type 2 Diabetes in Pima Indians. *The American Journal of Human Genetics, 66*, 527-538.

Wilson, M. J., Lopez, M., Vargas, M., Julian, C., Tellez, W., Rodriguez, A., . . . Moore, L. G. (2007). Greater uterine artery blood flow during pregnancy in multigenerational (Andean) than shorter-term (European) high-altitude residents. *American journal of physiology. Regulatory, integrative and comparative physiology, 293*, R1313-1324. doi: 10.1152/ajpregu.00806.2006

Wilson, P. W., Anderson, K. M., & Castelli, W. P. (1991). Twelve-year incidence of coronary heart disease in middle-aged adults during the era of hypertensive therapy: the Framingham offspring study. *American Journal of Medicine, 90*(1), 11-16.

Winslow, R. M., Chapman, K. W., Gibson, C. C., Samaja, M., Monge, C. C., Goldwasser, E., . . . Santolaya, R. (1989). Different hematologic responses to hypoxia in Sherpas and Quechua Indians. *Journal of applied physiology (Bethesda, Md. : 1985), 66*, 1561-1569.

Wong, W. W., Butte, N. F., Ellis, K. J., Hergenroeder, a. C., Hill, R. B., Stuff, J. E., & Smith, E. O. (1999). Pubertal African-American girls expend less energy at rest and during physical activity than Caucasian girls. *The Journal of clinical endocrinology and metabolism, 84*, 906-911.

Woods, R. (2009). *Death before Birth*: Oxford University Press.

Wrangham, R. W., & Conklin-Brittain, N. (2003). Cooking as a biological trait. *Comparative Biochemistry and Physiology Part A, 136*, 35-46.

Wright, S. (1950a). Genetical Structure of Populations. *Nature, 166*, 247-249.

Wright, S. (1950b). Genetical Structure of Populations. *Nature, 166*(4215), 247-249.

Wu, D. D., Li, G. M., Jin, W., Li, Y., & Zhang, Y. P. (2012). Positive selection on the osteoarthritis-risk and decreased-height associated variants at the gdf5 gene in east asians. *PLoS ONE, 7*, 1-9. doi: 10.1371/journal.pone.0042553

Xu, M., Bi, Y., Xu, Y., Yu, B., Huang, Y., Gu, L., . . . Ning, G. (2010). Combined effects of 19 common variations on type 2 diabetes in Chinese: results from two community-based studies. *PloS one, 5*, 1-10. doi: 10.1371/journal.pone.0014022

Yamauchi, T., Hara, K., Maeda, S., Yasuda, K., Takahashi, A., Horikoshi, M., . . . Kadowaki, T. (2010). A genome-wide association study in the Japanese population identifies susceptibility loci for type 2 diabetes at UBE2E2 and C2CD4A-C2CD4B. *Nature genetics, 42*, 864-868. doi: 10.1038/ng.660

Yang, J., Chi, Y., Burkhardt, B. R., Guan, Y., & Wolf, B. a. (2010). Leucine metabolism in regulation of insulin secretion from pancreatic beta cells. *Nutr. Rev., 68*, 270-279. doi: 10.1111/j.1753-4887.2010.00282.x

Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Cuo, Z. X. P., Pool, J. E., . . . Wang, J. (2010). Sequencing of 50 Human Exomes Reveals Adaptation to High Altitude. *Science, 329*, 75-78. doi: 10.1126/science.1190371

Yong, H. E. J., Murthi, P., Borg, a., Kalionis, B., Moses, E. K., Brennecke, S. P., & Keogh, R. J. (2014). Increased decidual mRNA expression levels of candidate maternal pre-

eclampsia susceptibility genes are associated with clinical severity. *Placenta, 35*, 117-124. doi: 10.1016/j.placenta.2013.11.008

Yuan, X., Waterworth, D., Perry, J. R. B., Lim, N., Song, K., Chambers, J. C., . . . Mooser, V. (2008). Population-based genome-wide association studies reveal six loci influencing plasma levels of liver enzymes. *American journal of human genetics, 83*, 520-528. doi: 10.1016/j.ajhg.2008.09.012

Zhang, C., Liu, S., Solomon, C. G., & Hu, F. B. (2006). Dietary fiber intake, dietary glycemic load, and the risk for gestational diabetes mellitus. *Diabetes care, 29*, 2223-2230. doi: 10.2337/dc06-0266

Zhang, C., & Ning, Y. (2011). Effect of dietary and lifestyle factors on the risk of gestational diabetes: review of epidemiologic evidence. *The American journal of clinical nutrition, 94*, 1975S-1979S. doi: 10.3945/ajcn.110.001032

Zhang, Y., Guo, K., LeBlanc, R. E., Loh, D., Schwartz, G. J., & Yu, Y. H. (2007). Increasing dietary leucine intake reduces diet-induced obesity and improves glucose and cholesterol metabolism in mice via multimechanisms. *Diabetes, 56*, 1647-1654. doi: 10.2337/db07-0123

Zhao, L., Bracken, M. B., & DeWan, A. T. (2013). Genome-Wide Association Study of Pre-Eclampsia Detects Novel Maternal Single Nucleotide Polymorphisms and Copy-Number Variants in Subsets of the Hyperglycemia and Adverse Pregnancy Outcome (HAPO) Study Cohort. *Annals of Human Genetics, 77*, 277-287. doi: 10.1111/ahg.12021