A Natural History of Celiac Disease: Insights Into the Evolution of Complex Phenotypes.

By

Aaron J. Sams

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy
(Anthropology)

at the
UNIVERSITY OF WISCONSIN-MADISON
2012

Date of final oral examination: 12/12/2012

The dissertation is approved by the following members of the Final Oral Committee:

John Hawks, Associate Professor, Anthropology
Henry Bunn, Professor, Anthropology
Bret Payseur, Associate Professor, Medical Genetics
Travis Pickering, Professor, Anthropology
Karen Strier, Professor, Anthropology

TABLE OF CONTENTS

ABSTRACT

Celiac disease (CD) is a highly heritable small intestinal inflammatory condition induced by wheat gluten and related proteins from rye and barley. Left untreated, the clinical presentation of CD can include failure to thrive, malnutrition, and distension in juveniles. Therefore, CD potentially had a negative effect on fitness in past populations utilizing wheat, barley, and rye agriculture.

CD is common (>1%) in several populations with long histories of wheat agriculture. Therefore, it represents an evolutionary paradox. Previous analyses of CD risk variants have uncovered evidence for positive selection on some of these loci. These studies suggest the possibility that risk for common autoimmune conditions such as CD may be the result of positive selection on immune related loci in the genome. Under this evolutionary scenario, disease phenotypes may be a trade-off from positive selection on immunity. If this hypothesis is generally true, we can expect to find a signal of natural selection when we survey across the network of loci known to influence CD risk.

This project examines the total autosomal network of non-HLA genetic loci known to be associated with CD risk in Europe. We reject the null hypothesis of neutrality on this network of CD risk loci and attempt to localize evidence of selection in time and space by adding information from Tyrolean Iceman genome. We cannot reject the hypothesis of recent selection in Europe for a portion of CD risk loci. Additionally, our analysis revealed differentiation at CD risk loci early in the divergence of European and East Asian populations.

While the evolutionary history of these loci involves deterministic forces like natural selection, the selection is not uniform with respect to time or geography and is therefore not the result of adaptation to a single ancestral environment. Therefore, the connection of genetic loci to CD does not inform us about the selective pressures that shaped their current distribution. This project also demonstrates that future projects utilizing ancient DNA will be able to localize the past selective events that shaped the genetic variation underlying complex traits today.

ACKNOWLEDGEMENTS

I wish to express my sincere gratitude to Dr. John Hawks for his mentorship and encouragement during the course of my graduate studies. The lessons I have learned from working with John are too numerous to mention but he has been an exceptional role model in both teaching and research. I look forward to our ongoing friendship and collaboration. Special thanks are also due to Drs. Karen Strier, Travis Pickering, Henry Bunn, and Bret Payseur for their advice, encouragement, and willingness to serve on my guidance committee.

Furthermore, my graduate studies at the University of Wisconsin would not have been possible without the eight semesters of teaching assistantships and two lectureships provided by the Department of Anthropology. Also, an early stage of my dissertation work was funded by a generous grant from the Robert Wood Johnson Foundation.

I would also like to convey my appreciation to all of my friends and colleagues in Madison, both past and present. To those that have shared in both my happiest and most stressful moments and that have provided steadfast friendship. Special thanks are due to Zach Throckmorton and Marc Kissel who have been present and supportive during all of my academic challenges and successes. I am also grateful to Alia Gurtov, Peggy Boone, and Sarah Traynor who have provided friendship, laughs, and many inspiring conversations. Life in graduate school would have been much more challenging without such incredible friends.

The greatest thanks are due to my family for their sacrifices and unwavering support of my choices and aspirations. To my dad, Carl Sams, for encouraging me to pursue the path that brought me joy and for being the best role model, in both academia

and as a man and father, that I could hope for. To my mom, Debra Sams, for teaching me to write, for bearing the burden of my worries in life while remaining encouraging throughout, and for sharing in the joys of my accomplishments. To my sisters, Karen Lawson and Amanda Brock, and their families for their love, support, and encouragement and for putting up with me being so far away from home.

Chapter 1

GENERAL INTRODUCTION

The agricultural transition began around ten thousand years ago, changing human diet, health, and demography (Lambert, 2009). These changes had many biological effects, documented by evidence from human skeletal material (Larsen, 2006), genetic studies of human pathogens (Armelagos and Harper, 2005a) and human population genetics (Hancock et al., 2010). New agricultural diets and new pathogen pressures were among the strongest evolutionary pressures in Holocene human populations. The genes that affect immune responses and digestive and metabolic processes are not independent of each other, and in many cases are the same (Soranzo et al., 2009; Zhernakova et al., 2010). Genetic connections between diet and immunity suggest the hypothesis that human adaptations to pathogens and adaptations to diet may have coevolved in Holocene humans. Coevolution may in some cases reflect fitness tradeoffs in agricultural populations, as genetic changes with adaptive effects against pathogens may have had deleterious side effects in certain dietary environments, or vice versa.

I approached the hypothesis of diet-immune coevolution in agricultural populations by pursuing a coordinated approach that brings an evolutionary perspective to the framework of systems genetics (Nadeau and Dudley, 2011). This approach recognizes that (a) phenotypes related to immunity and diet are multigenic, and (b) there

are potentially many different configurations of genes that may be adaptive in a particular environment. The systems genetics approach uses gene-phenotype correlations and gene expression data to illuminate the interaction networks among genes, recognizing that evolution may have affected several parts of such networks.

To address the relation of diet and immunity as evolving systems, I focused on the genetic underpinnings of celiac disease (CD), an autoimmune disorder that affects dietary responses in living (industrial) populations that eat wheat gluten and related proteins from barley and rye. CD is a case study of the phenotypic intersection of diet and immunity, providing information about trade-offs that may emerge in contexts removed from those relevant in early agriculturalists or in Pleistocene hunter-gatherers.

Anthropology adds two powerful sources of evidence that can combine with the systems genetics approach. Each early agricultural population provides a natural experiment in which the environmental contexts related to both immunity and diet changed across a span of a few thousand years. To the extent that these changes induced selection, the living people who descended from these early agricultural populations may have had similar evolutionary dynamics. Comparative evidence from recent hunter-gatherers provides control cases in which populations were subject to different dietary and immune contexts until recent historic times. Additionally, ancient genomes of Neandertals, Denisovans, and Holocene aged individuals (Green et al., 2010; Rasmussen et al., 2010; Reich et al., 2010; Keller et al., 2012; Sánchez-Quinto et al., 2012; Skoglund et al., 2012) directly represent the genetics of pre-agricultural populations.

In this project I tested whether patterns of genetic variation at loci underlying CD deviate from genome-wide expectations in a hierarchical series of population

comparisons. If the evolution of CD is dominated by diet and immune evolution in agricultural contexts, the network of loci underlying CD should demonstrate a surplus of evidence for recent selection.

Diet as a selection pressure

Anthropology provides geographic and temporal evidence of dietary changes resulting from the development of agriculture (Armelagos and Harper, 2005b; a). In addition to skeletal and archaeological markers, molecular archaeological techniques have become more advanced in detecting the types of foods eaten by ancient populations (Evershed et al., 2008; Namdar et al., 2009). Diet change provides a clear avenue for testing biocultural models because of the diversity of diets adopted by different human populations and the importance of diet for fitness outcomes, particularly in energetically stressed populations. Lactase and salivary amylase are two examples of new genetic adaptations to new diets, in which some agricultural or pastoral populations have undergone large changes in the frequency of a single allele (Bersaglieri et al., 2004; Tishkoff et al., 2006; Perry et al., 2007). Both examples are notable for their parallelism: In the case of lactase, at least five independent mutations of similar effect were selected in different populations with dairying (Tishkoff et al., 2006), but more independent cases will likely be discovered (Peng et al., 2012). These are clear successes of a biocultural approach to studying recent genetic evolution related to diet.

But the broader questions about health and diet transitions are much more challenging. The evolution of health-related phenotypes must often have involved multigenic adaptations, with relatively slight changes to the frequency of many alleles.

This pattern of genetic change is very difficult to demonstrate statistically (Pritchard and Di Rienzo, 2010; Peter et al., 2012). Anthropological evidence can, however, greatly increase the statistical power of analyses by presenting evolutionarily informative comparisons and highlighting key phenotypic changes.

Bioarchaeological evidence shows that the Holocene dietary transition affected human health (Larsen, 2006), making both caloric and nutritional factors into potential selective pressures driving human evolution. Across many human cultures during the last ten thousand years, the transition to agricultural production appears to have greatly reduced human health. This reduction in human health due to nutritional deficits, dietary change, and more dense living conditions is documented by increased rates of porotic hyperostosis, cribra orbitalia, dental caries, linear enamel hypoplasias, tuberculosis, and trepanematoses in populations after the onset of agricultural production (Armelagos and Harper, 2005a; b; Larsen, 2006). Additionally, genetic and demographic evidence support an increase in the rate of adaptive evolution concurrent with large-scale demographic change over the last 10,000 years (Hawks et al., 2007). Prehistoric variation in human dietary behavior can be used to better understand the pattern and context of underlying genetic variation.

The complexity of hypotheses involving diet and health can be illustrated by the "thrifty genotype" hypothesis, which has a prominent role in the history of comparative population health. Neel (1962) proposed that foragers are more likely to experience regular food shortages than agriculturalists, so that genes that adapt them to dietary shortfalls would be adaptive in hunter-gatherer contexts, but maladaptive upon adoption of agricultural foods. Benyshek and Watson (2006) dispute the thrifty genotype

hypothesis on the basis of evidence about from past and modern ethnographic samples. Type II diabetes mellitus (T2DM) has import far beyond diet, and dozens of genetic variants are associated with slight increases in risk of T2DM in different populations. Without any single strong genetic association, this simple hypothesis becomes difficult to test.

Anthropological data are key to the question. Past populations that transitioned from foraging to agricultural production experienced a reduction in diet variation, including reduced meat consumption and increased reliance on a few plant resources, particularly cereals (Richards et al., 2003; Larsen, 2006). In contrast, the typical forager diet is characterized by more food variety including wild game, fruits, roots, legumes, nuts, and other non-cereals (Eaton et al., 1997). The hunter-gatherer diet was an evolutionary innovation earlier in the hominin lineage. Gorilla and orangutan diets consist mostly of lower quality items such as mature foliage, barks, and unripe fruit; and chimpanzees consume more fruit and less animal protein compared to humans (Milton, 1999). Dietary Westernization, with increased consumption of agricultural products, such as a high proportion of starchy carbohydrates and increased fat from domesticated animals, is thought to be the environmental cause of many dietary diseases in humans. Given this shift, we may expect not merely a "thrifty genotype" model of genetic response to diet, but a richer, more complex relation of culture, diet, and selection on genes.

The combination of evidence of increased adaptive evolution at the genetic level and of large-scale changes in human biology and culture during the Holocene suggests

that changes in human diet, culture, and demography led to large-scale genetic changes in human populations.

Immune genes

Almost all antigen-specific immune responses involve T lymphocytes. Encounters between T cells and antigens involve interactions between human leukocyte antigen (HLA) molecules and T cell receptors. HLA molecules are involved in binding antigenic peptides, which are recognized by T cell receptors. Because of this relationship with T cell receptors, HLA molecules are essential components of the immune system (Lawlor et al., 1990). The HLA system is a genic region spanning ~4Mb across the short arm of Chromosome 6 and includes over 160 protein-coding genes (Traherne, 2008). The HLA system is divided into three subregions, beginning with the telomeric class I, class III, and centromeric class II region (Fernando et al., 2008). Genes in the HLA system are known to influence chronic inflammatory and autoimmune conditions, such as type I diabetes, multiple sclerosis, Crohn's disease, and celiac disease, as well as confer susceptibility to many infectious diseases such as malaria, tuberculosis, hepatitis C and human immunodeficiency virus (HIV) (Burton et al., 2007; Blackwell et al., 2009).

A severe problem with the association of diseases with specific causal HLA variants is the extensive linkage disequilibrium (LD) (non-random associations in the inheritance of multiple loci) across the HLA region. Although the fine-scale structure of LD in the HLA system appears to be similar to the rest of the genome (extending approximately 22 kilobases (kb), there is a higher amount of LD between small-scale

segments of LD often spanning as much as ~900 kb (Walsh et al., 2003; Miretti et al., 2005; Fernando et al., 2008).

In addition to its high gene density and high LD, HLA is also one of the most variable regions of the genome. A recent analysis of eight complete HLA haplotypes (Horton et al., 2008) identified 44,544 variations including single-nucleotide substitutions and insertions and deletions (indels), while nearly 20,000 more variations have been submitted by other analyses of HLA. Several hypotheses have been proposed to explain the high degree of variability within HLA, including pathogen-driven balancing selection, HLA-dependent mate selection, and preferential abortion (Traherne, 2008). Pathogen-driven balancing selection is supported by a recent study which demonstrated increased HLA diversity relative to average genomic diversity on populations living in areas with high pathogen diversity (Prugnolle et al., 2005).

While the incredible variation within the HLA system may suggest that it is subject to balancing selection, patterns of long-range LD indicate that some HLA haplotypes have been subject to recent positive selection. Several recent studies utilizing tests of selective neutrality such as long-range haplotype (LRH) and extended haplotype homozygosity (EHH) tests support the hypothesis that the HLA system has been subject to recent positive selection (Altshuler et al., 2005; De Bakker et al., 2006; Voight et al., 2006; Frazer et al., 2007; Sabeti et al., 2007; Traherne, 2008). Two of the most common HLA haplotypes among Northern Europeans (COX and PGF) comprise about 10% frequency among Caucasian Europeans, far greater frequency than expected from the frequencies of individual loci. Traherne (2008) has suggested that the high sequence similarity of long-range haplotypes such as COX and PGF suggests that their high

frequency is due to significant expansions in recent times. In support of this, Smith and colleagues (2006) have estimated the time of origin of the COX haplotype at around 23,500 years ago. A recent common origin of long-range haplotypes may suggest that there does not need to be a functional significance to the linkage of HLA types at different loci, but rather, recent positive selection at a single locus could be sufficient to produce the expansion of a haplotype.

A few hypotheses for the selective factors that drove some common long-range HLA haplotypes have been proposed. Many of these selective factors are related to challenges that humans spreading from Africa into Eurasia between fifty and twenty thousand years ago may have faced, such as new environments, changes in diet and nutrition, and novel pathogens. For example, Moalem and colleagues (2005) posit that genetic variants contributing to type I diabetes mellitus, which is strongly associated with genetic variants in HLA (Burton et al., 2007) may have been actively selected in populations inhabiting colder climates of Northern Europe during the Younger Dryas ~13-11 thousand years ago, as a means of lowering the freezing point of body fluids. Some hypotheses are more straightforward. For example, some HLA variants are associated with protection from pathogens. However, other common HLA variants have been associated with susceptibility to pathogen infection (Blackwell et al., 2009), and the prevalence of such variants is not completely understood.

Although it is clear that the HLA region has undergone strong recent selection (Albrechtsen et al., 2010), the exact selection pressures which drove the evolution of HLA variants related to autoimmune diseases, pathogens, and haemotological factors remain elusive. The emerging field of systems genetics, which focuses on networks of

interactions between systems of genes and phenotypic traits (Nadeau and Dudley, 2011), combined with an anthropological evolutionary approach provides a framework from which to address such questions. Novel methods which attempt to correlate evolutionary history and global genetic variation with environmental conditions, demographic history, and subsistence patterns may have more success in identifying the selective factors most highly associated with specific systems of genes and phenotypic traits (Hancock et al., 2010).

Pleiotropy of diet and immune associated loci

Some evolutionary case studies of dietary change as a selection pressure are straightforward. The recent adaptive evolution of the lactase and salivary amylase genes (described above) involve single gene changes or duplications. These studies involved the identification of genetic changes in single genes with known function. On the other hand, many phenotypic traits associated with diet, metabolism, and immune functions are, in their own right, complex and often controlled by many polymorphisms at numerous genetic loci. Complicating our understanding of these systems even further are various associations, which link genes typically associated with one genetic pathway and its associated phenotypic trait, with phenotypes in other systems.

Celiac disease (CD) is a common chronic inflammatory condition of the small intestine that is induced by ingestion of dietary wheat, rye, and barley. This heritable dietary condition involves a well-understood inflammatory response that is activated when cereal peptides are presented CD4+ T cells. While polymorphisms in genes in the HLA system (HLA-DQA1 and DQB1) have long been known to contribute to disease

risk, they do not completely explain disease development (Green and Cellier, 2007; Kagnoff, 2007; Wolters and Wijmenga, 2008). In a recent genome wide association study (GWAS) of CD by Dubois and colleagues (2010), which included five European CD case and control sample collections, the researchers confirmed 13 previously reported CD risk alleles and identified 26 new regions. Interestingly, the majority of loci in the newly identified regions are known to also have other immune functions.

As another recent example of pleiotropy of genetic associations for complex disorders, Soranzo and colleagues (2009) present evidence that risk alleles for type-1 diabetes, coronary artery disease, and celiac disease are carried on a derived haplotype of the 12q24 region, a region also linked with immune function. Various tests of neutrality (linkage disequilibrium, Fst, iHS, and XP-EHH scores) provide evidence that this haplotype arose from a selective sweep in European and nearby populations ~3,400 years ago, possibly as a result of an adaptive immune response to living in high-density (agricultural) populations.

In addition to the close relationships demonstrated between diet and immune genetic pathways, there is also reason to believe that changes in diet in past populations may have provided novel selection pressures to not only diet and metabolic pathways, but also immune pathways. For instance, several essential micronutrients are known to have a direct role in immune function. Iron and zinc deficiencies have been associated with immune dysfunction (Kaput, 2003) and are precisely the same micronutrients that are thought to have been lacking in grain-based, low-meat agricultural diets (Larsen, 2006). This example further illustrates the complex interconnectivity between diet, pathogen, and immune related phenotypes and genotypes.

As hypothesized by Soranzo and colleagues (2009) we may expect there to have been evolutionary trade-offs, particularly at pleiotropic loci involved in multiple phenotypic and genetic pathways. While the genetic and phenotypic intermingling of diet, metabolic, and immune related traits complicates our ability to predict health outcomes from genetic material alone, it is precisely these relationships we can identify by taking advantage of information from comparisons of human populations at different scales, thereby utilizing the demographic history of human populations.

Celiac disease as a model to test for recent diet/immune trade-offs

Celiac disease is a complex trait related to both diet and the immune system and provides an excellent case study to explore the evolution of a genetically well-understood complex trait for which the primary environmental trigger, gluten, is known. Knowledge of the worldwide prevalence of CD is incomplete but growing (Gandolfi et al., 2000; Catassi et al., 2001; Mäki et al., 2003; Greco et al., 2012; Riddle et al., 2012; Rubio-Tapia et al., 2012). Further, there is much variation across populations in the consumption of CD causing cereals (Abadie et al., 2011). The general consensus, however, is that CD is more common in populations with European and Near Eastern ancestry. This is somewhat of an evolutionary enigma because these are precisely the populations with the longest history of wheat, rye, and barley agriculture (Tresset and Vigne, 2011; Zeder, 2011), which might lead us to expect adaptations to have arisen to prevent CD. A long history of CD risk in these populations requires some evolutionary explanation. Founder effects in ancient populations might have increased the frequency of deleterious CD risk alleles in at-risk populations or these alleles may have been

influenced by positive selection on their other pleiotropic effects. Additionally, the genetic architecture of CD may be such that most loci contributing to CD risk have only minor effects on risk (fitness), meaning that selection against CD would have miniscule to nonexistent effects on any single risk locus (Stranger et al., 2011).

Consideration of the evolution of such a complex trait incites a larger philosophical question. What is an evolutionary genomic network? Can selection across multiple genetic loci be studied in a systematic way? Furthermore, should we expect sets of interacting loci to exhibit common evolutionary histories, particularly with regards to selection? There is a established academic history (explored in Chapter III) of understanding the role of selective forces in human evolution by analyzing traits that pose the same type of evolutionary paradox as celiac disease (CD), that is, reduced fitness in a common present environments.

Content of thesis

An important question is, does genetic variation at the non-HLA risk loci reveal a pattern of recent selection in high CD risk areas? In other words, do HLA and non-HLA risk loci represent an evolutionarily integrated network, as we might expect if CD risk represents an evolutionary response to health changes occurring during the Neolithic transition? In order to investigate this question I investigated patterns of genetic variation in CD background (non-HLA) risk loci and explored the role of ancient DNA in understanding the timing of recent selection. Chapter II consists of a hierarchical comparison of CD genetic variation using genome-wide controls and utilizing six populations from the 1000 Genomes Project. This chapter examines the evidence for

recent selection across the CD background risk genetic network, with special emphasis on Europe, and uses data from the Tyrolean Iceman, Ötzi, to explore previously reported evidence about the timing of selection on several CD risk loci. Chapter III is a comprehensive review of the evidence about selection on CD and frames this literature within the larger discussion about the evolution of complex traits and genetic networks. Finally, Chapter IV uses genetic variation from Ötzi more explicitly in a genome-wide analysis of estimated allele age. The purpose of this study was to explore the accuracy of predictions about the timing of selection made from present human DNA and the role of demography in biasing these estimations. Chapter V is a general discussion of the challenges faced when investigating the evolution of complex genetic networks. This discussion integrates the previous chapters and explores future directions for further investigation of this topic, with special emphasis on the role of ancient DNA.

REFERENCES

Abadie V, Sollid LM, Barreiro LB, and Jabri B. 2011. Integration of Genetic and Immunological Insights into a Model of Celiac Disease Pathogenesis. Annual Review of Immunology 29:493–525.

Albrechtsen A, Moltke I, and Nielsen R. 2010. Natural Selection and the Distribution of Identity-by-Descent in the Human Genome. Genetics 186:295–308.

Altshuler D, Brooks LD, Chakravarti A, Collins FS, Daly MJ, and Donnelly P. 2005. A haplotype map of the human genome. Nature 437:1299–1320.

Armelagos GJ, and Harper KN. 2005a. Genomics at the origins of agriculture, part two. Evol Anthropol 14:109–121.

Armelagos GJ, and Harper KN. 2005b. Genomics at the origins of agriculture, part one. Evol Anthropol 14:68–77.

Benyshek DC, and Watson JT. 2006. Exploring the thrifty genotype's food-shortage assumptions: A cross-cultural comparison of ethnographic accounts of food security among foraging and agricultural societies. Am J Phys Anthropol 131:120–126.

Bersaglieri T, Sabeti P, Patterson N, Vanderploeg T, Schaffner S, Drake J, Rhodes M, Reich D, and Hirschhorn J. 2004. Genetic signatures of strong recent positive selection at the lactase gene. The American Journal of Human Genetics 74:1111–1120.

Blackwell JM, Jamieson SE, and Burgner D. 2009. HLA and Infectious Diseases. Clinical Microbiology Reviews 22:370–385.

Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, Duncanson A, Kwiatkowski DP, Mccarthy MI, Ouwehand WH, et al. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661–678.

Catassi C, Doloretta Macis M, Ratsch IM, De Virgiliis S, and Cucca F. 2001. The distribution of DQ genes in the Saharawi population provides only a partial explanation for the high celiac disease prevalence. Tissue Antigens 58:402–406.

De Bakker PIW, Mcvean G, Sabeti PC, Miretti MM, Green T, Marchini J, Ke X, Monsuur AJ, Whittaker P, et al. 2006. A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. Nature Genetics 38:1166–1172.

Dubois PCA, Trynka G, Franke L, Hunt KA, Romanos J, Curtotti A, Zhernakova A, Heap GAR, Ádány R, et al. 2010. Multiple common variants for celiac disease influencing immune gene expression. Nature Genetics 42:295–302.

Eaton S, Eaton S, and Konner M. 1997. Paleolithic nutrition revisited: a twelve-year retrospective on its nature and implications. European Journal of Clinical Nutrition 51:207–216.

Evershed R, Payne S, Sherratt A, Copley M, Coolidge J, Urem-Kotsu D, Kotsakis K, and Özdo gbreve M. 2008. Earliest date for milk use in the Near East and southeastern Europe linked to cattle herding. Nature 455:528–531.

Fernando MMA, Stevens CR, Walsh EC, De Jager PL, Goyette P, Plenge RM, Vyse TJ, and Rioux JD. 2008. Defining the Role of the MHC in Autoimmunity: A Review and Pooled Analysis. PLoS Genetics 4:e1000024.

Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, et al. 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851–861.

Gandolfi L, Pratesi R, Cordoba J, Tauil P, Gasparin M, and Catassi C. 2000. Prevalence of celiac disease among blood donors in Brazil. The American Journal of Gastroenterology 95:689–692.

Greco D, Pisciotta M, Gambina F, and Maggio F. 2012. Celiac disease in subjects with type 1 diabetes mellitus: a prevalence study in western Sicily (Italy). Endocrine 1–4.

Green P, and Cellier C. 2007. Celiac Disease. New England Journal of Medicine 357:1731–1743.

Green R, Krause J, and Paabo S. 2010. A draft sequence of the Neandertal genome. Science 328:710-722.

Hancock AM, Witonsky DB, Ehler E, Alkorta-Aranburu G, Beall C, Gebremedhin A, Sukernik R, Utermann G, Pritchard J, et al. 2010. Colloquium Paper: Human adaptations to diet, subsistence, and ecoregion are due to subtle shifts in allele frequency. Proceedings of the National Academy of Sciences 107:8924–8930.

Hawks J, Wang ET, Cochran GM, Harpending HC, and Moyzis RK. 2007. Recent acceleration of human adaptive evolution. Proceedings of the National Academy of Sciences 104:20753–20758.

Horton R, Gibson R, Coggill P, Miretti M, Allcock RJ, Almeida J, Forbes S, Gilbert JGR, Halls K, et al. 2008. Variation analysis and gene annotation of eight MHC haplotypes: The MHC Haplotype Project. Immunogenetics 60:1–18.

Kagnoff MF. 2007. Celiac disease: pathogenesis of a model immunogenetic disease. Journal of Clinical Investigation 117:41–49.

Kaput J. 2003. Nutritional genomics: the next frontier in the postgenomic era. Physiological Genomics 16:166–177.

Keller A, Graefen A, Ball M, Matzas M, Boisguerin V, Maixner F, Leidinger P, Backes C, Khairat R, et al. 2012. New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing. Nat Comms 3:698.

Lambert PM. 2009. Health versus Fitness. Current Anthropology 50:603–608.

Larsen CS. 2006. The agricultural revolution as environmental catastrophe: Implications for health and lifestyle in the Holocene. Quaternary International 150:12–20.

Lawlor D, Zemmour J, Ennis P, and Parham P. 1990. Evolution of class-I MHC genes and proteins: from natural selection to thymic selection. Annual Review of Immunology 8:23–63.

Mäki M, Mustalahti K, and Kokkonen J. 2003. Prevalence of Celiac Disease among Children in Finland. The New England Journal of Medicine 348:2517–2524.

Milton K. 1999. A hypothesis to explain the role of meat-eating in human evolution. Evol Anthropol 8:11–21.

Miretti M, Walsh E, Ke X, and et al. 2005. A high-resolution linkage-disequilibrium map of the human major histocompatibility complex and first generation of tag single-nucleotide polymorphisms. The American Journal of Human Genetics 76:634-646.

Moalem S, Storey KB, Percy ME, Peros MC, and Perl DP. 2005. The sweet thing about type 1 diabetes: a cryoprotective evolutionary adaptation. Medical Hypotheses 65:8–16.

Nadeau JH, and Dudley AM. 2011. Systems Genetics. Science 331:1015–1016.

Namdar D, Stacey RJ, and Simpson SJ. 2009. First results on thermally induced porosity in chlorite cooking vessels from Merv (Turkmenistan) and implications for the formation and preservation of archaeological lipid residues. Journal of Archaeological Science 36:2507–2516.

Neel J. 1962. Diabetes Mellitus: A "Thrifty" Genotype Rendered Detrimental by "Progress"? American Journal of Human Genetics 14:353.

Peng M-S, He J-D, Zhu C-L, Wu S-F, Jin J-Q, and Zhang Y-P. 2012. Lactase persistence may have an independent origin in Tibetan populations from Tibet, China. Journal of Human Genetics 57:394–397.

Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, Werner J, Villanea FA, Mountain JL, et al. 2007. Diet and the evolution of human amylase gene copy number variation. Nature Genetics 39:1256–1260.

Peter BM, Huerta-Sanchez E, and Nielsen R. 2012. Distinguishing between Selective Sweeps from Standing Variation and from a De Novo Mutation. PLoS Genetics 8:e1003011.

Pritchard J, and Di Rienzo A. 2010. Adaptation–not by sweeps alone. Nat Rev Genet 11:665–667.

Prugnolle F, Manica A, Charpentier M, Guégan JF, Guernier V, and Balloux F. 2005. Pathogen-driven selection and worldwide HLA class I diversity. Current Biology 15:1022–1027.

Rasmussen M, Li Y, Lindgreen S, Pedersen JS, Albrechtsen A, Moltke I, Metspalu M, Metspalu E, Kivisild T, et al. 2010. Ancient human genome sequence of an extinct Palaeo-Eskimo. Nature 463:757–762.

Reich D, Green R, Kircher M, Krause J, Patterson N, Durand EY, Viola B, Briggs AW, Stenzel U, et al. 2010. Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature 468:1053–1060.

Richards MP, Schulting RJ, and Hedges REM. 2003. Archaeology: Sharp shift in diet at onset of Neolithic. Nature 425:366–366.

Riddle MS, Murray JA, and Porter CK. 2012. The Incidence and Risk of Celiac Disease in a Healthy US Adult Population. The American Journal of Gastroenterology 107:1248-1255.

Rubio-Tapia A, Ludvigsson JF, Brantner TL, Murray JA, and Everhart JE. 2012. The prevalence of celiac disease in the United States. The American Journal of Gastroenterology 107:1538-1544.

Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, Mccarroll SA, et al. 2007. Genome-wide detection and characterization of positive selection in human populations. Nature 449:913–918.

Sánchez-Quinto F, Schroeder H, Ramirez O, Avila-Arcos MC, Pybus M, Olalde I, Velazquez A, Marcos MEP, Encinas JMV, et al. 2012. Genomic Affinities of Two 7,000-Year-Old Iberian Hunter-Gatherers. Current Biology 22:R631–R633.

Skoglund P, Malmström H, Raghavan M, Storå J, Hall P, Willerslev E, Gilbert MTP, Götherström A, and Jakobsson M. 2012. Origins and Genetic Legacy of Neolithic Farmers and Hunter-Gatherers in Europe. Science 336:466–469.

Smith W, Vu Q, Li S, Hansen J, and Zhao L. 2006. Toward understanding MHC disease associations: partial resequencing of 46 distinct HLA haplotypes. Genomics. 87:561-571.

Soranzo N, Spector TD, Mangino M, Kühnel B, Rendon A, Teumer A, Willenborg C, Wright B, Chen L, et al. 2009. A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. Nature Genetics 41:1182–1190.

Stranger BE, Stahl EA, and Raj T. 2011. Progress and Promise of Genome-Wide Association Studies for Human Complex Trait Genetics. Genetics 187:367–383.

Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, Powell K, Mortensen HM, Hirbo JB, et al. 2006. Convergent adaptation of human lactase persistence in Africa and Europe. Nature Genetics 39:31–40.

Traherne JA. 2008. Human MHC architecture and evolution: implications for disease association studies. Int J Immunogenet 35:179–192.

Tresset A, and Vigne JD. 2011. Last hunter-gatherers and first farmers of Europe. Comptes rendus biologies 334:182–189.

Voight BF, Kudaravalli S, Wen X, and Pritchard JK. 2006. A Map of Recent Positive Selection in the Human Genome. Plos Biol 4:e72.

Walsh EC, Mather KA, Schaffner SF, Farwell L, Daly MJ, Patterson N, Cullen M, Carrington M, Bugawan TL, et al. 2003. An integrated haplotype map of the human major histocompatibility complex. American Journal of Human Genetics 73:580.

Wolters VM, and Wijmenga C. 2008. Genetic Background of Celiac Disease and Its Clinical Implications. The American Journal of Gastroenterology 103:190–195.

Zeder MA. 2011. The Origins of Agriculture in the Near East. Current Anthropology 52(S4):S221-S235.

Zhernakova A, Elbers CC, Ferwerda B, Romanos J, Trynka G, Dubois PC, de Kovel CGF, Franke L, Oosting M, et al. 2010. Evolutionary and Functional Analysis of Celiac Risk Loci Reveals SH2B3 as a Protective Factor against Bacterial Infection. The American Journal of Human Genetics 86:970–977.

Chapter II

Patterns of Population Differentiation and Natural Selection on the Celiac Disease Background Risk Network.

Aaron Sams and John Hawks

ABSTRACT

Celiac disease is a common, strongly heritable small intestinal inflammatory condition induced by wheat gluten and related proteins from rye and barley. Left untreated, the clinical presentation of CD can include failure to thrive, malnutrition, and distension in juveniles. These symptoms are referred to as the tip of the "celiac disease iceberg," as the disease can additionally lead to severe vitamin deficiencies, anemia, and osteoporosis. Therefore, CD potentially had a negative effect on fitness in past populations utilizing wheat, barley, and rye agriculture.

Previous analyses of CD risk variants have uncovered evidence for positive selection on some of these loci. These studies also suggest the possibility that risk for common autoimmune conditions such as CD may be the result of positive selection on immune related loci in the genome to fight infection. Under this evolutionary scenario, disease phenotypes may be a trade-off from positive selection on immunity. If this hypothesis is

generally true, we can expect to find a signal of natural selection when we survey across the network of loci known to influence CD risk.

The current study examines the total autosomal network of gene loci known to be associated with CD risk in Europe. We reject the null hypothesis of neutrality on this network of CD risk loci. In addition to rejecting neutrality, we can localize evidence of selection in time and space by adding information from the genome of the Tyrolean Iceman. While we can show significant differentiation between continental regions across the CD network, the pattern of evidence is not consistent with primarily recent (Holocene) selection across this network in Europe. Instead, the loci presently associated with CD risk were collectively influenced by selection that acted much earlier in the emergence of modern humans, either as they left Africa or within more recent ancestors of present populations. Further localizing the distribution of ancient selection on this network may illuminate the ecological pressures acting on the human immune system during this critically interesting phase of our evolution.

Keywords: natural selection, celiac disease, autoimmune, Europe, ancient DNA

INTRODUCTION

Celiac disease (CD) is a common, highly heritable (Nistico, 2006), small intestinal inflammatory condition induced by wheat gluten and related proteins from rye and barley (Sollid, 2000). Specific risk alleles of the HLA-DQA1 and DQB1 genes encoding the DQ2 and DQ8 heterodimers appear to be necessary, but not solely responsible for development of CD (Wolters and Wijmenga, 2008). These specific HLA genotypes

explain approximately 40 percent of the genetic risk for CD. HLA-DQ heterodimers encoded by these risk alleles present cereal peptides to CD4+ T cells, activating an inflammatory immune response in the small intestine (Kagnoff, 2007). Recent genome-wide association studies (GWAS) and linkage analyses have identified at least forty non-HLA loci, many of which lie in genomic regions that earlier work had already associated with immune function (Van Heel et al., 2007; Hunt et al., 2008; Dubois et al., 2010; Zhernakova et al., 2010; Trynka et al., 2011; Zhernakova et al., 2011). Unlike the case of HLA, the exact causal mechanisms for these other CD risk associations are mostly unknown.

Bioarchaeological evidence shows that dietary and demographic transitions of the Holocene were detrimental to human health (Larsen, 2006). Dietary specialization led to nutritional deficits and denser living conditions created ideal conditions for the spread of new pathogens, as documented by higher rates of porotic hyperostosis, cribra orbitalia, dental caries, linear enamel hypoplasias, tuberculosis, and trepanematoses in populations after the onset of agricultural production (Armelagos and Harper, 2005; Larsen, 2006). Genetic and archaeological evidence show the importance of adaptive evolution concurrent with large-scale demographic change over the last 10,000 years (Hawks et al., 2007). These observations provoke the hypothesis that the Holocene transition to agriculture was a time of strong natural selection on genes important to diet and immunity.

CD appears to be an evolutionary paradox. It is common today (>1%) in several populations with long histories of wheat agriculture including Europe and the Near East. The effectiveness of a gluten-free diet as a treatment has been clinically recognized only

since the discovery of gluten as the CD trigger in the 1950's (Barker and Liu, 2008). In past populations CD should have had a high potential to reduce the fitness of its sufferers directly, or by interfering with nutrient absorption, or by other effects that reduce fertility (Soni and Badawy, 2010). The idea that CD would have reduced past fitness was recently bolstered by skeletal and DNA evidence from a roughly 2000 year-old female from the archaeological site of Cosa, near Tuscany, Italy. This skeleton includes classic paleopathological signs of infection and malnutrition leading to death prior to age 20 and also contains the most common HLA-DQ risk haplotype (DQ2.5) for CD (Gasbarrini et al., 2012). A long history of CD risk European and Near Eastern populations requires some evolutionary explanation. Founder effects in ancient populations might have increased the frequency of deleterious CD risk alleles, or these alleles may have been influenced by positive selection on their other pleiotropic effects. Much evidence supports the role of recent natural selection in shaping the HLA region (Albrechtsen et al., 2010). Less well understood is the influence of selection on the more than 40 non-HLA loci associated with CD risk.

One study has suggested strong recent selection on some of these GWAS CD risk loci. Zhernakova and colleagues (2010) found evidence of recent selection in or around the genes IL12A, IL18RAP, and SH2B3. The risk variant in the SH2B3 gene is functionally involved in the NOD2 recognition pathway, suggesting that it may have been positively selected to protect against bacterial infection. The authors inferred a very recent onset of selection (between 1,200 and 1,700 years ago) by extended haplotype homozygosity (EHH) (Sabeti et al., 2002).

We investigated whether the signals on these three loci are consistent with the broader genetic network influencing CD risk. A null hypothesis is that CD-associated gene loci have undergone the same dynamics of demography and selection as the rest of the genome. We tested this hypothesis by evaluating the differentiation of the chromosomal regions flanking CD-associated loci, both between and within continental populations. Owing to the spread of humans around the globe and their subsequent rapid population growth, evidence of recent selection tends to be geographically localized within continents or smaller geographic regions (Voight et al., 2006; Hawks et al., 2007). For this reason, gene networks whose components have been subject to recent selection may show significant regional differentiation when compared to loci randomly drawn from across the genome (Hancock et al., 2010). In the case of CD selection on standing variants may have been just as important as selection on new mutations and a test across many loci may find evidence missed in the examination of single loci with linkage disequilibrium (LD) based approaches. GWAS risk loci are not necessarily causal variants, so we tested flanking regions for evidence of population differentiation. By comparing a hierarchical set of populations both within continents and between continents, we investigated the time and geographic distribution of adaptive variation in this gene network.

RESULTS

To investigate broad-scale continental differentiation between continental regions we assessed the fraction of SNPs in each CD risk locus that fall in the top 1% of the genome-wide F_{ST} distribution (assessed from over 2 million randomly chosen SNPs) for that

between-continent comparison. Further, that fraction of high F_{ST} SNPs at each locus was assigned a significance (P) value based on a genome-wide random sample of 11,000 loci.

The regions around CD risk loci are much more likely to show significant intercontinental differentiation than are SNP loci randomly chosen across the genome. Averaged over all CD loci, the proportion of high F_{ST} outliers exceeded 1% in all three between-continent comparisons (Appendix). In our bootstrapping procedure, thirteen of fifty-four (~24%) of the regions flanking independent CD risk loci are significantly differentiated (have more than expected high- F_{ST} SNPs, $P < 0.05$) for at least one between-continent comparison. This may be a slight overestimate due to the fact that flanking regions overlap in a few loci.

The differentiation of the CD risk network stands out most prominently from other loci when comparing European and Asian samples (Figure 1). In this case, the simplest comparison is the overall fraction of high- F_{ST} SNPs across all regions. For Europe and Asia, 1.5 percent of SNPs across the regions flanking CD risk loci exceed the 1 percent genome-wide threshold (Appendix). In other words, fifty percent more high-F_{ST} SNPs are found in these regions than expected from the genome-wide distribution. These SNPs are linked into regions and so the appropriate significance test on the network is based on the proportion of regions with an excess of high F_{ST} SNPs. High F_{ST} SNP loci are especially clustered within nine of the chromosomal regions around CD risk-associated loci. Nine of these regions are significant in a Europe-Asia pairwise comparison genome-wide at $P < 0.05$, two of which are significant at $P < 0.01$ (Figure 1). This high proportion of high F_{ST} blocks is strongly significant ($P < 0.01$) when considering the total number of comparisons.

The Europe-Africa comparison also shows a highly elevated fraction of high F_{ST} SNPs. However, these are clustered mainly in four of the chromosomal regions (Figure 1). The number of regions showing elevated F_{ST} between Europe and Africa is therefore not significant. The four high F_{ST} regions include two loci that show significant evidence of selection by long-range linkage (Barreiro and Quintana-Murci, 2010; Zhernakova et al., 2010; Abadie et al., 2011). To the extent that these loci do show evidence of selection, our comparison across the network does not show that other loci have been significantly selected in the history of European and African population differentiation.

We consider this pattern of between-continent results a rejection of the null hypothesis (H_0) that the CD background risk network has experienced the same pattern of evolution compared to the genome as a whole. The results point in an interesting direction. Loci identified by earlier long-range haplotype tests as recently selected (such as the SH2B3 locus on chr 12) account for a high fraction of the high- F_{ST} SNPs in all comparisons. Furthermore, evidence of elevated F_{ST} across the broader CD network that has not been shown to be subject to positive selection is concentrated between Europe and East Asia. Importantly, the loci with elevated F_{ST} between Europe and East Asia are, for the most part, not the same loci that are highly differentiated between Europe and Africa. Therefore, we are picking up a signal primarily of selection on standing variation, not strong positive selection, in addition to previously demonstrated evidence of recent positive selection on two of these loci. Evolution of the CD risk network within Europe might account for some of these observations.

We turned to within-continent comparisons to address whether evolution of the CD risk-associated loci occurred uniformly over time. Previous work suggested that a

proportion of autoimmune genetic risk (including CD) may reflect positive selection on the immune system within the last 10,000 years (Soranzo et al., 2009; Barreiro and Quintana-Murci, 2010; Zhernakova et al., 2010; Abadie et al., 2011). If the selection that we identified with between-continent comparisons were very recent, we might expect additional loci to show high- F_{ST} fractions. If there were more recent (Holocene) selection in Europe or East Asia, we might expect additional loci to show up as significant outliers when comparing populations within each continental region. What we observe is the opposite. Within Africa, approximately five independent regions show a significant excess of high- F_{ST} SNPs. Within Asia, only four regions have a significant excess of high- F_{ST} SNPs. In Europe there is a strong deficit of high- F_{ST} SNPs. No within continent comparisons show a significant ($P < 0.05$) excess of regions with a significant ($P < 0.05$) ratio of high- F_{ST} SNPs.

DISCUSSION

Our within-continent comparisons show that the overall pattern of CD risk does not match the predictions of strong recent selection in Europe. Considered together, the CD risk loci had a higher than expected proportion of outlier SNPs in Africa and Asia, but not in Europe (Figure 1). Compared to the genome-wide value, the CD risk network showed the highest relative differentiation between Luhya (LWK) and Yoruba (YRI) samples. The two Chinese samples (CHB and CHS) are also more highly diverged than the genome-wide expectation. In contrast, the CEPH (CEU) and Tuscani (TSI) samples show a much smaller fraction of genomic high-F_{ST} outliers in the CD risk regions. Our intercontinental comparisons showed that selection likely differentiated East Asia and

Europe at several CD risk loci, if that selection acted mostly early in the differentiation of Europe and East Asia, we could predict that European populations today might show low differentiation across this network. That is the result we observe. However, very recent selection on standing variation within Europe that acted in the same direction across this network in these European populations might also give rise to low differentiation of these populations as observed today. Therefore, the within-continent F_{ST} comparison by itself cannot rule out recent selection within Europe.

This result may appear to contradict previous evidence of strong recent selection on at least four of these loci (Barreiro and Quintana-Murci, 2010; Zhernakova et al., 2010; Abadie et al., 2011). The contradiction might be a unique property of the evolution of these loci for which our comparisons, directed to the broader issue of selection across the entire network, may be less informative. We chose to investigate one of these loci further. The Iceman genome (Keller et al., 2012) provides an alternative test of recent selection on the CD risk network. The strongest prior evidence of selection in this network is the risk variant rs3184504 in SH2B3 reported by Zhernakova and colleagues (2010). This locus is represented in the data from the 5,300-year-old Ötzi genome. Zhernakova and colleagues estimated an age for this locus using the EHH statistic of only 1,200-1,700 years ago. This age estimate makes the clear prediction that the iceman genome should not have the risk allele. However, Ötzi is a heterozygote carrying this allele. For this reason, we propose that the evidence of selection on this locus may actually pertain to a time period well before 5,300 years ago.

We do not know the extent to which GWAS SNPs contribute to CD risk, but the Iceman genotypes can provide a rough test of positive selection across the network even

without this information. The Ötzi genotype was drawn from the European population of 5,000 years ago. If the ancient population was different from today's European population in the frequencies of SNP alleles associated with CD risk, then Ötzi will carry some genotypes that may be unlikely given their frequencies today in Europe. Ötzi is a heterozygote at nine out of forty-nine total GWAS risk sites (mean heterozygosity ≈ 0.184). The average heterozygosity in the Ötzi genome across these loci is identical to the minimum heterozygosity among Europeans in the 1000 Genomes Project sample today and closest to the mean heterozygosity observed in Africa (Figure 2). Low coverage in the Ötzi genome could lead to a bias in the ascertainment of heterozygous sites due to chance. We tested whether a coverage bias could explain the reduction in heterozygosity across the GWAS risk sites (avg. coverage = 5.6) by randomly generating sets of genotypes based on allele frequencies in modern Europeans and used these sets of genotypes to resample reads from each genotype randomly based on read number in each risk site in Ötzi. The resulting distribution of over one million randomly generated average heterozygosities demonstrates a marked reduction in ascertainment of heterozygous sites (Figure 3). Nonetheless, the average number of heterozygous sites that we measure for Ötzi remains in the bottom of the heterozygosity distribution for Europe. Because Ötzi is an outlier compared to present-day Europeans for genotypes across these CD risk loci, we cannot reject the hypothesis that strong positive selection has affected the frequencies of any large proportion of these loci during the past 5,000 years.

However, recent genetic evidence from ancient mitochondrial and autosomal DNA (Fu et al., 2012; Sánchez-Quinto et al., 2012; Skoglund et al., 2012) supports the hypothesis of large-scale demographic turnover in Europe associated with the spread of

agriculture between 10,000 and 5,000 years ago. If Ötzi is not closely related to the ancestors of present Europeans, then a difference in genetic composition at GWAS sites may (perhaps more likely) reflect demography rather than selection. More refined resolution of the Neolithic transition will help to resolve this issue. These results are equivocal but illustrate both the value of and issues associated with using ancient DNA for testing hypotheses about the timing of recent genetic changes. These issues are discussed further in Chapter IV.

The goal of this study was to address the evolutionary paradox of celiac disease risk, by assessing whether non-HLA CD risk loci have been affected by selection in human prehistory. The HLA-DQ haplotypes that contribute most substantially to individual risk of CD are already strong candidates for selection due to their functional roles in immunity, but the functional interactions are less well understood for the wider background of loci associated with CD risk. Our results demonstrate that these loci, when considered as a network, also have been influenced by selection during the evolution of human populations. That selection increased the differentiation of CD risk loci between Europeans and other continental groups.

The linkage block that includes the SH2B3 locus identified by Zhernakova and colleagues (2010) shows the strongest sign of selection in our analysis, in terms of differentiation between continents. Therefore, this locus is likely subject to increased selection compared to other CD associated risk loci. However, both its presence in these far-flung populations and the presence of the specific risk allele (rs3184504) in the 5,300-year-old Tyrolean Iceman show that it is much older than initially estimated by Zhernakova and colleagues.

The pattern at SH2B3 is not typical of the CD risk network as a whole, which shows significant but much weaker indications of past selection. The pattern of increased differentiation between continents is clear only when considering a relatively large number of CD risk loci and their flanking regions. Setting aside the large risk conferred by HLA-DQ CD risk alleles, the known genetic CD background risk seems to have been affected by a combination of evolutionary patterns. Many of the CD risk loci are consistent with the pattern of differentiation of the genome as a whole; a substantial (and significant) fraction have been subject to changes in frequency that increased differentiation between populations without creating the kind of linkage disequilibrium that triggers significant EHH or iHS tests for positive selection, while a handful of loci may have been affected by strong positive selection.

No loci are excessively differentiated within Europe and the genotypes of the 5,300-year-old Tyrolean Iceman potentially suggest a recent shift in allele frequencies at some CD risk loci. These observations lead to two alternatives. First, it is possible that there has been a great deal of recent selection across the CD network in Europe which has either been in the form of uniform selection on standing variation or positive selection in combination with recent demographic turnover in Europe. Alternatively, our between and within continent comparisons may simply represent selection earlier in the expansion of modern humans from Africa. We can improve our resolution of the recent history of CD loci in Europe with either data from more European and Near Eastern populations or with more ancient DNA. Ötzi illustrates that our hierarchical framework of modern comparisons are lacking in resolution with respect to recent selection. Additionally, we

still cannot rule out that recent selection in Asia and/or Africa primarily explains our between-continent results.

Moving forward, additional analysis of CD risk loci utilizing a geographically broader sample of population data and perhaps more importantly a larger number of ancient DNA samples will be necessary to resolve the timing of selection across the CD network in Europe. That additional information will reveal the extent to which the cultural and demographic shifts associated with the spread of agriculture from the Near East influenced the current distribution of genetic variation at non-HLA CD risk loci.

METHODS

1000 Genomes Project data

We obtained variant calls for single nucleotide polymorphisms from the June 2011 data release of the 1000 Genomes Project (http://1000genomes.org) (Durbin et al., 2010). For our comparisons, we considered only the autosomes, chromosomes 1-22, and excluded the X chromosome from consideration. We utilized the subset of individuals from the following population samples: Utah (CEU), Finn (FIN), Spanish (IBS), British (GBR), Tuscan (TSI), North and South ethnic Han Chinese (CHB, CHS), Japanese (JPT), Luhya (LWK), and Yoruba (YRI).

Celiac regions

We utilized the list in Table 2 of Trynka and colleagues (2011) of all autosomal genomic regions with a highly correlated genome-wide significant signal of celiac risk. In our F_{ST} analyses we chose the center of each independent locus and assessed F_{ST} in the region

spanning 25Kb upstream and downstream of this central position. For loci much longer than 50Kb we broke the locus into non-overlapping 50Kb intervals. The resulting regions were mapped to human genome build 19 (hg19) using the UCSC genome build liftover utility (genome.ucsc.edu). The original regions from Trynka and colleagues (2011), nearby genes, our loci mapped to hg19, and the results of our F_{ST} comparisons can be found in Appendix 1. Because we used physical distance as a measure of locus length rather than recombination map distance, we assessed average recombination rates in the set of CD loci to ensure that they do not have systematically lower recombination rates than the genome-wide expectation and that low recombination rates are not correlated with elevated F_{ST} ratios (data not shown).

Iceman genome

We examined the recently published genome of the 5,300 year old Tyrolean Iceman (Ötzi) (Keller et al., 2012) to provide a further test of the hypothesis of recent selection across the CD risk network in Europe. We obtained aligned genome reads from the Ötzi from the European Short Read Archive. At the time of our download, the Ötzi data were provided as three BAM files; we used samtools (http://samtools.org) to merge these into a single dataset. This allowed us to examine the probability of heterozygosity at a sample of SNP sites in the catalog of Genome-Wide Association Studies (accessed 3 July 2012) (Hindorff and Sethupathy, 2009) associated with CD risk. We extracted all unique CD risk sites for which data was available in the Iceman genome. The final sample included forty-nine SNP sites. Since the risk variant is not reported for all sites in the GWAS catalog, we considered average heterozygosity across all forty-nine sites as an estimate of

the degree of background risk present in a single individual and compared the Iceman to all individuals in the 1000 Genomes Project dataset samples listed above.

Sample statistic

We calculated sample weighted F_{ST} for all comparisons using in-house Python scripts (www.python.org) according to the method outlined by Akey and colleagues (2002).

Pairwise F_{ST} was calculated between each Old World continental region using two samples from each geographic region (LWK x YRI, CHB x CHS, CEU x TSI). Additionally, we assessed the differentiation within each region by comparing a single pair of populations from the region.

Differentiation at each locus

We pursued an empirical approach similar to Pagani and colleagues (Pagani et al., 2011) to determine the level of differentiation at each CD risk locus we calculated pairwise F_{ST} for all sites in the region that were segregating in at least one sample in the comparison ($F_{ST} > 0$). We used an empirical approach instead of a model-based approach because of the heterogeneity of human population histories and our lack of detailed knowledge about them across the entire timescale important to CD risk. Demography affects all loci across the genome, so a demonstration that a group of loci is surprisingly different from the genome-wide pattern of differentiation would allow us to reject the hypothesis that demography alone is sufficient to explain their distribution. The 99% upper boundary for a randomly chosen set of over 2 million (2,200,000) SNPs from the entire genome was used as the empirical significance level for each comparison. Therefore, we considered

as outliers all SNPs exceeding this upper boundary. A region was considered excessively differentiated if more than 1% of SNPs in the region exceeded this upper boundary.

While the above approach is a first step to determine if a region has an elevated proportion of high F_{ST} variants, this method does not include a formal statistical test. Yu and colleagues (2009) demonstrated the value of empirical comparison of a candidate locus to randomly selected regions of the genome ascertained in a similar manner to detect signs of natural selection. Therefore, to determine the significance of each CD locus, for each population comparison we randomly selected eleven thousand 50Kb regions from the genome and calculated the number of excessively differentiated SNPs for each of these (using the same process as for the CD loci). The number of randomly selected genomic regions with a percentage of excessively differentiated SNPs greater than or equal to that of the CD locus was taken as an empirical P-value for the CD locus. All comparisons were analyzed using in-house Python scripts (www.python.org).

Our test is a one-tailed comparison screening for an excess fraction of SNP loci with high F_{ST} across a region. Low- F_{ST} regions may also reflect interesting evolutionary dynamics, but we have chosen to test the hypothesis that regions have experienced the same evolutionary dynamics. This hypothesis can be rejected if directional selection has acted in different directions on the allele frequencies in different populations, leading to high differentiation.

Reduction of heterozygosity in Ötzi from low-coverage

To determine the effect of low coverage in the Ötzi genome on average heterozygosity across a small number of sites, we used the average weighted derived allele frequency in

Europe (from 1000 Genomes Project) for each site to randomly generate a genotype for each site (repeated 500 times). For each genotype we randomly chose reads for each site according to the number of reads present at that site in the Ötzi genome (repeated 2500 times). This sampling procedure resulted in a set of 500 average heterozygosities expected based on allele frequency and 1,250,000 average heterozygosities based on allele frequency and random sampling of reads.

REFERENCES

Abadie V, Sollid LM, Barreiro LB, and Jabri B. 2011. Integration of Genetic and Immunological Insights into a Model of Celiac Disease Pathogenesis. Annual Review of Immunology 29:493–525.

Akey JM. 2002. Interrogating a High-Density SNP Map for Signatures of Natural Selection. Genome Research 12:1805–1814.

Albrechtsen A, Moltke I, and Nielsen R. 2010. Natural Selection and the Distribution of Identity-by-Descent in the Human Genome. Genetics 186:295–308.

Armelagos GJ, and Harper KN. 2005. Genomics at the origins of agriculture, part two. Evol Anthropol 14:109–121.

Barker J, and Liu E. 2008. Celiac Disease: Pathophysiology, Clinical Manifestations, and Associated Autoimmune Conditions. Advances in pediatrics 55:349–365.

Barreiro LB, and Quintana-Murci LIS. 2010. From evolutionary genetics to human immunology: how selection shapes host defence genes. Nat Rev Genet 11:17–30.

Dubois PCA, Trynka G, Franke L, Hunt KA, Romanos J, Curtotti A, Zhernakova A, Heap GAR, Ádány R, et al. 2010. Multiple common variants for celiac disease influencing immune gene expression. Nature Genetics 42:295–302.

Durbin RM, Altshuler DL, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, Collins FS, La Vega De FM, et al. 2010. A map of human genome variation from population-scale sequencing. Nature 467:1061–1073.

Fu Q, Rudan P, Pääbo S, and Krause J. 2012. Complete Mitochondrial Genomes Reveal Neolithic Expansion into Europe. PLoS ONE 7:e32473.

Gasbarrini G, Rickards O, and Martínez-Labarga C. 2012. Origin of celiac disease: How old are predisposing haplotypes? World Journal of Gastroenterology 18:5300–5304.

Hancock AM, Alkorta-Aranburu G, Witonsky DB, and Di Rienzo A. 2010. Adaptations to new environments in humans: the role of subtle allele frequency shifts. Philosophical Transactions of the Royal Society B: Biological Sciences 365:2459–2468.

Hawks J, Wang ET, Cochran GM, Harpending HC, and Moyzis RK. 2007. Recent acceleration of human adaptive evolution. Proceedings of the National Academy of Sciences 104:20753–20758.

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

Hunt KA, Zhernakova A, Turner G, Heap GAR, Franke L, Bruinenberg M, Romanos J, Dinesen LC, Ryan AW, et al. 2008. Newly identified genetic risk variants for celiac disease related to the immune response. Nature Genetics 40:395–402.

Kagnoff MF. 2007. Celiac disease: pathogenesis of a model immunogenetic disease. Journal of Clinical Investigation 117:41–49.

Keller A, Graefen A, Ball M, Matzas M, Boisguerin V, Maixner F, Leidinger P, Backes C, Khairat R, et al. 2012. New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing. Nat Comms 3:698.

Larsen CS. 2006. The agricultural revolution as environmental catastrophe: Implications for health and lifestyle in the Holocene. Quaternary International 150:12–20.

Nistico L. 2006. Concordance, disease progression, and heritability of coeliac disease in Italian twins. Gut 55:803–808.

Pagani L, Ayub Q, Macarthur DG, Xue Y, Baillie JK, Chen Y, Kozarewa I, Turner DJ, Tofanelli S, et al. 2011. High altitude adaptation in Daghestani populations from the Caucasus. Hum Genet 131:423–433.

Sabeti P, Reich D, Higgins J, Levine H, Richter D, Schaffner S, Gabriel S, Platko J, Patterson N, et al. 2002. Detecting recent positive selection in the human genome from haplotype structure. Nature 419:832–837.

Sánchez-Quinto F, Schroeder H, Ramirez O, Avila-Arcos MC, Pybus M, Olalde I, Velazquez A, Marcos MEP, Encinas JMV, et al. 2012. Genomic Affinities of Two 7,000-Year-Old Iberian Hunter-Gatherers. Current Biology 22:R631–R633.

Skoglund P, Malmström H, Raghavan M, Storå J, Hall P, Willerslev E, Gilbert MTP, Götherström A, and Jakobsson M. 2012. Origins and Genetic Legacy of Neolithic Farmers and Hunter-Gatherers in Europe. Science 336:466–469.

Sollid LM. 2000. Molecular basis of celiac disease. Annual Review of Immunology

18:53–81.

Soni S, and Badawy S. 2010. Celiac disease and its effect on human reproduction. The Journal of Reproductive Medicine 55.

Soranzo N, Spector T, Mangino M, Kühnel B, Rendon A, Teumer A, Willenborg C, Wright B, Chen L, et al. 2009. A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. Nature Genetics 41:1182–1190.

Trynka G, Hunt KA, Bockett NA, Romanos J, Mistry V, Szperl A, Bakker SF, Bardella MT, Bhaw-Rosun L, et al. 2011. Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. Nature Genetics 43:1193–1201.

Van Heel DA, Franke L, Hunt KA, Gwilliam R, Zhernakova A, Inouye M, Wapenaar MC, Barnardo MCNM, Bethel G, et al. 2007. A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21. Nature Genetics 39:827–829.

Voight BF, Wen X, and Pritchard JK. 2006. A Map of Recent Positive Selection in the Human Genome. Plos Biol 4:e72.

Wolters VM, and Wijmenga C. 2008. Genetic Background of Celiac Disease and Its Clinical Implications. The American Journal of Gastroenterology 103:190–195.

Yu F, Keinan A, Chen H, Ferland RJ, Hill RS, Mignault AA, Walsh CA, and Reich D. 2009. Detecting natural selection by empirical comparison to random regions of the genome. Human Molecular Genetics 18:4853–4867.

Zhernakova A, Elbers CC, Ferwerda B, Romanos J, Trynka G, Dubois PC, de Kovel CGF, Franke L, Oosting M, et al. 2010. Evolutionary and Functional Analysis of Celiac Risk Loci Reveals SH2B3 as a Protective Factor against Bacterial Infection. The American Journal of Human Genetics 86:970–977.

Zhernakova A, Stahl EA, Trynka G, Raychaudhuri S, Festen EA, Franke L, Westra H-J, Fehrmann RSN, Kurreeman FAS, et al. 2011. Meta-Analysis of Genome-Wide Association Studies in Celiac Disease and Rheumatoid Arthritis Identifies Fourteen Non-HLA Shared Loci. PLoS Genetics 7:e1002004.

Figure 1. Genome-Wide High- F_{ST} Ratio Decay and Non-HLA Risk Loci Values

In each plot, the histogram represents the genome-wide distribution (11,000 samples) of the fraction of SNPs in each locus with F_{ST} higher than the 99% genome-wide upper boundary (black line) this distribution is plotted against the high- F_{ST} ratio for each CD risk locus bin in this analysis ordered from smallest to largest (full data set in Appendix). Dots above the line represent loci with significantly elevated high- F_{ST} ratios ($P < 0.05$). Raw data for CD regions is in Appendix 1. a. within Africa, b. within East Asia, c. within Europe, d. between Europe/East Asia, e. between Africa/East Asia, f. between Africa/Europe.

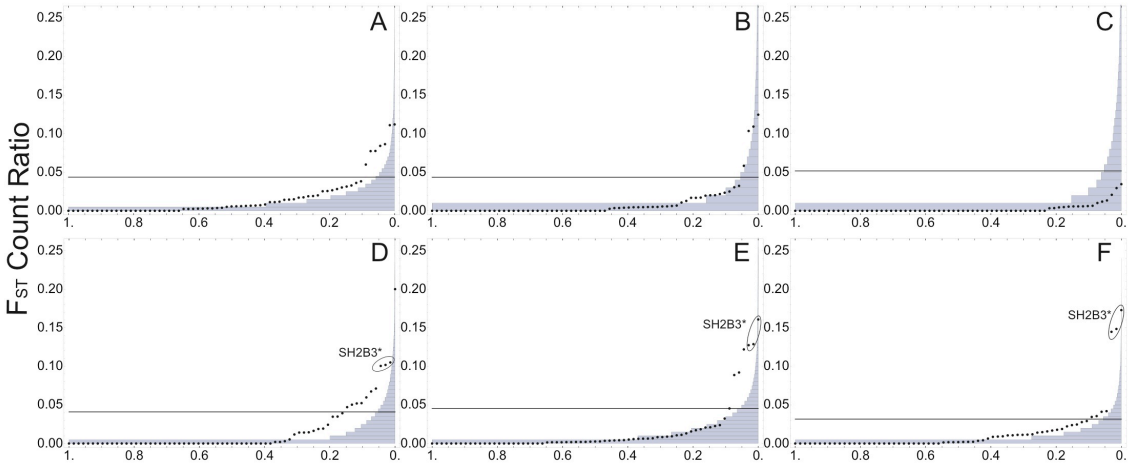


Figure 2. Mean Observed Heterozygosity Across CD GWAS SNPs from 1000 Genomes and Iceman

Histograms represent the distributions of observed average heterozygosities across 49 GWAS SNPs associated with CD in Europeans (Red), East Asians (Green), Africans (Blue), and the Iceman (black line). Illustrates that the Iceman's mean heterozygosity across CD associated loci is low compared to present Europeans.

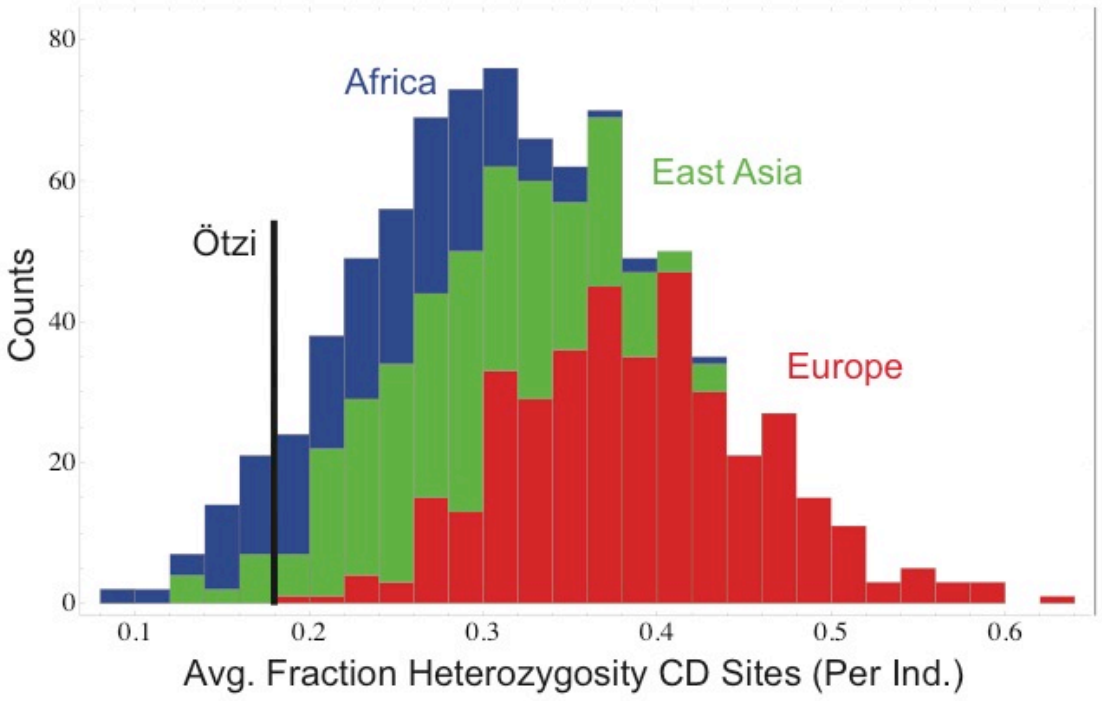
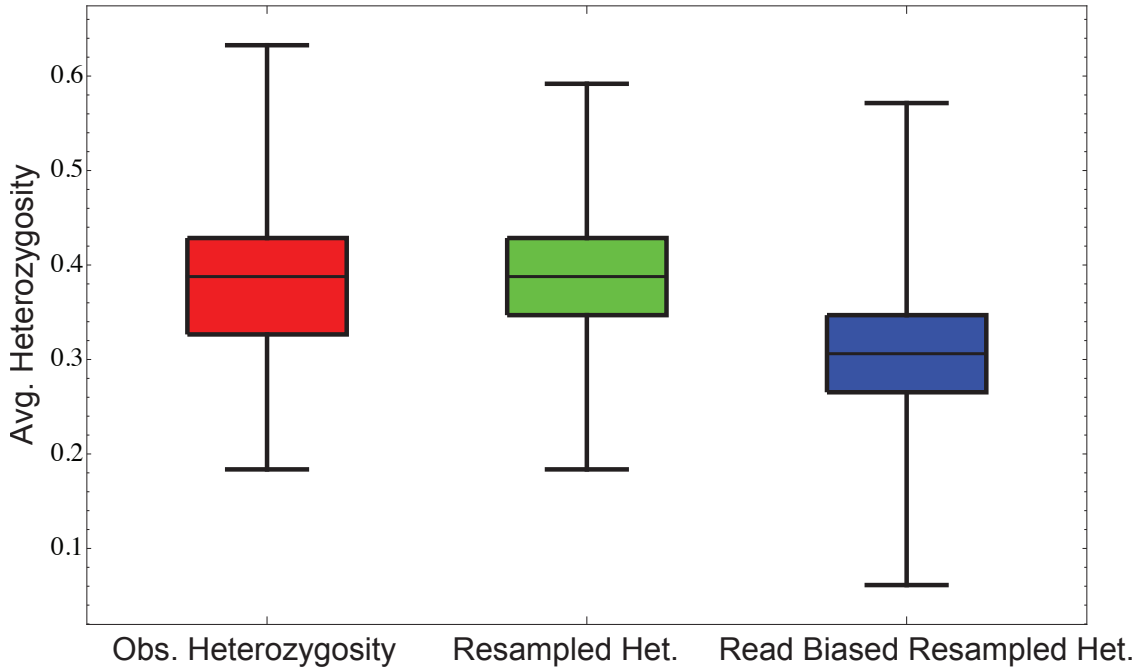


Figure 3. Estimated effect of low coverage in the Iceman genome on mean heterozygosity at GWAS loci.

Box plot illustrating the distribution of observed mean heterozygosity across 49 GWAS SNPs in present Europeans from the 1000 Genomes Project (Red), the expected distribution of mean heterozygosity based on resampling from allele frequencies in Europe (Green), and the expected distribution of mean heterozygosity based on allele frequencies in Europe and random sampling of alleles according to the distribution of read coverage in the Iceman (Blue). Although there is a significant reduction in mean heterozygosity, the range of values is similar, suggesting that the observed reduction of heterozygosity in the Iceman compared to present Europeans is genuine.



CHAPTER III

Celiac Disease as a Model for the Evolution of Multifactorial Diseases in Humans.

Aaron Sams and John Hawks

ABSTRACT

Celiac disease (CD) is a multifactorial chronic inflammatory condition that results in injury of the mucosal lining of the small intestine upon ingestion of wheat gluten and related proteins from barley and rye. Although the exact mechanisms leading to CD are not fully understood, the genetic basis of CD has been relatively well characterized. In this review we briefly review the history of discovery, clinical presentation, pathophysiology, and current understanding of the genetics underlying CD risk. Then, we discuss what is known about the current distribution and evolutionary history of genes underlying CD risk in light of other evolutionary models of disease. Specifically, we conclude that the set of loci underlying CD risk did not cohesively evolve as a response to a single past selection event such as the development of agriculture. Rather, deterministic and stochastic evolutionary processes have both contributed to the present distribution of variation in CD risk loci. Selection has shaped some components of this network, but this selection appears to have occurred at different points in the past. Other

parts of the CD risk network have likely arisen due to stochastic processes such as genetic drift.

Key words: celiac disease, gene networks, complex traits, natural selection

INTRODUCTION

Celiac disease (CD) is a multifactorial chronic inflammatory condition that results in injury of the mucosal lining of the small intestine. The earliest descriptions of a condition with symptoms like CD date back to the first and second century writings of the Greek physician Arataeus (Simoons, 1981; Losowsky, 2008). The first modern description of CD is attributed to English physician Samuel Jones Gee, who was familiar with the writings of Arataeus. In 1888 (Dowd and Walker-Smith, 1974; Simoons, 1981; Losowsky, 2008)**,** Gee described the 'coeliac affection' as a condition of chronic indigestion associated with diarrhea, wasting, and weakness and common in people of all ages. Gee's report was the first to hypothesize dietary factors as a primary cause of the condition (Dowd and Walker-Smith, 1974). Although earlier researchers noted the value of a starch free diet for prevention of CD (Haas, 1924), the role of diet as a cause of CD was experimentally confirmed in 1953 by Dicke and colleagues (1950; 1953) who noted the detrimental effect of wheat flour (but not wheat starch) on CD patients. Shortly thereafter, Anderson and colleagues (Anderson et al., 1952; Alvey et al., 1957) confirmed the effect of wheat on CD and identified wheat gluten as the primary culprit.

CD appears to be an evolutionary paradox. It is common today (>1%) in several populations with long histories of wheat agriculture including Europe and the Near East

(Catassi et al., 1996; Corrao et al., 1996; Mäki et al., 2003; Rubio-Tapia et al., 2012) and has been documented in frequencies around 5-6% among the Saharawi of northern Africa (Catassi et al., 1999; 2001). The effectiveness of a gluten-free diet as a treatment has been clinically recognized only since the discovery of gluten as the CD trigger in the 1950's (Barker and Liu, 2008). In past populations CD should have had a high potential to reduce the fitness of its sufferers directly, or by interfering with nutrient absorption, or by other effects that reduce fertility (Corrao et al., 2001; Soni and Badawy, 2010). A long history of CD risk in these populations requires some evolutionary explanation. Founder effects in ancient populations might have increased the frequency of deleterious CD risk alleles, or these alleles may have been influenced by positive selection on their other pleiotropic effects. Much evidence supports the role of recent natural selection in shaping the HLA region (Albrechtsen et al., 2010). Less well understood is the influence of selection on the more than 40 non-HLA loci associated with CD risk.

In this review we briefly cover the clinical presentation, pathophysiology, and current understanding of the genetics underlying CD risk. Then, we discuss what is known about the current distribution and evolutionary history of genes underlying CD risk in light of other evolutionary models of disease.

Clinical presentation and treatment

The clinical presentation of CD is highly variable and depends on age (Fasano, 2005; Barker and Liu, 2008). The classical symptoms or tip of the "celiac disease iceberg" include abdominal symptoms such as pain and distension, diarrhea, malnutrition, and failure to thrive within the first few years of life (Barker and Liu, 2008). Marsh (1992)

developed a scoring system for assessing damage to the mucosal lining of the small intestine in CD. This system ranges from normal upper intestinal histology to increasing levels of crypt hyperplasia often accompanied by intraepithelial lymphocytes and ending in any degree of villous atrophy. In addition to intestinal histology, diagnosis is typically first assessed by the presence of tissue transglutaminase (TGase) autoantibodies, such as immunoglobulin A (IgA) and confirmed after intestinal biopsy by treatment with a gluten-free diet (GFD) (Farrell and Kelly, 2002).

Many complications from CD have been noted. CD can be associated with vitamin (D and K) and micronutrient (calcium, iron, folate) deficiencies, which often result in conditions such as rickets, osteoporosis, and anemia in prolonged cases (Farrell and Kelly, 2002; Fasano, 2005; Green and Cellier, 2007). Linear enamel hypoplasias on the teeth have been noted in greater than twenty percent of CD cases in children (Bucci et al., 2007; Campisi et al., 2007; Cheng et al., 2010) but see Procaccini and colleagues (2007) for a contrasting result. Neurological abnormalities including hypotonia, developmental delay, learning disorders and ADHD, headache, and cerebellar ataxia have also been reported to be more common in children with CD (Zelnik et al., 2004). Patients with CD have a slightly increased risk for developing gastrointestinal malignancies such as adenocarcinoma(Barker and Liu, 2008). Finally, as noted earlier, complications from CD may also include reduced fertility and increased mortality (Corrao et al., 2001; Soni and Badawy, 2010).

Celiac disease is more common among certain disease cohorts than expected. In particular, several researchers have noted a high occurrence of CD in patients diagnosed with type 1 diabetes (T1D) (Greco et al., 2012; Pham Short et al., 2012). Increased

prevalence of CD among autoimmune thyroid disease sufferers has also been frequently noted (reviewed in (Ch'ng et al., 2007)). These associations may be expected due to the fact that these conditions share the HLA-DQ2 and HLA-DQ8 risk genotypes (Ch'ng et al., 2007; Pham Short et al., 2012). In addition to T1D, shared (non-HLA) genetic risk factors with CD have been noted for other autoimmune conditions such as Crohn's disease (CrD) (Festen et al., 2011), rheumatoid arthritis (RA) (Zhernakova et al., 2011).

The wide range of clinical symptoms and complications associated with CD, particularly those that are more severe in children, support the notion of CD as an evolutionary paradox. Prior to the recent discovery of gluten as the trigger for CD, we might expect CD to have lowered evolutionary fitness in individuals in the past afflicted with the condition.

Diet, HLA, and the pathophysiology of CD

Celiac disease is activated by proline-rich and glutamine-rich gluten proteins found in wheat and related proteins in rye and barley (Kagnoff, 2007) (and in rare cases Oats (Koning, 2012)). Gluten is composed of two major protein types, gliadins and glutenins. The high proline content of these peptides prohibits their complete digestion by gastric, pancreatic, and brush border enzymes in the gastrointestinal tract (Shan et al., 2002). Once ingested, gluten is partially digested into large high-proline and high-glutamine peptides, such as the 33 amino acid α-gliadin peptide, which can accumulate in the small intestine. Although the primary environmental trigger of CD is wheat gluten and related proteins, diet is not sufficient to explain the occurrence of the disease (Green and Cellier, 2007; Kagnoff, 2007).

The pathogenesis of CD is rooted in genetic factors. This fact became evident and has been confirmed by various family and twin studies of CD (Risch, 1987; Petronzelli et al., 1997; Bevan et al., 1999; Greco, 2002; Nistico, 2006). Human leukocyte antigen (HLA) genes have been securely established as the primary genetic component of CD. HLA molecules are antigen-presenting molecules, which bind antigenic peptides and present them to samples of circulating T-cells (Lawlor et al., 1990). More precisely, with respect to the cell, HLA class I molecules bind endogenous antigenic peptides and class II molecules bind exogenous antigenic peptides (Townsend and Bodmer, 1989; Lawlor et al., 1990). Although early research showed association to HLA class I alleles (Falchuk et al., 1972) and the class II HLA-DR locus (DeMarchi et al., 1979), further research has demonstrated that CD is most strongly associated with specific HLA class II alleles at the HLA-DQ locus. Specifically, the alleles encoding the HLA-DQ2 and HLA-DQ8 molecules are necessary, but not singularly responsible for CD risk. The majority of CD cases (90-95%) are associated with the HLA-DQ2 heterodimer, which is encoded by the alleles HLA-DQA1*05 and HLA-DQB1*02 either in cis, on the same haplotype, or in trans (Sollid, 1989; Sollid and Thorsby, 1990; Spurkland et al., 1990). Additionally, a gene dosage effect has been observed for individuals homozygous for HLA-DQB1*02 (Karinen et al., 2006; Murray et al., 2007). Of the approximately ten percent of cases not carrying the HLA-DQ2 heterodimer, half carry the HLA-DQ8 heterodimer (HLA-DQA1*03/HLA-DQB1*03:02) and the other half carry either HLA-DQA1*05 or HLA-DQB1*02 alone (Karell et al., 2003). Further research has demonstrated the likely involvement of genes in the extended HLA region in CD risk (Margaritte-Jeannin et al., 2004; Ahn et al., 2012).

The role of the HLA-DQ2 and HLA-DQ8 heterodimers in CD pathogenesis is well understood. HLA-DQ2 and HLA-DQ8 molecules contain positively charged pockets, which preferentially bind to negatively charged peptides. HLA-DQ heterodimers on antigen-presenting cells bind gluten peptides in which glutamine has been deamidated (converted to negatively charged glutamic acid) via tissue transglutaminase (Dieterich et al., 1997; Molberg et al., 1998; van de Wal et al., 1998; Fleckenstein et al., 2002; Fleckenstein, 2004). The bound gluten-HLA-DQ complexes are then presented to populations of CD4+ T-cells in the lamina propria of the small intestine. The T-cells that recognize the gluten-HLA-DQ complexes then produce pro-inflammatory cytokines such as interferon-γ, which likely play a role in the mucosal damage that is characteristic of CD (Nilsen et al., 1998). Importantly, different wheat, rye, and barley each contain a variable repertoire of immunogenic peptides and not all CD patients respond to the same sets of peptides (Koning, 2012), which makes the development of non-immunogenic grains difficult (Kagnoff, 2007; Koning, 2012).

In addition to the adaptive immune response described above, gliadin peptides can also activate an innate immune response. This innate immune response, which is characterized by an increase in the number of intra-epithelial lymphocytes (IEL's) in CD patients is probably related to ongoing pathogenesis of CD such as the development of refractory CD (non-responsive villous atrophy and malabsorption) and the development of enteropathy-associated T-cell lymphoma (Kagnoff, 2007). Interleukin-15 (IL-15) is up-regulated in the lamina propria and the epithelium during the course of CD pathogenesis (Mention et al., 2003). The increased presence of IL-15 results in the activation of IEL's expressing the natural killer immunoreceptor NK-G2D. Activated

IEL's then become cytotoxic and kill enterocytes with MIC cell-surface expression (Hüe et al., 2004; Meresse et al., 2004; Green and Cellier, 2007). For a more thorough review of CD pathogenesis see (Green and Cellier, 2007; Kagnoff, 2007; Abadie et al., 2011).

Non-HLA genetic factors-GWAS and fine-mapping studies

No compelling non-HLA loci were associated with CD prior to the advent of genome-wide association studies (GWAS) (Kumar et al., 2012). The first GWAS study of CD in 2007 (Van Heel et al.) included a cohort of 778 CD individuals and 1,422 controls from the UK. This study tested for association to approximately 310,000 single-nucleotide polymorphisms (SNPs) and the top 1,500 SNPs were re-analyzed in a replication cohort of approximately twice the size of the original (Van Heel et al., 2007; Hunt et al., 2008; Trynka et al., 2009). This first GWAS study identified 13 non-HLA CD risk loci. A later CD GWAS by Dubois and colleagues (2010) included over 4,500 CD cases and over 11,000 controls from four populations (Finland, Italy, the Netherlands, the UK). This study included replication of the 131 top SNPs in seven European cohorts comprising approximately 5,000 CD cases and 5,500 controls and added another thirteen risk loci to the list of CD associated genomic regions. Importantly, this study also demonstrated that nearly fifty percent of CD risk SNPs affect nearby expression quantitative trait loci (eQTLs), suggesting a likely role for downregulation of gene expression in CD. As mentioned above, additional GWA studies have identified risk loci shared between CD and other autoimmune conditions such as T1D, RA, CrD, and ulcerative colitis (UC) (Zhernakova et al., 2007; Barrett et al., 2009; Festen et al., 2009; 2011; Gutierrez-Achury et al., 2011; Zhernakova et al., 2011). These studies have

illuminated the tremendously pleiotropic nature of loci underlying autoimmune and other immune related conditions.

A major step forward in the identification and resolution of risk loci underlying CD was a fine-mapping study conducted by Trynka and colleagues (2011). Rather than using the same genotype arrays as earlier CD studies, which contained 300,000 to 550,000 SNPs ascertained primarily in Europeans and dispersed throughout the genome, this study utilized the Immunochip (Cortes and Brown, 2011), a custom Illumina array designed to increase SNP coverage density in regions with existing associations to twelve immune-mediated diseases (including CD) in Europeans (~200,000 SNPs total). Trynka and colleagues (2011) identified a total of 40 genomic loci (including HLA) associated with CD. However, the additional resolution from the Immunochip array allowed the researchers to identify multiple independent signals in 13 loci, bringing the total number of non-HLA CD associated loci to 57 with 29 signals confined to single genes.

Genetic architecture and environmental factors

A large proportion of risk for CD can be attributed to genetic factors. The most recent large-scale twin-study of heritability has estimated the heritability (h^2) of CD, given a population prevalence of 1/91 in Europeans at approximately 87%(Nistico, 2006). The fine-mapping study by Trynka and colleagues (2011) has brought the total percentage of that genetic variance currently explained to ~57%, with 40% due to the primary HLA risk variants. Some combination of other effects likely explains the remainder of the genetic variance that is expected from population data.

Traditional GWAS approaches are statistically underpowered to identify causal rare-variants due to the fact that sample sizes needed to establish a significant association with rare variants (MAF < 1%) are much larger than needed for common variants (Gorlov et al., 2008). Further, traditional genotyping arrays primarily consist of common variants. It has been proposed that rare-variants may explain larger proportions of risk than common variants in certain cases (Bodmer and Bonilla, 2008; Asimit and Zeggini, 2010), particularly those that are population specific (Kumar et al., 2012). Therefore, although much of CD risk can be explained by a common disease/common variant model, new methods for identifying rare-variant associations hold the promise to discover additional rare risk loci underlying this condition (Asimit and Zeggini, 2010). Similar statistical factors make it impossible for current GWAS approaches to identify common variants with very small effect sizes. Doing so would require extremely large cohort sizes (Kumar et al., 2012).

Further, traditional models used to estimate h^2 from population occurrence data typically assume that the trait does not involve gene-gene interactions (epistasis). Failure to include epistasis in models of h^2 where it does exist likely inflates the estimate of h^2 for the trait (Zuk et al., 2012). Recent developments in identifying epistatic interactions (Rao et al., 2011; Ma et al., 2012; Rajapakse et al., 2012) hold promise for explaining some of the "missing" h^2 underlying CD risk.

The remainder of CD risk must lie in environmental risk factors and gene by environment interactions. Although the pathophysiology of CD is well understood, mysteries remain. For example, environmental factors such as enteric viruses (Kagnoff, 1984; Stene et al., 2006; Kagnoff, 2007) may in some cases affect the permeability of the

mucosal lining of the small intestine, allowing gluten peptides to access the underlying mucosal layers. Additionally, variation in immunogenic peptides in wheat, rye, and barley may explain some of the variation in CD symptoms (Koning, 2012). A more complete understanding of the non-genetic factors that contribute to CD will improve our ability to study the evolutionary history of this condition.

Fortunately, the primary environmental risk factor underlying CD risk (gluten) is known, which makes CD a useful model for studying the evolutionary genetics of complex disease, particularly within a biocultural evolutionary framework. The well-understood history of grain domestication (Kilian et al., 2009; Willcox, 2012) by humans gives us additional information that we can use to illuminate the evolutionary history of CD. Several researchers have argued that CD likely rose to its current frequency due to positive selection on risk variants from some beneficial effect on the immune system within the context of increasing human population density (Barreiro and Quintana-Murci, 2010; Zhernakova et al., 2010; Abadie et al., 2011). However, a more complex model of CD may be necessary to explain the history of the full network of genes eliciting the CD response.

EVOLUTION OF CD

Knowledge of the worldwide prevalence of CD is incomplete but growing (Gandolfi et al., 2000; Catassi et al., 2001; Mäki et al., 2003; Greco et al., 2012; Riddle et al., 2012; Rubio-Tapia et al., 2012). Further, there is much variation across populations in the consumption of cereals that can induce CD (Abadie et al., 2011). CD is more common among populations with European and Near Eastern ancestry. This is somewhat of an

evolutionary enigma because these are precisely the populations with the longest history of wheat, rye, and barley agriculture (Tresset and Vigne, 2011; Zeder, 2011). Although the population incidence of CD increases with age, it is nonetheless notable in children and young adults before and during the reproductive lifespan. Mortality and morbidity associated with CD would have exerted fitness costs in past populations. The persistence of CD therefore requires some explanation.

Several hypotheses present possible explanations:

1. Founder effects in ancient populations might have increased the frequency of deleterious CD risk alleles in at-risk populations or these alleles may have been influenced by positive selection on their other pleiotropic effects.
2. The genetic architecture of CD may be such that most loci contributing to CD risk have only minor effects on risk (fitness), meaning that selection against CD would have miniscule to nonexistent effects on any single risk locus (Stranger et al., 2011).
3. Balancing or frequency-dependent selection may have maintained some risk-associated variants at high frequencies in past populations. In particular, it seems likely that either long-term balancing selection, including overdominance, frequency-dependent selection, or selection in a fluctuating (pathogen) environment has led to the current distribution of HLA risk variants in the ancestors of present at-risk populations (Albrechtsen et al., 2010).

HLA and balancing selection

Balancing selection is a common evolutionary mechanism underlying risk for many diseases in present populations. The classic single-gene disease models with balancing selection are haemoglobinopathies, structural abnormalities of globin proteins. The most famous case is sickle-cell disease (Frenette and Atweh, 2007), in which the sickle-cell gene has been selected due to overdominance of heterozygous individuals in malarial environments.

Much evidence supports the role of recent natural selection in shaping the HLA region (Albrechtsen et al., 2010). Some HLA haplotypes have been directionally selected in recent human populations, but there is no clear pattern of directional selection on CD-risk-associated HLA haplotypes (Abadie et al., 2011). HLA has long been observed to be under balancing selection in human populations and ancestral hominins. Recently, Solberg and colleagues (2008) applied the Ewens-Watterson test of homozygosity on a large composite global dataset of HLA allele frequencies to demonstrate that balancing selection is widespread among HLA loci. The class II genes HLA-DQA1 and HLA-DQB1 are among the most strongly selected loci in this study, showing signals of balancing selection in Europe, North Africa, and Southwest Asia, precisely where CD is most common.

Some researchers have argued that the cause of this balancing selection in HLA is overdominance (Takahata and Nei, 1990; Takahata et al., 1992) within environments with particular pathogen environments. Recent evidence pertaining to this hypothesis is mixed. Balancing selection based on fitness overdominance does not tend to increase identity-by-descent (IBD) in the region linked to a selected locus, yet Albrechtsen and

colleagues (2010) found strong IBD signals for HLA. Strong IBD is itself a sign of selection, but is more likely produced by fluctuating selection due to coevolution of immunity with pathogens, for example in a frequency-dependent pattern. HLA risk alleles account for approximately 40% of the additive variance underlying CD risk, suggesting that balancing selection of some kind may in part explain the high frequency of CD in European and North African populations today.

Under this hypothesis, CD is not a target of selection but instead is a fitness-reducing side effect of selection for pathogen resistance. Pathogen pressure has not been uniform across all of human evolution. Instead, there is substantial evidence that new pathogen pressures intensified during the last 10,000 years. Bioarchaeological evidence demonstrates that dietary and demographic transitions of the Holocene were detrimental to human health (Larsen, 2006). Dietary specialization led to nutritional deficits and denser living conditions created ideal conditions for the spread of new pathogens, as documented by higher rates of porotic hyperostosis, cribra orbitalia, dental caries, linear enamel hypoplasias, tuberculosis, and trepanematoses in populations after the onset of agricultural production (Armelagos and Harper, 2005; Larsen, 2006). Genetic and archaeological evidence show the importance of adaptive evolution concurrent with large-scale demographic change over the last 10,000 years (Hawks et al., 2007). These observations provoke the hypothesis that the Holocene transition to agriculture was a time of strong natural selection on genes important to diet and immunity. Providing that the genetic locus of largest effect in CD is HLA (Trynka et al., 2011), and that genetic loci associated with autoimmune diseases in general are enriched for long-range haplotype signals of recent positive selection (Barreiro and Quintana-Murci, 2010), a

plausible hypothesis to explain the present distribution of CD is that the recent shift in demography and settlement patterns in human populations led to increased selection on immune-related genes due to elevated pathogen pressure.

GWAS, genetic architecture, and selection on complex traits

Recent strong selection does appear consistent with the evolution of HLA, but HLA is only one component of the overall additive genetic variance in CD risk. A large network of other loci is known to explain some of the additive variance. We refer to these loci as "background risk" loci, because (a) their effect on CD risk is conditioned on the presence of the major HLA risk alleles, and (b) the effect of any single one of these loci on CD risk is relatively slight. The evolution of this set of loci may have been shaped mainly by other phenotypic associations or by random genetic drift, meaning that they may show very heterogeneous patterns of variation with respect to each other. In order to discuss the evolution of background risk of CD, we must more deeply probe the dynamics of this broader set of loci.

Genome-wide association studies (GWAS) have produced nearly all the data that link non-HLA loci with CD risk. GWAS is an unbiased statistical approach to identify phenotype/allelic correlations, with respect to genomic structure and the etiology of the phenotype under consideration (Stranger et al., 2011). A major goal of GWAS is to provide information about the underlying biology of a trait. Systems genetics approaches, which analyze GWAS and other association data in combination with functional genomics data, can be combined with database information approaches, such as gene ontology (GO) (Botstein et al., 2000), gene-set enrichment analysis (GSEA)

(Subramanian et al., 2005) and similar types of analysis (Stranger et al., 2011), to provide avenues for identifying the complete network that underlies a complex trait.

At this point it is important to clarify the term "network." CD risk loci that have been identified by GWAS represent a network only in their joint association with the CD phenotype. Pathway analysis can additionally examine whether some or all of these genes interact via co-expression or co-involvement in other phenotypes. In the case of CD, pathway analysis of GWAS loci has facilitated the identification of several biochemical pathways underlying CD including T-cells, NK-cells, B-cells, and neutrophils (Kumar et al., 2012), greatly contributing to our understanding of CD pathogenesis. In that sense, different CD background risk loci are parts of different functionally integrated networks of genes, each related to a discrete biochemical pathway. Dietary gluten can, within a certain biochemical context, have major effects on these different pathways, and the pathogenesis of CD is modified by certain genes within those pathways. Yet these pathways have many phenotypic consequences beyond CD, and many of the genes in these pathways are not associated with CD risk. In that respect, the functional integration of CD risk loci is unclear.

Functional integration is only one way of defining a genetic system. A different way of considering genetic networks is evolutionary integration. A set of genes that respond together to a single selection pressure may cause several distinct functional pathways to exhibit strong patterns of similarity in genetic variation or differentiation. Such an evolutionarily integrated network may be recognized by a common pattern of evidence for positive selection, balancing selection, loss of functional constraint, or differentiation among populations. Different populations that share common

environmental factors may also help to illuminate networks of genes that respond to selection associated with similar environments (Hancock et al., 2010a).

Evolutionary integration of a network of genes requires a fairly narrow set of conditions. The genetic architecture (number of loci and effect sizes of each locus) underlying a trait influences the ability of selection to shape the genetic variation underlying the trait. The classic model of quantitative trait variation is Fisher's (1918) infinitesimal model, which holds that many loci contribute only a small amount to the genetic variance of a trait. In contrast to this model, CD and other immune-related conditions tend to have at least one locus contributing a majority of the genetic variance (HLA) in addition to many other loci with small effect (Stranger et al., 2011; Kumar et al., 2012). This genetic architecture of CD more closely fits an exponential model of effect size.

A logical consequence of an exponential distribution of effect-sizes underlying CD risk is that we should expect the major-effect locus, HLA, to bear the majority of the negative fitness effects of CD. The effect size of a genetic locus on a trait is proportional to the contribution of that locus to the narrow-sense heritability (h^2), or the proportion of phenotypic variation explained by additive-genetic variation (Crow and Kimura, 2009). Further, the response of a locus to selection on a trait is directly proportional to the contribution of that locus to the heritability of the trait. This relationship can be described quantitatively by the prediction equation referred to as the Breeder's Equation (Falconer and MacKay, 1996):

$$R = Sh^2$$

Where R is the response to selection on a quantitative trait (difference in mean trait value between generations), S is the selection differential (difference in mean value of trait in original population and mean of selected parents). HLA haplotypes account for approximately 40% of the additive genetic variance of CD and the currently known non-HLA variants together account for at most 14% (Trynka et al., 2011). Therefore, any response to selection associated with CD will affect HLA genetic variation 40% / 14% = 2.86 times more than the background alleles combined. Alternatively, if we consider that the largest contribution to cumulative heritability of any single background locus is approximately 1%, then the response to selection on HLA is approximately 40% / 1% = 40 times greater than the response on the most heritable background locus.

We can conclude from these considerations that CD itself is unlikely to have driven evolutionary integration of a network of associated loci. However, it is implicit from the hypothesis that CD risk may be a *side-effect* of selection on some other phenotype that evolutionary integration may have arisen on CD risk loci from non-CD causes. The test of evolutionary integration is the pattern of evolutionary history of CD risk-associated genes. If these genes exhibit a common pattern of evolution, we may be able to discover which environmental or historical factors generated the present pattern of CD risk. Alternatively, if the evolutionary history of these loci has been heterogeneous, we may conclude that CD is an accident of unrelated evolutionary causes.

Evolution of non-HLA risk loci

We can evaluate the evolutionary history of CD risk-associated loci by several methods. Some researchers (Pickrell et al., 2009; Barreiro and Quintana-Murci, 2010; Zhernakova

et al., 2010; Abadie et al., 2011) have used the iHS method to infer a role for strong recent selection on some non-HLA GWAS risk loci in Europeans. For example, Zhernakova and colleagues (2010) found evidence of recent selection in or around the genes *IL12A*, *IL18RAP*, and *SH2B3*. The risk variant in the *SH2B3* gene is functionally involved in the NOD2 recognition pathway, suggesting that it may have been positively selected to protect against bacterial infection. The authors inferred a very recent onset of selection (between 1,200 and 1,700 years ago) by extended haplotype homozygosity (EHH)(Sabeti, P.C. et al., 2002).

In a Chapter II, we employed data from the 1000 Genomes Project (Durbin et al., 2010) to investigate whether the signals on these three loci are consistent with the broader genetic network influencing CD risk. A null hypothesis is that the evolutionary history of these loci matches that of randomly chosen loci across the genome. We tested this hypothesis by evaluating the differentiation of the chromosomal regions flanking CD-associated loci, both between and within continental populations. Gene networks whose components have been subject to recent selection may show significant regional differentiation when compared to loci randomly drawn from across the genome (Hancock et al., 2010a). This is in part because recent natural selection in human populations tends to have been geographically localized, owing to the recent spread of humans around the globe and their subsequent population growth (Voight et al., 2006; Hawks et al., 2007). Between-population comparisons across many loci can provide a powerful test of neutrality even when selection has affected standing variation instead of new mutational variants, and even when the causal loci themselves are unknown.

Non-HLA regions of the genome associated with CD risk are more likely to show significant differences between continents than are loci chosen randomly from the genome. Surprisingly, the comparison showing the most differentiation across CD risk loci was Europe against East Asia, rather than between Africans and non-Africans as might be predicted from the genome-wide average. Additionally, we compared population differentiation within continents to assess whether or not the signal of population differentiation has occurred uniformly over time. Previous work suggested that a proportion of autoimmune genetic risk (including CD) may reflect positive selection on the immune system within the last 10,000 years (Soranzo et al., 2009; Barreiro and Quintana-Murci, 2010; Zhernakova et al., 2010; Abadie et al., 2011). If the differentiation of CD risk loci between continents can be explained by recent selection within Europe, then we predicted that the variation within continents should show a distinctive pattern: Africa and Asia should show differentiation of the CD risk network no different from the genome-wide average, while Europe may show greater differentiation because rapid changes in allele frequency may have pulled populations apart for these loci. Contrary to our initial prediction, the CD risk loci showed the highest differentiation within Africa (between the Luhya and Yoruba samples). The two samples from China also diverged more highly than the genome-wide expectation. Unexpectedly, CD risk loci between the European (CEPH and Tuscani) samples did not diverge significantly from the genome-wide expectation. This observation weakens the hypothesis that the broad network of CD risk loci has experienced selection within the last 10,000 years. We recognized, however, that demographic factors might explain the lack of differentiation across European populations. For example, if a significant portion

of the ancestry of present Europeans is found among early Neolithic Near Eastern agriculturalists as has been predicted from archaeological and genomic data (Sokal et al., 1991; Haak et al., 2010; Fu et al., 2012; Gamba et al., 2012; Pinhasi and Cramon-Taubadel, 2012; Sánchez-Quinto et al., 2012; Skoglund et al., 2012), given the genetic uniformity of European samples in our analysis, CD risk loci may have been selected early enough in the history of this demographic event to have spread to all present European populations, contrary to previous estimates (Zhernakova et al., 2010).

To address this problem we utilized the genome of the Tyrolean Iceman, Ötzi (Keller et al., 2012). In Chapter II we took a first-round approach to determine the extent to which GWAS risk alleles underlying CD risk today are present in the 5,300 year-old Iceman. We measured the degree of heterozygosity across all GWAS sites in present populations and the Iceman. Our result, after correcting for a low genomic coverage bias in the Iceman genome, suggested that many of the risk loci most tightly associated with CD risk today are not present in this pre-Neolithic individual. Therefore, we cannot reject the hypothesis that some CD risk loci have been subject to recent selection.

Our previous results suggest that selection on the CD joint association network was not uniform with respect to geography or some specific ancient environment. However, it would be informative to know if past selection has played a larger role in shaping variation underlying present CD risk in Europe than stochastic forces. To this end we used the empirical p-values from our previous F_{ST} test (Appendix) on the CD loci between Africans and Europeans and the Odds-Ratios (effect sizes) of each of those loci from Trynka and colleagues (2011) and performed a Spearman rank order correlation between the two datasets, which consisted of 54 data points each. Due to constraints

from our previous F_{ST} analysis we had to combine values at a few independent CD associated loci. The resulting rank-order correlation coefficient is $r^2 = 0.268$. The salient point of this correlation is that the signal of selection at a locus does not predict the effect-size of the locus on present risk (Figure 1).

Our previous results reject the hypothesis that the joint association network underlying CD risk has experienced the same evolutionary dynamics as the rest of the genome. Those results and the results presented here also allow us to reject the hypothesis that selection on the joint association network underlying CD risk was uniform in time and space, as we might expect if the loci represented an evolutionarily integrated network.

EVOLUTIONARY NETWORKS

Based on our comparisons, CD risk does not show strong evolutionary integration. CD risk loci constitute a genetic network only in the loosest sense of common association with a trait. By considering other examples of evolutionary integration of human genetic networks, we can bring some insight into the role of natural selection and genetic drift on CD risk in human populations. Can selection across loci (from GWAS results for example) be studied in a systematic way? Furthermore, should we expect sets of interacting loci to exhibit common evolutionary histories, particularly with regards to selection?

Mendelian diseases

Natural selection in past environments has been firmly established as the cause of many common Mendelian disorders among present populations. Hemoglobinopathies and other disorders of red blood cells (RBCs) have long been recognized as likely adaptations to malaria, which typically involve mutations to red blood cells (RBCs). As with CD, it has been proposed that major changes in human demography and settlement patterns associated with the agricultural lifestyle led to an increase in human malarial burden in many populations over the past 10,000 years (Carter and Mendis, 2002). The first suggestion that thalassemias (among the disorders of RBCs) might be adaptations to past endemic malarial environments came from J. B. S. Haldane (1949) who noted the high frequencies of these disorders in coastal populations around the Mediterranean Sea where malaria was formerly endemic. A few years prior to Haldane's observation, a similar comparison between a disorder of RBCs (sickle-cell) and malaria was made by Beet in 1946, who observed a lower incidence of malaria in sickle-cell heterozygotes. The protective effect of the sickle-cell trait in heterozygous form was experimentally confirmed shortly a few years later by Allison (1954). It is now clear that several common disorders in present populations represent adaptations to endemic malarial environments of the past. These include various thalassemias (mutations to *HBA1*, *HBA2*, and *HBB* globin genes), enzymopathies (G6PD-deficiency), and ovalocytosis (structural disorders of hemoglobin caused by abnormal transmembrane proteins). Most such cases that result in severely anemic or lethal phenotypes in homozygotes, such as the sickle-cell trait, have risen to high frequency via balancing selection (Carter and Mendis, 2002; Hedrick, 2011). Other adaptations to malaria that do not have known

detrimental effects have also arisen recently in human evolution including the Duffy blood group antigen null allele and the HLA Bw53 and DQw5 haplotypes (Carter and Mendis, 2002; Hedrick, 2011). Further, many of these mutations are present within the same population.

The case of recent human adaptations to malaria provides a useful comparison for CD. Responses to malaria have evolved in several underlying biochemical pathways including hemoglobin structure and regulation, erythrocyte enzymatic activity and cytoadherence, antigen recognition, antibody response, and the proinflammatory response (Kwiatkowski, 2005). Although many of these pathways are involved with immune response, they represent several independent underlying functional genetic networks. They are united as a joint-association network insofar as they are all related to malarial resistance. In this case, the joint-association network appears to have evolved to a great extent as an evolutionarily integrated genomic network because many loci experienced geographically and temporally restricted selection due to the past and recent patterns of malarial endemicity (Hedrick, 2011). In short, Haldane was able to recognize the evolutionary causes of these disorders because their present pattern of geographic variation suggested evolutionary integration.

In contrast to responses to simple genetic disorders that have become common due to past selection, others have arisen due to stochastic factors. A now classic example of such a disorder is variegate porphyria among the Afrikaner population of South Africa. Variegate porphyria (VP) a low-penetrance autosomal dominant genetic condition of heme biosynthesis characterized by photosensitivity and in acute cases a high predisposition to develop neuropsychiatric attacks accompanied by abdominal pain,

constipation, vomiting, and high blood pressure (Brenner and Bloomer, 1980; Meissner et al., 1996). VP results from a partial deficiency of protoporphyrinogen oxidase encoded by the PPOX gene. A study of 108 unrelated VP patients from the UK and France (Whatley et al., 1999) identified 66 non-synonymous polymorphism in exons of the PPOX gene. Although VP appears to have less allelic heterogeneity in Western Europe than other porphyrias, it is fare more heterogenous in France and the UK than among South Africans of Western European ancestry, the Afrikaners. While VP has a prevalence of approximately 1 in 75,000 in Finland (Mustajoki, 1980) and probably lower occurrence in the remainder of Western Europe (Whatley et al., 1999), it has a much greater occurrence in South African Afrikaners of around 3 in 1000 (Dean, 1971; Warnich et al., 1996). Most cases in South Africa (as many as 20,000 individuals) have a single mutation (R59W) in the PPOX gene (Dean, 1971; Meissner et al., 1996; Warnich et al., 1996; Whatley et al., 1999). Unlike the case of hemoglobinopathies described above, the high prevalence of VP in South Africa can be attributed to stochastic forces, specifically a founder effect. In fact the mutation was traced back to a single Dutch couple in the late 17th century (Dean, 1971; Zeegers et al., 2004).

We have described cases of Mendelian diseases that today are far more common than we might expect based on their effects on health in present environments. In these cases the present evolutionary dynamics necessitate a functional explanation with reference to fitness in ancestral populations. These explanations are sometimes deterministic and sometimes stochastic but the simple nature of the underlying genetics facilitates our ability to understand the past dynamics that produced the present distribution of these genes in a more straightforward way. For complex genetic traits,

this process is not as simple. Nonetheless there have been several attempts to arrive at similar explanations to explain complex deleterious traits in relation to present environments and evolution.

Complex diseases: Type 2 diabetes and the "thrifty-genotype"

Neel (1962) proposed the "thrifty-genotype" hypothesis to explain another evolutionary conundrum. Obesity and type-2 diabetes (T2D) have clear negative impacts on health. They also appear to have a large genetic component (Maes et al., 1997; Clément, 2005; Mutch and Clément, 2006; Scuteri et al., 2007; Hasstedt et al., 2011; Lu et al., 2012; Palmer, 2012; Saxena et al., 2012). As with CD, why would natural selection favor alleles that predispose individuals to such diseases? Neel proposed that foragers are more likely to experience regular food shortages than agriculturalists, so that genes that adapt them to dietary shortfalls would be adaptive in hunter-gatherer contexts, but maladaptive upon adoption of agricultural foods. Therefore, the genes underlying T2D and obesity were adaptations to past dietary conditions and only detrimental in modern industrial contexts. More specifically, "thrifty" genes are predicted to work primarily through adaptations for more efficient food intake such as appetite, satiety, and an enhanced ability to store adipose fat (Neel, 1962; Prentice et al., 2005) and to a lesser extent energy expenditure. By making an individual store more adipose fat during periods of plenty, "thrifty" genes enhance survival during periods of famine. The "thrifty-genotype" hypothesis has spurred much research and is still considered by many an attractive hypothesis to explain modern rates and distribution of T2D and obesity (Eaton et al., 1988; Lev-Ran, 1999; Campbell and Cajigal, 2001; Lev-Ran, 2001; Ravussin, 2002;

Chakravarthy and Booth, 2004; Scott and Grant, 2006; Wells, 2007). Despite the fact that T2D and obesity satisfy the basic requirements for natural selection (they are heritable, they vary genetically, and likely confer effects on fitness), many researchers find the "thrifty-genotype" hypothesis to be a superficially attractive but insufficient model to explain the current distribution of T2D and obesity (Allen and Cheer, 1996; Benyshek and Watson, 2006; Speakman, 2006; Bouchard, 2007; Gibson, 2007; Speakman, 2008).

A persuasive critic of the "thrifty-genotype" hypothesis, Speakman (2006; 2008), has eloquently outlined the major flaws of the hypothesis. If human prehistory was characterized by periods of famine such that "thrifty" genes have always been under selection, based on a fairly weak amount of selection (a selective advantage of 0.1%) and basic principles of population genetics we should predict that many of these genes would have already risen to fixation and rates of T2D and obesity should be much higher in present populations. However, if the selection on "thrifty" genes only began at the start of the Neolithic (Prentice, 2005), selection would not have acted strongly enough on "thrifty" genes to explain current rates of T2D and obesity alone (Speakman, 2008). To solve this apparent problem, we might suggest that selection had simply been stronger, allowing more rapid recent change. But Speakman notes that we have historical data on famines that suggests a limit to their selective intensity. Famines large enough to invoke a much greater selective force are too rare (once every 4-6 generations) and the small increases in mortality in these times are more often due to infectious disease than low body weight and unequally affect the young and elderly (Speakman, 2006; Gibson, 2007; Speakman, 2008). This fact has been supported by Benyshek and Watson (2006), who

used ethnographic reports from 94 forager and agricultural populations to demonstrate that the quantity of available food and frequency and extent of food shortages is not statistically different between preindustrial foragers, recent foragers, and agricultural populations. Thus, selection coefficients during the Neolithic would not have been high enough to result in the current distribution of obesity and T2D.

Recent genetic discoveries have also contributed to the unraveling of the "thrifty genotype" hypothesis. Case-control and GWAS studies have contributed greatly to our knowledge of the genes underlying risk for T2D and obesity (Vander Molen et al., 2005; Dupuis et al., 2010; Saxena et al., 2010; 2012). A natural prediction from the "thrifty-genotype" hypothesis is that these genes should be enriched for selection. More specifically, the haplotypes conferring risk for obesity and T2D should appear to be under recent positive selection. In fact, many genetic loci associated with T2D appear to have been under recent positive selection (Vander Molen et al., 2005; Pickrell et al., 2009). However, for many of these implicated genes, the selected alleles are the protective variants (Vander Molen et al., 2005), the risk variants are not associated with body mass (Helgason et al., 2007), and in some cases the risk variants may only appear selected due to selection on related phenotypes (pleiotropy) (Pickrell et al., 2009).

The "thrifty-genotype" is a hypothesis of evolutionary integration for the genes underlying a complex trait. Yet, upon closer scrutiny it appears that this simple explanation is inadequate. The genetic loci underlying T2D are not integrated in an evolutionary way as we might expect from a set of genes under strong, local selective pressure. Some selection has affected parts of this joint-association network, but the selection has not been uniform in direction. Interestingly, the case for CD closely

parallels that of T2D. Both are genetically complex and have heritable components, both are dominated by a single genetic locus (*TCF7L2* for T2D and HLA for CD), both are environmentally triggered by dietary factors, and both have been hypothesized to be the result of the demographic, subsistence, and settlement shifts associated with the agrarian transformation.

Is there a way to investigate the extent to which recent environments have shaped human evolution? A recent novel approach has been pioneered by Hancock and colleagues (2008; 2010a; b; 2011) and is an unbiased genome-wide method to detect genetic variants that correlate with environmental variables across a large range of human populations. In particular, Hancock's approach is adept at detecting subtle shifts in allele frequency whereas traditional tests of selection based on population differentiation or linkage disequilibrium rely on strong selection and rapid shifts in allele frequency. For example, Hancock and colleagues (Hancock et al., 2010b) found that some of the genomic loci associated with dietary variables (main dietary component of cereals) overlap with previously identified regions of the genome associated with T2D and energy metabolism related traits. These results suggest that some networks underlying complex traits may have evolved in response to recent environmental shifts via subtle increases and decreases in allele frequency. Hancock's results suggest that adaptations to major cultural and environmental changes in the past should produce a signal of an evolutionarily integrated genomic network. Can such integration explain the current distribution of CD risk?

Conclusions about CD

CD is a trait on the border between simple and complex. Many risk loci are known, but one (HLA) accounts for much more of the additive variance than others (Trynka et al., 2011). Given the strong evidence of ongoing balancing selection at the HLA-DQA1 and HLA-DQB1 haplotypes (cited above), HLA may explain the majority of the evolutionary paradox of CD. HLA may behave largely as the sickle-cell model would suggest. In that case, refined balancing selection and variation at class II HLA loci in addition to a few modifier loci may also explain much of the variation in CD risk across Europe, the Near East, and North Africa. The Saharawi of North Africa, for example, have a 5.6% occurrence of CD, in comparison to 1-2% in Southern Europe, despite similar frequencies of the primary risk HLA-DQ haplotype (Catassi et al., 1999; 2001). What explains this difference in risk? In contrast to many other populations of Europe and the Near East, wheat only recently (within the last century) became a staple food among the Saharawi and is now introduced to the diet much earlier (in infancy) (Catassi et al., 2001). An intriguing possibility is that selection on HLA-DQ may be have led to a balance between CD risk and pathogen resistance after the spread of wheat, rye and barley agriculture during the Neolithic. More refined analysis of HLA haplotypes in CD patients from multiple combinations and tests for gene-gene interactions may reveal such a pattern.

Does genetic variation at the non-HLA risk loci similarly reveal a pattern of recent selection in high CD risk areas? In other words, do HLA and non-HLA risk loci represent an evolutionarily integrated network? While balancing selection on HLA may explain a large proportion of CD risk, the evolutionary integration of the network falls apart when we consider the temporal and geographic aspects of selection on the broader

network of non-HLA background risk loci. Some of these loci do appear to have responded to selection on a recent timescale, but most were influenced either by neutral evolution or by selection at earlier times or different places than Europe.

The CD risk joint-association network consists of different functional biochemical pathways (Kumar et al., 2012). Each may have been influenced by different environmental and historical factors, and overall there is no evidence for evolutionary integration among them. The pathogenesis of CD shows that the condition results from the functional and regulatory activity of these loci. Despite this, their connection to CD does not inform us about the selective pressures that affected them in the past in any systematic way.

In closing, we argue that CD is not a side effect of adaptation to an ancient environment. It is a side effect of many unrelated things, apparently in multiple ancient environments, many of which are just random chance. To the extent that other autoimmune conditions share a similar pattern of risk, dominated by HLA class II variation and overlapping greatly with CD risk in many cases (Rioux et al., 2009; Gutierrez-Achury et al., 2011; Ahn et al., 2012; Cifuentes et al., 2012), we should consider the likelihood that the model that we have outlined to explain the paradoxical risk of CD may very well explain the same paradox in other autoimmune diseases.

REFERENCES

Abadie V, Sollid LM, Barreiro LB, and Jabri B. 2011. Integration of Genetic and Immunological Insights into a Model of Celiac Disease Pathogenesis. Annual Review of Immunology 29:493–525.

Ahn R, Ding Y, Murray J, Fasano A, and Green P. 2012. Association Analysis of the Extended MHC Region in Celiac Disease Implicates Multiple Independent Susceptibility Loci. PLoS ONE. 7:e36926.

Albrechtsen A, Moltke I, and Nielsen R. 2010. Natural Selection and the Distribution of Identity-by-Descent in the Human Genome. Genetics 186:295–308.

Allen JS, and Cheer SM. 1996. The non-thrifty genotype. Current Anthropology 37:831–842.

Allison AC. 1954. Protection Afforded by Sickle-cell Trait Against Subtertian Malarial Infection. British Medical Journal 1:290–294.

Alvey C, Anderson CM, and Freeman M. 1957. Wheat gluten and coeliac disease. Archives of Disease in Childhood 32:434–437.

Anderson CM, French JM, Sammons HG, and Frazer AC. 1952. Coeliac disease; gastrointestinal studies and the effect of dietary wheat flour. Lancet 1:836–842.

Armelagos GJ, and Harper KN. 2005. Genomics at the origins of agriculture, part two. Evol Anthropol 14:109–121.

Asimit J, and Zeggini E. 2010. Rare variant association analysis methods for complex traits. Annu. Rev. Genet. 44:293–308.

Barker J, and Liu E. 2008. Celiac Disease: Pathophysiology, Clinical Manifestations, and Associated Autoimmune Conditions. Advances in pediatrics 55:349–365.

Barreiro LB, and Quintana-Murci LIS. 2010. From evolutionary genetics to human immunology: how selection shapes host defence genes. Nat Rev Genet 11:17–30.

Barrett JC, Barrett JC, Clayton DG, Clayton DG, Concannon P, Concannon P, Akolkar B, Akolkar B, Cooper JD, et al. 2009. Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. Nature Genetics 41:703–707.

Beet EA. 1946. Sickle cell disease in the Balovale District of Northern Rhodesia. East African Medical Journal 23:75–86.

Benyshek DC, and Watson JT. 2006. Exploring the thrifty genotype's food-shortage assumptions: A cross-cultural comparison of ethnographic accounts of food security among foraging and agricultural societies. Am J Phys Anthropol 131:120–126.

Bevan S, Popat S, Braegger CP, Busch A, O'Donoghue D, Falth-Magnusson K, Ferguson A, Godkin A, Hogberg L, et al. 1999. Contribution of the MHC region to the familial risk of coeliac disease. Journal of Medical Genetics 36:687–690.

Bodmer W, and Bonilla C. 2008. Common and rare variants in multifactorial susceptibility to common diseases. Nature Genetics 40:695–701.

Botstein D, Cherry JM, Ashburner M, Ball CA, Blake JA, Butler H, Davis AP, Dolinski K, Dwight SS, et al. 2000. Gene Ontology: tool for the unification of biology. Nature

Genetics 25:25–29.

Bouchard C. 2007. The biological predisposition to obesity: beyond the thrifty genotype scenario. Int J Obes Relat Metab Disord 31:1337–1339.

Brenner DA, and Bloomer JR. 1980. The Enzymatic Defect in Variegate Porphyria. N Engl J Med 302:765–769.

Bucci P, Carile F, Sangianantoni A, D'Angiò F, Santarelli A, and Muzio L. 2007. Oral aphthous ulcers and dental enamel defects in children with coeliac disease. Acta Paediatrica 95:203–207.

Campbell BC, and Cajigal A. 2001. Diabetes: energetics, development and human evolution. Medical Hypotheses 57:64–67.

Campisi G, Di Liberto C, Iacono G, Compilato D, Di Prima LL, Calvino F, Di Marco V, Muzio Lo L, Sferrazza C, et al. 2007. Oral pathology in untreated coelic disease. Alimentary pharmacology & therapeutics 26:1529–1536.

Carter R, and Mendis KN. 2002. Evolutionary and historical aspects of the burden of malaria. Clinical Microbiology Reviews 15:564–594.

Catassi C, Doloretta Macis M, Ratsch IM, De Virgiliis S, and Cucca F. 2001. The distribution of DQ genes in the Saharawi population provides only a partial explanation for the high celiac disease prevalence. Tissue Antigens 58:402–406.

Catassi C, Fabiani E, Ratsch IM, Coppa GV, Giorgi PL, Pierdomenico R, Alessandrini S, Iwanejko G, Domenici R, et al. 1996. The coeliac iceberg in Italy. A multicentre antigliadin antibodies screening for coeliac disease in school-age subjects. Acta Paediatrica 85:29–35.

Catassi C, Ratsch I, Gandolfi L, Pratesi R, and Fabiani E. 1999. Why is coeliac disease endemic in the people of the Sahara? The Lancet 354:647–648.

Ch'ng CL, Jones MK, and Kingham JGC. 2007. Celiac Disease and Autoimmune Thyroid Disease. Clinical Medicine & Research 5:184–192.

Chakravarthy MV, and Booth FW. 2004. Eating, exercise, and “thrifty” genotypes: connecting the dots toward an evolutionary understanding of modern chronic diseases. Journal of Applied Physiology 96:3–10.

Cheng J, Malahias T, Brar P, Minaya MT, and Green PHR. 2010. The Association Between Celiac Disease, Dental Enamel Defects, and Aphthous Ulcers in a United States Cohort. Journal of Clinical Gastroenterology 44:191–194.

Cifuentes RA, Restrepo-Montoya D, and Anaya J-M. 2012. The Autoimmune Tautology: An In Silico Approach. Autoimmune Diseases 2012:1–10.

Clément K. 2005. Genetics of human obesity. Proceedings of the Nutrition Society 64:1–10.

Corrao G, Corazza GR, Bagnardi V, Brusco G, and Ciacci C. 2001. Mortality in patients with coeliac disease and their relatives: a cohort study. The Lancet 358:356–361.

Corrao G, Usai P, Galatola G, Ansaldi N, Meini A, Pelli MA, Castellucci G, Corazza GR, and del Tenue TWGOTIC. 1996. Estimating the Incidence of Coeliac Disease with Capture-Recapture Methods within Four Geographic Areas in Italy. Journal of Epidemiology and Community Health (1979-) 50:299–305.

Cortes A, and Brown MA. 2011. Promise and pitfalls of the Immunochip. Arthritis Res Ther 13:101.

Crow JF, and Kimura M. 2009. An Introduction to Population Genetics Theory. Blackburn Press.

Dean G. 1971. The porphyrias. 2nd ed. London: Pitman Medical.

DeMarchi M, Borelli I, Olivetti E, Richiardi P, Wright P, Ansaldi N, Barbera C, and Santini B. 1979. Two HLA-D and DR Alleles are Associated with Coeliac Disease. Tissue Antigens 14:309–316.

Dicke WK, WEIJERS HA, and KAMER JHVD. 1953. Coeliac Disease The Presence in Wheat of a Factor Having a Deleterious Effect in Cases of Coeliac Disease. Acta Paediatrica 42:34–42.

Dicke WK. 1950. Coeliakie. Een onderzoek naar de nadelige invloed van sommige graansoorten op de lijder aan coeliakie.

Dieterich W, Ehnis T, Bauer M, Donner P, Volta U, Riecken EO, and Schuppan D. 1997. Identification of tissue transglutaminase as the autoantigen of celiac disease. Nature medicine 3:797–801.

Dowd B, and Walker-Smith J. 1974. Samuel Gee, Aretaeus, and the coeliac affection. British Medical Journal 2:45.

Dubois PCA, Trynka G, Franke L, Hunt KA, Romanos J, Curtotti A, Zhernakova A, Heap GAR, Ádány R, et al. 2010. Multiple common variants for celiac disease influencing immune gene expression. Nature Genetics 42:295–302.

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

Durbin RM, Altshuler DL, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, Collins FS, La Vega De FM, et al. 2010. A map of human genome variation

from population-scale sequencing. Nature 467:1061–1073.

Eaton SB, Konner M, and Shostak M. 1988. Stone agers in the fast lane: Chronic degenerative diseases in evolutionary perspective. The American journal of medicine 84:739–749.

Falchuk ZM, Rogentine GN, and Strober W. 1972. Predominance of histocompatibility antigen HL-A8 in patients with gluten-sensitive enteropathy. Journal of Clinical Investigation 51:1602–1605.

Falconer DS, and MacKay TFC. 1996. Introduction to Quantitative Genetics. 4th ed. Harlow, England: Longman Group Ltd.

Farrell RJ, and Kelly CP. 2002. Celiac sprue. New England Journal of Medicine 346:180–188.

Fasano A. 2005. Clinical presentation of celiac disease in the pediatric population. Gastroenterology 128:S68–S73.

Festen EAM, Goyette P, Green T, Boucher G, Beauchamp C, Trynka G, Dubois PC, Lagacé C, Stokkers PCF, et al. 2011. A Meta-Analysis of Genome-Wide Association Scans Identifies IL18RAP, PTPN2, TAGAP, and PUS10 As Shared Risk Loci for Crohn's Disease and Celiac Disease. PLoS Genetics 7:e1001283.

Festen EAM, Goyette P, Scott R, Annese V, Zhernakova A, Lian J, Lefèbvre C, Brant SR, Cho JH, et al. 2009. Genetic variants in the region harbouring IL2/IL21 associated with ulcerative colitis. Gut 58:799–804.

Fisher RA. 1918. The correlation between relatives on the supposition of Mendelian inheritance. Transactions of the Royal Society of Edinburgh 52:399–433.

Fleckenstein B, Molberg Ø, Qiao S-W, Schmid DG, Mülbe von der F, Elgstøen K, Jung G, and Sollid LM. 2002. Gliadin T Cell Epitope Selection by Tissue Transglutaminase in Celiac Disease. Journal of Biological Chemistry 277:34109–34116.

Fleckenstein B. 2004. Molecular Characterization of Covalent Complexes between Tissue Transglutaminase and Gliadin Peptides. Journal of Biological Chemistry 279:17607–17616.

Frenette PS, and Atweh GF. 2007. Sickle cell disease: old discoveries, new concepts, and future promise. Journal of Clinical Investigation 117:850–858.

Fu Q, Rudan P, Pääbo S, and Krause J. 2012. Complete Mitochondrial Genomes Reveal Neolithic Expansion into Europe. PLoS ONE 7:e32473.

Gamba C, Fernández E, Tirado M, Deguilloux MF, Pemonge MH, Utrilla P, Edo M, Molist M, Rasteiro R, et al. 2012. Ancient DNA from an Early Neolithic Iberian

population supports a pioneer colonization by first farmers. Mol Ecol 21:45–56.

Gandolfi L, Pratesi R, Cordoba J, Tauil P, Gasparin M, and Catassi C. 2000. Prevalence of celiac disease among blood donors in Brazil. The American Journal of Gastroenterology 95:689–692.

Gibson G. 2007. Human evolution: thrifty genes and the Dairy Queen. Current Biology 17:R295–R296.

Gorlov IP, Gorlova OY, Sunyaev SR, and Spitz MR. 2008. Shifting Paradigm of Association Studies: Value of Rare Single-Nucleotide Polymorphisms. The American Journal of Human Genetics 82:100–112.

Greco D, Pisciotta M, Gambina F, and Maggio F. 2012. Celiac disease in subjects with type 1 diabetes mellitus: a prevalence study in western Sicily (Italy). Endocrine:1–4.

Greco L. 2002. The first large population based twin study of coeliac disease. Gut 50:624–628.

Green P, and Cellier C. 2007. Celiac Disease. New England Journal of Medicine 357:1731–1743.

Gutierrez-Achury J, Coutinho de Almeida R, and Wijmenga C. 2011. Shared genetics in coeliac disease and other immune-mediated diseases. Journal of Internal Medicine 269:591–603.

Haak W, Balanovsky O, Sanchez JJ, Koshel S, Zaporozhchenko V, Adler CJ, Sarkissian Der CSI, Brandt G, Schwarz C, et al. 2010. Ancient DNA from European Early Neolithic Farmers Reveals Their Near Eastern Affinities. Plos Biol 8:e1000536.

Haas SV. 1924. The value of the banana in the treatment of celiac disease. Arch Pediatr Adolesc Med 28:421–437.

Haldane J. 1949. The Rate of Mutation of Human Genes. Hereditas 35:267–273.

Hancock AM, Alkorta-Aranburu G, Witonsky DB, and Di Rienzo A. 2010a. Adaptations to new environments in humans: the role of subtle allele frequency shifts. Philosophical Transactions of the Royal Society B: Biological Sciences 365:2459–2468.

Hancock AM, Witonsky DB, Alkorta-Aranburu G, Beall CM, Gebremedhin A, Sukernik R, Utermann G, Pritchard JK, Coop G, et al. 2011. Adaptations to Climate-Mediated Selective Pressures in Humans. PLoS Genetics 7:e1001375.

Hancock AM, Witonsky DB, Ehler E, Alkorta-Aranburu G, Beall C, Gebremedhin A, Sukernik R, Utermann G, Pritchard J, et al. 2010b. Colloquium Paper: Human adaptations to diet, subsistence, and ecoregion are due to subtle shifts in allele frequency. Proceedings of the National Academy of Sciences 107:8924–8930.

Hancock AM, Witonsky DB, Gordon AS, Eshel G, Pritchard JK, Coop G, and Di Rienzo A. 2008. Adaptations to Climate in Candidate Genes for Common Metabolic Disorders. PLoS Genetics 4:e32.

Hasstedt SJ, Hanis CL, Das SK, and Elbein SC. 2011. Pleiotropy of type 2 diabetes with obesity. Journal of Human Genetics 56:491–495.

Hawks J, Wang ET, Cochran GM, Harpending HC, and Moyzis RK. 2007. Recent acceleration of human adaptive evolution. Proceedings of the National Academy of Sciences 104:20753–20758.

Hedrick PW. 2011. Population genetics of malaria resistance in humans. Heredity 107:283–304.

Helgason A, Pálsson S, Thorleifsson G, Grant S, Emilsson V, Gunnarsdottir S, Adeyemo A, Chen Y, Chen G, et al. 2007. Refining the impact of TCF7L2 gene variants on type 2 diabetes and adaptive evolution. Nature Genetics 39:218–225.

Hunt KA, Zhernakova A, Turner G, Heap GAR, Franke L, Bruinenberg M, Romanos J, Dinesen LC, Ryan AW, et al. 2008. Newly identified genetic risk variants for celiac disease related to the immune response. Nature Genetics 40:395–402.

Hüe S, Mention JJ, Monteiro RC, Zhang SL, and Cellier C. 2004. A Direct Role for NKG2D/MICA Interaction in Villous Atrophy during Celiac Disease. Immunity 21:367–377.

Kagnoff MF. 1984. Possible role for a human adenovirus in the pathogenesis of celiac disease. Journal of Experimental Medicine 160:1544–1557.

Kagnoff MF. 2007. Celiac disease: pathogenesis of a model immunogenetic disease. Journal of Clinical Investigation 117:41–49.

Karell K, Louka AS, Moodie SJ, Ascher H, Clot F, Greco L, Ciclitira PJ, Sollid LM, and Partanen J. 2003. Hla types in celiac disease patients not carrying the DQA1* 05-DQB1* 02(DQ2) heterodimer: results from the european genetics cluster on celiac disease. Human immunology 64:469–477.

Karinen H, Kärkkäinen P, Pihlajamäki J, Janatuinen E, Heikkinen M, Julkunen R, Kosma V-M, Naukkarinen A, and Laakso M. 2006. Gene dose effect of the DQB1*0201allele contributes to severity of coeliac disease. Scand J Gastroenterol 41:191–199.

Keller A, Graefen A, Ball M, Matzas M, Boisguerin V, Maixner F, Leidinger P, Backes C, Khairat R, et al. 2012. New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing. Nat Comms 3:698.

Kilian B, Özkan H, Pozzi C, and Salamini F. 2009. Domestication of the Triticeae in the fertile crescent. In: Feuillet C, Muehlbauer GJ, editors. Genetics and Genomics of the

Triticeae. Vol. 7. London; New York: Springer. p 81–119.

Koning F. 2012. Celiac disease: quantity matters. Semin Immunopathol Online Preprint:1–9.

Kumar V, Wijmenga C, and Withoff S. 2012. From genome-wide association studies to disease mechanisms: celiac disease as a model for autoimmune diseases. Seminars in immunopathology Online Preprint:1–14.

Kwiatkowski DP. 2005. How Malaria Has Affected the Human Genome and What Human Genetics Can Teach Us about Malaria. The American Journal of Human Genetics 77:171–192.

Larsen CS. 2006. The agricultural revolution as environmental catastrophe: Implications for health and lifestyle in the Holocene. Quaternary International 150:12–20.

Lawlor D, Zemmour J, Ennis P, and Parham P. 1990. Evolution of class-I MHC genes and proteins: from natural selection to thymic selection. Annual Review of Immunology 8:23–63.

Lev-Ran A. 1999. Thrifty genotype: How applicable is it to obesity and type 2 diabetes". Diabetes Reviews 7:1–22.

Lev-Ran A. 2001. Human obesity: an evolutionary approach to understanding our bulging waistline. Diabetes Metab. Res. Rev. 17:347–362.

Losowsky MS. 2008. A History of Coeliac Disease. Digestive Diseases 26:112–120.

Lu Q, Wei C, Ye C, Li M, and Elston RC. 2012. A Likelihood Ratio-Based Mann-Whitney Approach Finds Novel Replicable Joint Gene Action for Type 2 Diabetes. Genetic Epidemiology 36:583–593.

Ma L, Brautbar A, Boerwinkle E, Sing CF, Clark AG, and Keinan A. 2012. Knowledge-Driven Analysis Identifies a Gene–Gene Interaction Affecting High-Density Lipoprotein Cholesterol Levels in Multi-Ethnic Populations. PLoS Genetics 8:e1002714.

Maes HHM, Neale MC, and Eaves LJ. 1997. Genetic and environmental factors in relative body weight and human adiposity. Behavior genetics 27:325–351.

Margaritte-Jeannin P, Babron MC, Bourgey M, Louka AS, Clot F, Percopo S, Coto I, Hugot JP, Ascher H, et al. 2004. HLA-DQ relative risks for coeliac disease in European populations: a study of the European Genetics Cluster on Coeliac Disease. Tissue Antigens 63:562–567.

Marsh MN. 1992. Mucosal pathology in gluten sensitivity. In: Marsh MN, editor. Coeliac disease. Oxford, UK: Blackwell Scientific Publishing. p 136–191.

Mäki M, Mustalahti K, and Kokkonen J. 2003. Prevalence of Celiac Disease among Children in Finland. The New England Journal of Medicine 348:2517–2524.

Meissner PN, Dailey TA, Hift RJ, Ziman M, Corrigall AV, Roberts AG, Meissner DM, Kirsch RE, and Dailey HA. 1996. A R59W mutation in human protoporphyrinogen oxidase results in decreased enzyme activity and is prevalent in South Africans with variegate porphyria. Nature Genetics 13:95–97.

Mention JJ, Ben Ahmed M, Bègue B, and Barbe U. 2003. Interleukin 15: a key to disrupted intraepithelial lymphocyte homeostasis and lymphomagenesis in celiac disease. Gastroenterology 125:730–745.

Meresse B, Chen Z, Ciszewski C, Tretiakova M, Bhagat G, Krausz TN, Raulet DH, Lanier LL, Groh V, et al. 2004. Coordinated Induction by IL15 of a TCR-Independent NKG2D Signaling Pathway Converts CTL into Lymphokine-Activated Killer Cells in Celiac Disease. Immunity 21:357–366.

Molberg Ø, McAdam SN, Körner R, Quarsten H, Kristiansen C, Madsen L, Scott H, Norén O, Fugger L, et al. 1998. Tissue transglutaminase selectively modifies gliadin peptides that are recognized by gut-derived T cells in celiac disease. Nature medicine 4:713–717.

Murray JA, Moore SB, van Dyke CT, Lahr BD, Dierkhising RA, Zinsmeister AR, Melton LJ III, Kroning CM, Yousseff El M, et al. 2007. HLA DQ Gene Dosage and Risk and Severity of Celiac Disease. Clinical Gastroenterology and Hepatology 5:1406–1412.

Mustajoki P. 1980. Variegate Porphyria: Twelve Years' Experience in Finland. QJM 49:191–203.

Mutch DM, and Clément K. 2006. Unraveling the Genetics of Human Obesity. PLoS Genetics 2:e188.

Neel J. 1962. Diabetes Mellitus: A "Thrifty" Genotype Rendered Detrimental by "Progress"? American Journal of Human Genetics 14:353.

Nilsen EM, Jahnsen FL, Lundin K, and Johansen FE. 1998. Gluten induces an intestinal cytokine response strongly dominated by interferon gamma in patients with celiac disease. Gastroenterology 115:551–563.

Nistico L. 2006. Concordance, disease progression, and heritability of coeliac disease in Italian twins. Gut 55:803–808.

Palmer N. 2012. A Genome-Wide Association Search for Type 2 Diabetes Genes in African Americans. PLoS ONE 7:1–14.

Petronzelli F, Bonamico M, and Ferrante P. 1997. Genetic contribution of the HLA region to the familial clustering of coeliac disease. Annals of Human Genetics

61:307–317.

Pham Short A, Donaghue KC, Ambler G, Chan AK, and Craig ME. 2012. Coeliac disease in Type 1 diabetes from 1990 to 2009: higher incidence in young children after longer diabetes duration. Diabetic Medicine 29:e286–e289.

Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, Absher D, Srinivasan BS, Barsh GS, Myers RM, et al. 2009. Signals of recent positive selection in a worldwide sample of human populations. Genome Research 19:826–837.

Pinhasi R, and Cramon-Taubadel von N. 2012. A craniometric perspective on the transition to agriculture in Europe. Human Biology 84:45–66.

Prentice AM, Rayco-Solon P, and Moore SE. 2005. Insights from the developing world: thrifty genotypes and thrifty phenotypes. Proceedings of the Nutrition Society 64:153–162.

Prentice AM. 2005. Starvation in humans: Evolutionary background and contemporary implications. Mechanisms of Ageing and Development 126:976–981.

Procaccini M, Campisi G, Bufo P, and Compilato D. 2007. Lack of association between celiac disease and dental enamel hypoplasia in a case-control study from an Italian central region. Head and Face Medicine 3:25.

Rajapakse I, Perlman MD, Martin PJ, Hansen JA, and Kooperberg C. 2012. Multivariate Detection of Gene-Gene Interactions. Genetic Epidemiology 36:622–630.

Rao S, Yuan M, Zuo X, Su W, Zhang F, Huang K, Lin M, and Ding Y. 2011. A Novel Evolution-Based Method for Detecting Gene-Gene Interactions. PLoS ONE 6:e26435.

Ravussin E. 2002. Cellular sensors of feast and famine. Journal of Clinical Investigation 109:1537–1540.

Riddle MS, Murray JA, and Porter CK. 2012. The Incidence and Risk of Celiac Disease in a Healthy US Adult Population. The American Journal of Gastroenterology 107:1248–1255.

Rioux J, Goyette P, Vyse T, Hammarström L, Fernando M, Green T, De Jager P, Foisy S, Wang J, et al. 2009. Mapping of Multiple Susceptibility Variants within the MHC Region for 7 Immune-Mediated Diseases. Proceedings of the National Academy of Sciences of the United States of America 106:18680–18685.

Risch N. 1987. Assessing the role of HLA-linked and unlinked determinants of disease. American Journal of Human Genetics 40:001–014.

Rubio-Tapia A, Ludvigsson JF, Brantner TL, Murray JA, and Everhart JE. 2012. The prevalence of celiac disease in the United States. The American Journal of

Gastroenterology 107:1538–1544.

Sabeti, P.C., Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, et al. 2002. Detecting recent positive selection in the human genome from haplotype structure. Nature 419:832–837.

Saxena R, Elbers CC, Guo Y, Peter I, Gaunt TR, Mega JL, Lanktree MB, Tare A, Castillo BA, et al. 2012. Large-Scale Gene-Centric Meta-Analysis across 39 Studies Identifies Type 2 Diabetes Loci. The American Journal of Human Genetics 90:410–425.

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

Sánchez-Quinto F, Schroeder H, Ramirez O, Avila-Arcos MC, Pybus M, Olalde I, Velazquez A, Marcos MEP, Encinas JMV, et al. 2012. Genomic Affinities of Two 7,000-Year-Old Iberian Hunter-Gatherers. Current Biology 22:R631–R633.

Scott EM, and Grant PJ. 2006. Neel revisited: the adipocyte, seasonality and type 2 diabetes. Diabetologia 49:1462–1466.

Scuteri A, Sanna S, Chen W-M, Uda M, Albai G, Strait J, Najjar S, Nagaraja R, Orrú M, et al. 2007. Genome-Wide Association Scan Shows Genetic Variants in the FTO Gene Are Associated with Obesity-Related Traits. PLoS Genetics 3:e115.

Shan L, Molberg Ø, Parrot I, Hausch F, Filiz F, Gray GM, Sollid LM, and Khosla C. 2002. Structural Basis for Gluten Intolerance in Celiac Sprue. Science, New Series 297:2275–2279.

Simoons FJ. 1981. Celiac disease as a geographic problem. In: Kretchmer N, Walcher DN, editors. Food, nutrition, and evolution: food as an environmental factor in the genesis of human variability. New York, NY: Masson. p 179–199.

Skoglund P, Malmström H, Raghavan M, Storå J, Hall P, Willerslev E, Gilbert MTP, Götherström A, and Jakobsson M. 2012. Origins and Genetic Legacy of Neolithic Farmers and Hunter-Gatherers in Europe. Science 336:466–469.

Sokal RR, Oden NL, and Wilson C. 1991. Genetic evidence for the spread of agriculture in Europe by demic diffusion. Nature 351:143–145.

Solberg OD, Mack SJ, Lancaster AK, Single RM, Tsai Y, Sanchez-Mazas A, and Thomson G. 2008. Balancing selection and heterogeneity across the classical human leukocyte antigen loci: A meta-analytic review of 497 population studies. Human immunology 69:443–464.

Sollid LM, and Thorsby E. 1990. The primary association of celiac disease to a given

HLA-DQ α/β heterodimer explains the divergent HLA-DR associations observed in various Caucasian populations. Tissue Antigens 36:136–137.

Sollid LM. 1989. Evidence for a primary association of celiac disease to a particular HLA-DQ alpha/beta heterodimer. Journal of Experimental Medicine 169:345–350.

Soni S, and Badawy S. 2010. Celiac disease and its effect on human reproduction. The Journal of Reproductive Medicine 55.

Soranzo N, Spector TD, Mangino M, Kühnel B, Rendon A, Teumer A, Willenborg C, Wright B, Chen L, et al. 2009. A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. Nature Genetics 41:1182–1190.

Speakman JR. 2006. Thrifty genes for obesity and the metabolic syndrome—time to call off the search? Diabetes and Vascular Disease Research 3:7–11.

Speakman JR. 2008. Thrifty genes for obesity, an attractive but flawed idea, and an alternative perspective: the "drifty gene" hypothesis. Int J Obes Relat Metab Disord 32:1611–1617.

Spurkland A, Sollid LM, Rønningen KS, Bosnes V, Ek J, Vartdal F, and Thorsby E. 1990. Susceptibility to develop celiac disease is primarily associated with HLA-DQ alleles. Human immunology 29:157–165.

Stene LC, Honeyman MC, Hoffenberg EJ, Haas JE, Sokol RJ, Emery L, Taki I, Norris JM, Erlich HA, et al. 2006. Rotavirus Infection Frequency and Risk of Celiac Disease Autoimmunity in Early Childhood: A Longitudinal Study. The American Journal of Gastroenterology 101:2333–2340.

Stranger BE, Stahl EA, and Raj T. 2011. Progress and Promise of Genome-Wide Association Studies for Human Complex Trait Genetics. Genetics 187:367–383.

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

Takahata N, and Nei M. 1990. Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci. Genetics 124:967–978.

Takahata N, Satta Y, and Klein J. 1992. Polymorphism and balancing selection at major histocompatibility complex loci. Genetics 130:925–938.

Townsend A, and Bodmer H. 1989. Antigen recognition by class I-restricted T lymphocytes. Annual Review of Immunology 7:601–624.

Tresset A, and Vigne JD. 2011. Last hunter-gatherers and first farmers of Europe. Comptes rendus biologies 334:182–189.

Trynka G, Hunt KA, Bockett NA, Romanos J, Mistry V, Szperl A, Bakker SF, Bardella MT, Bhaw-Rosun L, et al. 2011. Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. Nature Genetics 43:1193–1201.

Trynka G, Zhernakova A, Romanos J, Franke L, Hunt KA, Turner G, Bruinenberg M, Heap GA, Platteel M, et al. 2009. Coeliac disease-associated risk variants in TNFAIP3 and REL implicate altered NF- B signalling. Gut 58:1078–1083.

van de Wal Y, Kooy Y, van Veelen P, Peña S, Mearin L, Papadopoulos G, and Koning F. 1998. Cutting Edge: Selective Deamidation by Tissue Transglutaminase Strongly Enhances Gliadin-Specific T Cell Reactivity. The Journal of Immunology 161:1585–1588.

Van Heel DA, Franke L, Hunt KA, Gwilliam R, Zhernakova A, Inouye M, Wapenaar MC, Barnardo MCNM, Bethel G, et al. 2007. A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21. Nature Genetics 39:827–829.

Vander Molen J, Frisse LM, Fullerton SM, Qian Y, Del Bosque-Plata L, Hudson RR, and Di Rienzo A. 2005. Population genetics of CAPN10 and GPR35: implications for the evolution of type 2 diabetes variants. American Journal of Human Genetics 76:548.

Voight BF, Kudaravalli S, Wen X, and Pritchard JK. 2006. A Map of Recent Positive Selection in the Human Genome. Plos Biol 4:e72.

Warnich L, Kotze MJ, Groenewald IM, Groenewald JZ, van Brakel MG, van Heerden CJ, de Villiers JNP, van de Ven WJM, Schoenmakers EFPM, et al. 1996. Identification of three mutations and associated haplotypes in the protoporphyrinogen oxidase gene in South African families with variegate porphyria. Human Molecular Genetics 5:981–984.

Wells JCK. 2007. The evolution of human fatness and susceptibility to obesity: an ethological approach. Biological Reviews 81:183–205.

Whatley SD, Puy H, Morgan RR, Robreau A-M, Roberts AG, Nordmann Y, Elder GH, and Deybach J-C. 1999. Variegate Porphyria in Western Europe: Identification of PPOX Gene Mutations in 104 Families, Extent of Allelic Heterogeneity, and Absence of Correlation between Phenotype and Type of Mutation. The American Journal of Human Genetics 65:984–994.

Willcox G. 2012. The Beginnings of Cereal Cultivation and Domestication in Southwest Asia. In: Potts DT, editor. A Companion to the Archaeology of the Ancient Near East: Volume I. Chichester, West Sussex; Malden, MA: Blackwell Publishing Ltd. p 163–180.

Zeder MA. 2011. The Origins of Agriculture in the Near East. Current Anthropology 52.

Zeegers MP, van Poppel F, Vlietinck R, Spruijt L, and Ostrer H. 2004. Founder mutations among the Dutch. European Journal of Human Genetics 12:591–600.

Zelnik N, Pacht A, Obeid R, and Lerner A. 2004. Range of Neurologic Disorders in Patients With Celiac Disease. Pediatrics 113:1672–1676.

Zhernakova A, Alizadeh BZ, Bevova M, van Leeuwen MA, Coenen MJH, Franke B, Franke L, Posthumus MD, van Heel DA, et al. 2007. Novel association in chromosome 4q27 region with rheumatoid arthritis and confirmation of type 1 diabetes point to a general risk locus for autoimmune diseases. The American Journal of Human Genetics 81:1284–1288.

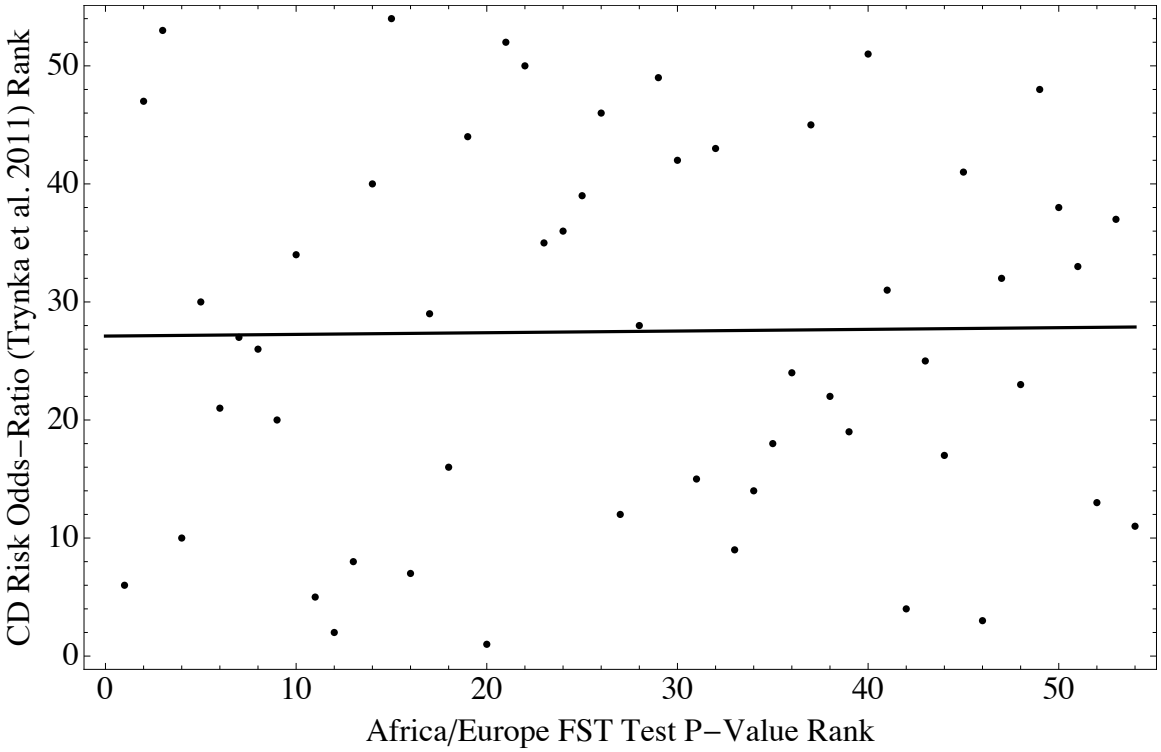
Zhernakova A, Elbers CC, Ferwerda B, Romanos J, Trynka G, Dubois PC, de Kovel CGF, Franke L, Oosting M, et al. 2010. Evolutionary and Functional Analysis of Celiac Risk Loci Reveals SH2B3 as a Protective Factor against Bacterial Infection. The American Journal of Human Genetics 86:970–977.

Zhernakova A, Stahl EA, Trynka G, Raychaudhuri S, Festen EA, Franke L, Westra H-J, Fehrmann RSN, Kurreeman FAS, et al. 2011. Meta-Analysis of Genome-Wide Association Studies in Celiac Disease and Rheumatoid Arthritis Identifies Fourteen Non-HLA Shared Loci. PLoS Genetics 7:e1002004.

Zuk O, Hechter E, Sunyaev SR, and Lander ES. 2012. The mystery of missing heritability: Genetic interactions create phantom heritability. Proceedings of the National Academy of Sciences of the United States of America 109:1193–1198.

Figure 1. Africa/Europe CD risk odds-ratio/ F_{ST} test p-value rank order correlation

Rank order correlation between Africa/Europe genome-wide F_{ST} test p-values and CD risk locus odds-ratios reported by Trynka and colleagues (2011) demonstrating the lack of correlation between these two variables. This pattern suggests that CD loci with larger effect-sizes have not been systematically subject to greater differentiation. The correlation coefficient is $r^2 = 0.268$.



CHAPTER IV

An Ancient Genome Reveals Biases in Allele Age Estimation From Measures of Intra-Allelic Variation

Aaron Sams and John Hawks

ABSTRACT

Estimates of allele age and timing of selection using intra-allelic variation in present human DNA depend heavily on assumptions about the demographic history of human populations and are often associated with large ranges of error. Ancient human DNA can be used to test hypotheses about the timing of selective changes in Holocene human populations. Ancient DNA can be used to demonstrate the presence/absence of specific genetic variants within a much smaller time frame, depending on the ability to accurately date ancient skeletal materials. In this study we utilize a method of moments estimator to estimate allele ages in a sample of Europeans. We then use the genome of the 5,300 year old Tyrolean Iceman, Ötzi, to determine how often such estimates based on levels of intra-allelic variation and recombination are incorrect. We find that sampling bias alone cannot account for the presence of seemingly young alleles in Ötzi's genome.

Key words: allele age, ancient DNA, Europe, demography, natural selection

INTRODUCTION

Over-representation of young alleles in present populations

Genome-wide comparisons during the past decade have revealed a surplus of linkage disequilibrium (LD) in some genomic regions in human populations compared to neutral expectations (Serre et al., 1990; 2002; Kim and Nielsen, 2004; Conrad et al., 2006; Voight et al., 2006; Wang et al., 2006; Hawks et al., 2007; Sabeti et al., 2007; Barreiro et al., 2008; Kelley et al., 2008; Pickrell et al., 2009; Grossman et al., 2010). LD has been applied as a test for recent natural selection, as selection on new alleles causes frequency increases that are faster than local recombination rates between the selected variant and nearby alleles. This produces long-range haplotypes that are found at disproportionately high frequency compared to other haplotypes of the same length. Over time, these long haplotypes are gradually degraded by recombination, so their continued presence and high frequencies in human populations are inconsistent with a very ancient origin for them. Instead, these long-range haplotypes have been argued to be the product of recent positive selection, mostly within the last 30,000 years.

In one of the earliest instances of an LD based test for neutrality, Sabeti and colleagues (2002) developed the extended haplotype homozygosity (EHH) statistic, which they define as the probability that two randomly chosen chromosomes carrying the core haplotype of interest are identical by descent. Sabeti and colleagues (2002) used EHH to demonstrate evidence of selection on the G6PD and TNFSF5 genes in Africa for malaria protection. Later, Voight and others (2006) further developed the EHH statistic into the integrated haplotype score (iHS) which is standardized by the genome-wide empirical distribution of haplotype lengths and frequencies. The researchers identified an

over-representation of young (long haplotype) moderate frequency haplotypes in European (CEU), Asian (CHB + JPT), and African (YRI) samples from the HapMap project (Altshuler et al., 2005). Similarly, Wang and colleagues (Wang et al., 2006), using another long-range haplotype method called linkage disequilibrium decay (LDD), found that roughly 1800 human genes show evidence for recent positive selection.

At any single locus, the effects of demography, selection, and stochastic fluctuations are difficult to disentangle. Nonequilibrium demographies, such as those affected by recent bottlenecks, have been shown to have inflated haplotype-based test statistics (Voight et al., 2006; Macpherson et al., 2008). Population bottlenecks produce common haplotypes that are extremely long, even in neutral regions. Nonetheless, extensive simulations in several genome-wide surveys of haplotype length (Voight et al., 2006; Hawks et al., 2007) have verified that the abundance of young alleles in human populations, particularly in Africa and Europe, are genuine. Furthermore, the high degree of adaptive evolution (and corresponding excess of apparently young alleles) observed in present human populations is consistent with explosive demographic growth during the past 40,000 years and major changes in ecology and subsistence associated with the Neolithic transition to agriculture (Hawks et al., 2007; Keinan and Clark, 2012; Fu et al., 2012b).

Although the surplus of long-range haplotypes across the genome does require a non-neutral explanation, for any single locus it can be difficult to substantiate selection. The coalescent process within a population under neutrality has a very high variance, recombination occurs as a stochastic process, and the population history of growth and replacement in recent humans is not known with very high fidelity. As a result, neutral

evolution has a long tail that overlaps with the distribution of LD expected for a positively selected allele. In these cases, ancient genomes provide an excellent avenue to differentiate between adaptive, stochastic, and demographic explanations for such haplotypes.

Presence of recently selected alleles in ancient European specimens

A straightforward means of testing hypotheses about long-range haplotype data in present populations is to analyze putatively selected haplotypes and single-nucleotide polymorphisms (SNPs) in ancient human specimens. To date, the most successful application of ancient DNA to a potential case of recent natural selection pertains to the genetic variants underlying lactase persistence. The ability to digest lactose, a disaccharide found in milk, slows around the time of weaning in most human populations, as in nearly all mammals. But many humans retain high production of lactase, which digests lactose into simple sugars, into adulthood. This phenotype is known as lactase persistence. Lactase persistence is not solely the result of a single genetic change within individuals, but mutations affecting the regulation of *LCT* are by far the most important influences on lactase persistence status. Several such mutations have reached high frequencies in human populations; the first of these to be characterized in terms of time and geographic distribution was the -13910*T allele in Europeans (Bersaglieri et al. 2004). The genomic region surrounding the LCT gene has been consistently reported as one of the most extreme long-range haplotype based examples of recent selection in Europe (Enattah et al., 2002; Bersaglieri et al., 2004; Tishkoff et al., 2006; Enattah et al., 2007; 2008). A critical discovery in the case of lactase is that the

LCT gene itself is not responsible for the phenotypic shift, rather intronic changes within the nearby *MCM6* gene reverse the cessation of lactase production after weaning that occurs in most mammals (Enattah et al., 2002). Similar disruptive changes to the MCM6 gene have been selected convergently in both African (Tishkoff et al. 2007) and Middle Eastern (Enattah et al. 2008) populations. In particular, the lactase example shows the relevance of comparing genetic data to known cultural practices in the past, such as the timing and geographic distribution of domesticated cattle and milk consumption(Holden and Mace, 1997; Gerbault et al., 2009).

The timing of the mutation and subsequent selective sweep underlying lactase persistence in Europeans (C/T-13910) is typically estimated (based on LD) between 3000 and 12000 years, which coincides with the presence of domesticated cattle (Bollongino et al., 2006) and a record of increasing pastoralism and dairying in several human populations, including Northern Europe. Tishkoff and colleagues (Tishkoff et al., 2006) estimated the age of the selective sweep using a coalescent simulation model that incorporated selection and recombination. Their estimated age of the selective sweep affecting the -13910*T European lactase persistence allele was between approximately 8,000 to 9,000 years depending on the degree of dominance assumed for the allele. Importantly, the confidence intervals around these estimates ranged from approximately 2,000 to 19,000 years. This wide range consistent with the large range of variation in coalescence times expected for a positively selected allele (Slatkin and Rannala, 2000).

Estimates of allele age and timing of selection using present human DNA depend heavily on assumptions about the demographic history of human populations and are often associated with large ranges of error (as illustrated above). Ancient human DNA

can be used to test hypotheses about the timing of selective changes in Holocene human populations (Burger et al., 2007; Malmström et al., 2010; Plantinga et al., 2012). Ancient DNA can be used to demonstrate the presence/absence of specific genetic variants within a much smaller time frame, depending on the ability to accurately date ancient skeletal materials. The derived allele (-13910*T) that has been shown to correspond completely to lactase persistence in Northern Europeans (Kuokkanen et al., 2003) was found in only one copy out of twenty (~5%) in a 5000 year old sample from Sweden (Malmström et al., 2010), was found at a frequency of approximately 27% in a sample of 26 Basque individuals dating between 4500 and 5000 years ago (Plantinga et al., 2012), and was completely absent from a smaller sample from Eastern Europe dating between 5000 and 5800 years ago (Burger et al., 2007).

Holocene demography of Europe and ancient DNA

Archaeological evidence suggests that the transition from a hunting and gathering lifestyle to a more sedentary agricultural "Neolithic" lifestyle, which began in the Near East by 10,000 years ago, spread across Europe between 8,000 and 4,000 years ago (Price, 2000). Archaeological and genetic evidence has traditionally been divided between two viewpoints regarding the spread of agricultural lifestyles from the Near East. The earlier and more popular viewpoint argues that this transition is characterized by a large amount of *demic diffusion* (Childe, 1958). In other words, the transition to agriculture in Europe involved a large influx of agricultural populations from the Near East, for which we would expect to see a distinct genetic signature. Others view this transition as being dominated by *cultural diffusion*, such that Mesolithic hunter-gatherers

in Europe embraced Neolithic lifestyles with little genetic input from Near Eastern farmers (Zvelebil and Dolukhanov, 1991).

Genetic arguments in favor of demic diffusion have persisted for several decades (Sokal and Menozzi, 1982; Ammerman and Cavalli-Sforza, 1984). For example, Sokal and colleagues (1982) used a set of allele frequencies from 21 loci drawn from the HLA-A and HLA-B genes from 58 populations in Europe in a spatial autocorrelation analysis and found support for a model of recent demic diffusion from the Near East. Nearly a decade later, Sokal and others (1991) extended their comparison to a set of loci drawn from 26 genetic systems and 3,373 geographic locations in Europe and found that 6 of 26 systems are consistent with a model of demic diffusion.

More recently, samples of present day mitochondrial (mtDNA), Y-chromosome, and autosomal DNA have been extensively applied to the problem of Neolithic origins, often with contradictory results. European populations demonstrate remarkable homogeneity, particularly in mtDNA (Puit et al., 1994; Simoni et al., 2000). However, the demographic event that generated the majority of this signal (either the initial Upper Paleolithic founding of Europe ~45,000 years ago or the Neolithic settlement of Europe ~10,000 years ago) is unclear. Additionally, while most geneticists agree that Near Eastern Neolithic populations contributed substantially to the ancestry of present Europeans, the estimated size of this contribution varies depending on the type of data used. Richards and colleagues (2000) implemented a phylogeographic founder analysis on a large sample of mtDNA from Europe and the Near East and estimated that Neolithic populations of the Near East contributed approximately 20% to the present day genetic diversity of Europe. The R1b1b2 Y-chromosome haplogroup is the most common

among present-day European men. This haplogroup ranges in frequency from ~12% Eastern Turkey to ~85% in Ireland. This frequency distribution led early studies of this haplogroup to conclude a pre-Neolithic origin for this haplogroup (Rosser et al., 2000; Semino et al., 2000). However, a more recent analysis of this haplogroup (Balaresque et al., 2010) concluded that the geographic diversity within this haplogroup is more consistent with a recent spread from the Near East and that as much as ~80% of present-day European Y-chromosome lineage were introduced to Europe during the Neolithic. Additionally, the Neolithic Near Eastern contribution to present European genomes was recently estimated to be between 46% and 66% using autosomal microsatellite variation (Belle et al., 2006). Thus, estimations of the relative contributions of Paleolithic European and Near Eastern Neolithic populations to the genomic diversity of present Europeans vary widely depending on the type of genetic data analyzed. It is important to note that these estimates based solely on phylogeographic analysis of genetic variation in living humans can be biased by several factors. Phylogeographic analyses of Europe and the Near East depend heavily on clines of genetic variation. When considering a limited number of genetic loci, these clines can be generated by stochastic processes such as isolation by distance (Novembre and Stephens, 2008). Additionally, phylogeographic methods can be heavily biased by recent demographic events, such as gene-flow and back-migration between Europe and the Near East (Richards et al., 2000; Balaresque et al., 2010).

Ancient DNA may circumvent these issues and help to settle this debate because genetic differences between skeletal samples associated with Mesolithic and Neolithic technologies can be directly observed. A model of cultural diffusion predicts little to no

major genetic differences associated with culture change in Holocene samples while the demic diffusion model predicts substantial influx of novel genetic variation. Currently, the majority of ancient DNA studies of Europe involve mtDNA analysis, primarily from hunter-gatherer and farming populations of north and central Europe. Studies performed to date (Haak et al., 2005; Bramanti, 2008; Haak et al., 2008; Bramanti et al., 2009; Haak et al., 2010; Guba et al., 2011; Lee et al., 2012; Nikitin et al., 2012; Fu et al., 2012a) suggest that approximately ~83% (19/23) of pre-Neolithic peoples of Europe carried mtDNA haplogroup U and none belong to haplogroup H, a composition that is very different from present samples in which haplogroup H is dominant. However, ~12% (13/105) of mtDNAs from early farming populations contain haplogroup U (which is similarly rare in modern Europeans), while haplogroup H is present in between 25 and 37% of mtDNAs from early farming samples (similar to the 30% in living Europeans). Overall, these results from ancient mtDNA analysis suggest that pre-Neolithic hunter-gatherers contributed ~20% at most to the mtDNA genetic composition of present European populations (Fu et al., 2012a). These ancient mtDNA results were recently combined with a dataset of 1,151 complete mtDNAs from across Europe (Fu et al., 2012a). Fu and colleagues (2012a) found evidence for a population expansion between 15,000 and 10,000 years ago in mtDNAs typical of pre-Neolithic hunter-gatherers and a subsequent contraction of these haplotypes between 10,000 and 5,000 years ago, consistent with the expansion of mtDNAs from agricultural populations expanding from the Near East.

In addition to ancient mtDNA data, results from whole genome sequencing of ancient Holocene aged individuals have been applied to the question of Neolithic population

replacement in Europe. Within the past year, substantial amounts of autosomal DNA have been reported for seven ancient individuals. These include Ötzi (the Tyrolean Iceman) dating to roughly 5,300 years ago (Keller et al., 2012), three hunter-gatherers associated with the Pitted Ware culture and one farmer associated with the Funnel Beaker culture from Scandinavia dating to approximately 5,000 years ago (Skoglund et al., 2012) and two 7,000 year old Iberian hunter-gatherers (Sánchez-Quinto et al., 2012). In each case, the amount of autosomal DNA allows for comparison of the average "genetic signature" of the ancient specimens to present-day samples. Genome-wide analysis of the Iceman's genome revealed that the present day people most closely related to that individual are Sardinians. Keller and colleagues (2012) have argued that this relationship may reflect recent common ancestry between the Sardinian and Alpine populations of ~5,000 years ago. The study of the four ancient Scandinavians surprisingly revealed that the hunter-gatherers of this sample are more similar to northern Europeans while the farmer in this sample is more genetically similar to present southern Europeans. While the ancient hunter-gatherers cluster most closely to present N. Europeans, it is important to note that their genetic signature is unique and does not precisely match any present populations. In contrast, the individual associated with the Funnel Beaker farming culture shows a high degree of genetic similarity to present-day individuals in Greece and Cyprus. Finally, analysis of the two Iberian individuals showed no close relationship between these individuals and present-day populations in Iberia or southern Europe. The general picture emerging from analysis of ancient DNA is that the spread of farming technology during the Neolithic was associated with significant population movements from the Near East across Europe. Given the likely large-scale replacement of pre-

Neolithic European genetic variation with that of migrating agriculturalists from the Near East it is important to consider how we apply data from ancient DNA to questions about recent natural selection in Europe.

Ancient DNA, demography, and selection

The results from genome-wide samples of ancient autosomal DNA largely support results from mtDNA in suggesting that the Neolithic of Europe was characterized by mass migrations and expansions of farming populations. In light of these demographic implications, should we question how we interpret the application of ancient DNA to questions of recent natural selection at single genetic loci? Consider the example of the lactase persistence marker described above. As a hypothetical example we can consider the analysis by Malmström and colleagues (2010), which revealed that the -13910*T lactase persistence allele was relatively rare (~5%) in a middle Neolithic sample of hunter-gatherers associated with the Pitted Ware culture. This result would seemingly strengthen an argument for the recent origin and rapid increase in frequency of the LP allele in populations of Europe. However, given that these mid-Neolithic hunter-gatherer populations may not be closely related to present day populations of northern Europe, the rapid shift in allele frequency at the single LP locus may represent the demographic shift rather than a rapid shift in allele frequency due to positive natural selection. In light of this, it will be important in future applications of ancient DNA to questions related to selection on individual loci to consider a wider geographic range of ancient DNA samples. For example, we do not know what the frequency of the -13910*T allele was in ancient populations of the Near East.

Estimating allele age

Given the uncertainties associated with using ancient DNA to assess recent selection we would like to be able to determine how closely estimates of the timing of selection actually represent reality. One way to utilize ancient DNA to approach this question would be to ask if sites that deviate significantly from neutral expectations (for example with extreme EHH or iHS scores) are underrepresented in ancient genomes compared to other sites across the genome at similar frequency. However, as discussed above, tests of neutrality are weakened by our lack of precise knowledge about demographic history. Additionally, tests of neutrality are weak at detecting selection at sites presently at low frequency (mostly sites that have arisen recently).

Another approach, which we have decided to employ, is to assess whether age estimates for alleles in present-day samples are consistent with evidence from ancient DNA samples. This approach does not depend on finding evidence for selection (or disproving neutrality) for any of the alleles in present populations. Both neutral and selected alleles have coalescence times, which represent the last time that the present-day copies of the allele shared a common ancestor. This coalescence time estimate is commonly termed “allele age”, or “mutation age”, although strictly speaking the coalescence time of present copies of an allele may represent a time long after the mutation originally occurred. In general, we expect that coalescence time estimates for alleles shared in an ancient genome will be older than the date of the ancient individual. For example, if we estimate that the coalescence time of an allele in present samples is 3000 years, and we discover that the allele is present in the genome of a 5,300-year-old

individual, these observations apparently contradict each other. Such mismatches between ancient DNA evidence and coalescence time estimates from samples of living people give us a way to assess the factors that may bias or contribute to error in allele ages, such as sampling variance, variance due to evolutionary stochasticity, or errors in the recombination map.

One of the earliest methods developed to estimate the age of an allele required only allele frequency to predict age. Kimura and Ohta (1973) used diffusion theory to derive the expected age and variance of the age of a neutral allele at a given frequency. For a given neutral allele at frequency p in a population of fixed effective size N the expected age (in units of $2N$ generations) is approximately (Kimura and Ohta, 1973; Slatkin and Rannala, 2000):

$$E(t_1) = \frac{-2p}{1-p}\ln(p)$$

Later, application of the coalescent process to the work of Kimura and Ohta produced formulae to estimate the probability distributions associated with allele age estimations based on frequency (Griffiths and Tavaré, 1998). The major revelation of this work, predicted by the theoretical work of Kimura and Ohta, was that age estimates of neutral alleles have extremely wide confidence intervals. This is particularly true for an allele at low frequency because it may be at low frequency because it arose recently, or because it arose long ago and is in decline from a previously higher frequency (Kimura and Ohta, 1973; Griffiths and Tavaré, 1998; 1999; Slatkin and Rannala, 2000). The wide error bars associated with such estimations make them unsuitable for our purposes.

In the past few decades, methods have been developed to estimate allele age based on the degree of intra-allelic variability between two variable sites (Serre et al., 1990; Risch

et al., 1995; Slatkin and Rannala, 2000). Consider a di-allelic site with alleles *A* and *a*. Also consider a nearby position that is fixed (*b*). If a mutation occurs at this nearby position on a chromosome carrying the *A* allele and we call the new allele *B* then initially the two sites will be in perfect LD (*AB/ab*). Over time, recombination events will shuffle the haplotypes. The theory of recombination, therefore, allows us to predict the time since the initial mutation occurred based on the proportions of haplotypes:

$$t = \frac{1}{\ln(1-c)} \ln\left(\frac{x_t - y}{1 - y}\right),$$

where t = allele age (in generations), c = recombination rate between two loci, x_t = proportion of the derived chromosomes carrying the linked allele, y = proportion of the ancestral chromosomes carrying the linked allele. This method is a method-of-moments estimator. Conveniently, it requires no population genetic or demographic assumptions, only the exponential decay of initially perfect LD because of recombination (Slatkin and Rannala, 2000).

When addressing the timing of recently selected alleles in the absence of ancient DNA, we would like to know if our estimates of the timing of selection are significantly biased. Ancient DNA provides an avenue to examine this problem. Consider an individual, like Ötzi, the Tyrolean Iceman, assumed to be approximately 5,000 years old. What proportion of alleles that we estimate to be younger than 5,000 years old using present DNA are ascertained in such an individual? The answer to this question is theoretically important because it provides us with an idea of how commonly our estimates of the timing of selection and age of mutations can be biased by demographic factors (such as population replacements, bottlenecks, and population growth), errors and randomness associated with the coalescent, linkage maps and recombination rates, and

other factors. To address this question we have used the simple estimator of allele age described above (and more fully in methods and in (Slatkin and Rannala, 2000)) and asked what proportion of derived alleles across the genome that are estimated to be younger than 5,000 years old are ascertained in the genome of Ötzi the Iceman(Keller et al., 2012). We estimated allele ages using the Tuscani (TSI) sample from phase I of the 1000 Genomes Project (see Methods) (Durbin et al., 2010).

RESULTS

General age distribution

The full distribution of ages (in generations) produced from our estimation procedure is displayed graphically in Figure 1. These ages range from 0 to approximately 13,000 generations. The full distribution of allele ages in a population should approximate an exponential decay curve. However, the HapMap sites used in this study are subject to an ascertainment bias, such that they are more common than a truly random genome-wide sample of SNPs. This bias results in a paucity of young alleles in our sample. For our purposes, the lack of these low frequency alleles still leaves a substantial fraction of ages below 200 generations (approximately 5,000 years given 25 year generations). Therefore, we can determine whether or not there is an excess of alleles that, given their estimated age, should not be present in the Iceman genome.

Excess of young alleles present in Iceman

If our allele age estimates were exact, we should expect to find no alleles younger than 200 generation in the Iceman genome. In reality allele age estimations are subject to

substantial variation and bias from small sample size. To assess the degree of sampling bias in the full sample of age estimates we divided the sample into four classes based on two categories (present/absent in Ötzi, less/greater than 200 generations). We observe in the full set of allele ages a significant ($p < 1 \times 10^{-100}$) excess of sites that are present in Ötzi with young (< 200 generations) estimated ages (Table 1). Exact age estimates should produce no alleles in this class. Therefore, we must explain the presence of these young alleles in Ötzi's genome. Our primary question is, are they the result of sampling error due to our small sample (196 haploid genomes) or other sources of error (variance in coalescence and/or recombination or demographic factors).

Examination of source of age estimation error

Error associated with sample size can substantially bias our allele age estimates in certain parts of parameter space. To address this issue we modeled the exact sampling probabilities and age estimates for every possible combination of allele count between two sites given a true allele age, recombination rate, allele frequency, and *y* (proportion of ancestral haplotypes at focal site carrying the linked allele) (see Methods). After exploring a wide range of sample space, we found that the greatest degree of sampling bias occurs under two conditions, when the focal allele frequency is extreme and when *y* is relatively high (Figure2).

To investigate the effect of this sampling bias we filtered the dataset to include only sites between 5% and 50% frequency, a range in which the two sources of sampling error mentioned above is minimized (Figures 1 and 2). This corrected set of ages has a slight but insignificant (chi-square test, p=0.749) excess fraction of sites that are absent

from Ötzi with young (< 200 generations) estimated ages (Table 1). The presence of young alleles in Ötzi after this filtering process suggests that bias on estimates of allele age from sources other than sampling is present.

To further demonstrate that sources of bias other than sampling contribute substantial error to our estimates we took the full distribution of allele age estimates and filtered it based on whether the linked allele frequency was moderately high (75-80%) or moderately low (20-25%) (which correlates with high and low values of *y*, respectively). If our prediction that high values of *y* are prone to bias allele age estimates is true (Figure 3), then our chi-square tests on the presence/absence of young alleles in Ötzi should be more strongly significant (more mis-classified ages) in the high-frequency case. This result is true. In both cases there is a significant excess of young alleles present in Ötzi. However, in the high frequency case, a much higher proportion of all present alleles are estimated to be younger than 200 generations (Table 2). This result is exactly opposite of that which would be predicted from allele frequency alone (older alleles are more often found at higher frequencies). Nonetheless, the over-representation of young alleles present in Ötzi in the lower frequency class suggests that even with optimal parameters for estimation, estimated ages are biased by sources other than sampling.

DISCUSSION

What explains the presence of alleles estimated to be younger than 5,000 years old in Ötzi's genome? There are two general explanations, one related to the process of estimation from present-day samples, and one rooted in population history.

Intra-allelic variation and the recombination map

Coalescence time estimates based on linkage disequilibrium depend both on the variation surrounding a focal site and the recombination distance between focal site and linked sites used in age estimation. The amount of intra-allelic variation surrounding a focal site can be biased by sample size. As sample sizes increase we should expect accuracy of our age estimations to increase. We estimate ages here based on data from the Tuscani (TSI) sample from the 1000 Genomes Project (Durbin et al., 2010) numbering 196 phased haploid genomes (see methods). While this sample is substantial, the restricted number of haplotypes means that for some loci, haplotype variation that is present in the true population (Europe) is missed by chance, thereby reducing allele age estimates. Our results show that the standard error of coalescence time estimates reduces markedly in larger samples. This factor alone is not expected to badly bias estimates of coalescence time, but it does lead to an increased dispersion of estimates, particularly for rare focal alleles.

Recombination distances between SNPs used in this study were derived from standard genetic maps, which provide the probability of recombination between any two sites on a chromosome. These probabilities are not exact and are themselves subject to biases from sample size and estimation. For this study we utilized the most recent (Jan 2011) release of the HapMap recombination map (hapmap.org). This map represents an average of three samples (CEU, YRI, CHB+JPT) from the second generation haplotype map (Frazer et al., 2007). Rates are estimated from phased haplotypes from these samples. In addition to being averaged across diverse populations, it is also sex-averaged. Much recent work has focused on variation in recombination rates both

between populations (Hinch et al., 2011; Wegmann et al., 2011) and between individuals within a population, particularly between sexes (Fledel-Alon et al., 2011). The appreciable variation in recombination rates across the genome, across populations, and across individuals means that average maps estimated from relatively small samples and averaged across populations and sexes are likely to introduce biases to our age estimations in an unpredictable manner. More basically, the recombination distance between two loci predicts the average of a stochastic process with substantial variance, which itself can bias point estimates of allele ages.

Population history

The importance of using an ancient genome in this context is that if our coalescence time estimates were absolutely accurate point estimates of mutation times, our model of population history would not matter. Whatever we assume about population history, we should never observe alleles that arose in the past 5,000 years in a 5,300 year-old genome.

However, coalescence times for linked haplotypes are not mutation times. In a population that fluctuates to a small effective size, it is quite possible for a mutation to repeatedly become rare enough for all but a single lineage to disappear. The coalescence time of such an allele may be substantially younger than the time the mutation appeared in the population.

The impact of population history in such cases is complex. If an allele was very rare in an ancient population, it should be unlikely to sample it in an ancient genome from that population. Nevertheless, in a population with small effective size, it becomes

relatively more likely for an allele to go from being relatively common to rare, and later recover to a higher frequency. For alleles segregating neutrally, this can happen several times during the evolution of an allele (Kimura and Ohta, 1973). Moreover, if populations were subdivided in the past, our ancient specimen may sample a subpopulation in which the allele was more common, while present-day populations may derive more of their ancestry from different subpopulations where the allele was rare. Hence, under a particular model of population history, a recent coalescence time may not be inconsistent with the presence of an allele in an ancient genome. Therefore, for some proportion of alleles, we should expect our age estimations to be biased towards younger ages. This phenomenon creates the bottleneck problem associated with tests of recent selection (Conrad et al., 2006; Macpherson et al., 2008; Pickrell et al., 2009), in which haplotype based test statistics are inflated due to recent population bottlenecks. If, as described above, Europe experienced one or more founding events from Neolithic Near Eastern populations, we may expect age estimates in European populations to be biased towards younger ages.

Impact of results on hypotheses of recent natural selection

Our results imply that various factors have biased estimates of allele age taken from relatively small (n<200) samples of human populations. Genome-wide tests of neutrality are based on the same logic as the simple method-of-moments allele age estimator that we utilize here. Alleles that have relatively high frequencies given their low degree of intra-allelic variability (haplotype diversity) are inferred to result from the action of recent natural selection. Our observation of seemingly young alleles in the 5,300 year-

old Ötzi genome implies that for a significant proportion of alleles, the age component of tests of neutrality (from haplotype variation) will be biased towards younger ages. The end result of this bias is to make some alleles appear to be much more divergent from neutral predictions than we expect. In other words, some alleles that were not recently under positive selection will appear as if they were. Other alleles may genuinely have been affected by recent selection, but the true timescale of selection on them may have been substantially more ancient than inferred from the degree of intra-allelic variation in a sample.

For example, an allele (rs3184504*A) in the *SH2B3* gene on chromosome 12 has been identified in multiple scans of recent selection (Barreiro and Quintana-Murci, 2010; Zhernakova et al., 2010; Abadie et al., 2011) and is thought to play a functional role in the immune system (Zhernakova et al., 2010). This allele was estimated to have been affected by a selective sweep in Europe between 1,200 and 1,700 years ago (Zhernakova et al., 2010). However, we have previously reported (Chapter 2 & 3) that this allele is present in Ötzi's genome. Our current results suggest that our observation of this allele in the Iceman's genome is plausibly not due to error from contamination or sequencing. Rather, many allele ages estimated from intra-allelic variation are younger than they can plausibly be. Future investigation of recent selection must increasingly consider these sources of bias. It remains unclear at present if other estimators of allele age based on intra-allelic variation, such as the maximum likelihood estimator (Slatkin and Rannala, 2000), will demonstrate a similar bias.

Not only positive selection but also neutral and deleterious variants are subject to these biases. Many of the alleles underlying common diseases in present populations are

likely to be rare. Our results strongly suggest that moving forward, improvements to estimated ages of alleles based on present human DNA will require substantially larger samples to alleviate the problems associated with sampling bias. For very rare alleles (with frequencies<0.01), sample sizes must be as large as 10,000 haplotypes in order to capture enough copies of the rare allele to gain useful estimates. Additionally, and perhaps more importantly, larger samples of ancient DNA must be increasingly integrated with results from present human genetic variation to understand the demographic, stochastic, and deterministic forces that have shaped present-day human genetic variation.

METHODS

1000 Genomes samples

We obtained variant calls for single nucleotide polymorphisms from the June 2011 data release of the 1000 Genomes Project (http://1000genomes.org). For our comparisons, we considered only the autosomes, chromosomes 1-22, and excluded the X chromosome from consideration. This sample comprises 1094 individuals from Utah (CEU), Yoruba (YRI), Luhya (LWK), British (GBR), Spanish (IBS), Tuscan (TSI), Finn (FIN) North and South ethnic Han Chinese (CHB, CHS), Puerto Rican (PUR), African-Americans (ASW).

Because each individual is diploid for the autosomes, the 1094 individuals are the equivalent of 2188 haploid genomes. The publicly available SNP dataset reports a diploid genotype for each individual imputed from the relatively low-coverage (4x) sequencing. For most of our comparisons, we considered individuals as diploid autosomes, counting each copy of an allele. Our allele age estimations required us to consider only a single

copy for each SNP locus, treating each individual as two haploid-equivalent genomes. We used the ordering in the original data files to assign alleles to haploid-equivalent genomes in these cases. For our final analysis we chose to focus only on individuals from the TSI samples.

The generation of the SNP genotypes in the low-coverage 1000 Genomes data release involved imputation of genotypes in many cases. Genotype imputation requires assumptions about population history, and relies on information from a small number of high-coverage genomes that may introduce ascertainment bias. We leave this as a potential shortfall of our results, which can be addressed in future work by high coverage genomes.

We also obtained genome sequence alignments for chimpanzee (PanTro2), orangutan (PonAbe1) and macaque to establish the polarity of the SNP alleles in humans. In most cases, we took the chimpanzee allele as the ancestral state for humans. In cases where the chimpanzee genome data did not represent the SNP (whether missing or not aligned) we took the orangutan allele as ancestral; if the orangutan was missing we took the macaque. In the cases where no alignment exists in other primates for a SNP, we removed that SNP from our analysis. It is worth noting that in a small fraction of cases polarity may be interpreted incorrectly, which will also lead to incorrect allele age estimations.

HapMap recombination map

We obtained the most recent (Jan 2011) HapMap recombination map from HapMap bulk data download site (http://hapmap.ncbi.nlm.nih.gov). We filtered the 1000 Genomes data to include only the sites present in this recombination map. The HapMap recombination

map was produced using patterns of linkage disequilibrium (LD) in the HapMap samples. Use of these recombination rates to estimate allele ages based on the expected amount of linkage decay may be theoretically problematic. A future implementation of the test presented in this chapter will need to utilize non-LD based estimates of recombination rates such as the pedigree based recombination maps produced by the decode project (Kong et al., 2010).

Allele age estimation

To estimate derived allele ages for each population we applied a moment estimator first applied by Serre and colleagues (Serre et al., 1990; Sabeti, P.C. et al., 2002; Kim and Nielsen, 2004; Conrad et al., 2006; Voight et al., 2006; Wang et al., 2006; Hawks et al., 2007; Sabeti et al., 2007; Barreiro et al., 2008; Kelley et al., 2008; Pickrell et al., 2009; Grossman et al., 2010) and outlined by Slatkin and Rannala (Slatkin and Rannala, 2000) (Equation 2). This method utilizes the level of intra-allelic variability between two SNPs and the expected recombination fraction to estimate a time since the origin of a single derived allele.

To estimate allele ages using the above method, it is necessary to know which nearby alleles a new mutation was originally linked to. This information comes from genetic survey. The simplest computational approach to apply this method of estimation to the entire genome is to rely only on cases in which a new mutation arose linked to a second derived allele. By estimating ages only for cases in which two derived alleles are linked to each other, we can computationally ascertain which is the older mutation using allele frequency.

For each SNP in our filtered 1000 Genomes dataset our algorithm searched half a Centimorgan (cM) upstream and downstream of the focal SNP for nearby sites in which the derived allele was linked to the focal derived allele. If there was significant evidence of linkage disequilibrium (LD), determined by a one-tailed 5% chi-square cutoff, we estimated the age of the focal derived allele.

Age (in generations) was computed using:

$$t = \frac{1}{\ln(1-c)} \ln\left(\frac{x_t - y}{1 - y}\right),$$

where t = allele age (in generations), c = recombination rate between two loci, x_t = proportion of the derived chromosomes carrying the linked allele, y = proportion of the ancestral chromosomes carrying the linked allele. This method is a method-of-moments estimator. It requires no population genetic or demographic assumptions, only the exponential decay of initially perfect LD because of recombination (Slatkin and Rannala, 2000).

This procedure produced multiple age estimates for each focal SNP. To average these estimates we first filtered out all SNP pairs within a recombination distance of 0.0001 of each other. Thus, all nearby SNPs used for age estimation were between 0.0001 and 0.005 recombination units of the focal SNP. We then averaged all remaining age estimates for a focal allele by weighting each age (in generations) by the derived allele frequency of each linked allele.

Testing for effects of sample size

A sample of 200 chromosomes in a whole-genome or genotyping study generates an impressive amount of data. However, it is in reality a small sample for evaluating the

ages of alleles using LD approaches. An allele at 20% frequency in such a sample is represented by only 40 copies. Suppose that this allele was initially completely linked to a background haplotype and the linkage decayed over time, with 0.001 recombination event expected per generation per chromosome. We want to estimate the time since these loci were completely linked, by observing the number of copies of the allele that occur on other haplotypes. Without considering the stochastic variability in the occurrence of recombination or coalescence of gene lineages, we should still expect substantial error in resulting estimates of time due to sampling only 40 copies of the allele. After 100 generations, an expected 36 copies should remain on the original background (an expected frequency of around 90%), after 300 generations only 30 will remain, on expectation (an expected frequency of around 75%). Observing 30 copies is substantially different from observing 36, and if we assume a p-value of 0.05, each would be unlikely from a population where the other was the expected value. Yet neither is an unlikely outcome if the true population frequency is intermediate (say, around 82%, corresponding to a bit more than 200 generations). In this case we may expect to see 33 copies, but under sampling considerations alone may expect to see values as low as 28 or as high as 36. If we require a more stringent p-value, as may be appropriate for genome-wide data, consistent observations occur across an even greater range.

The point is, by sampling error alone we may take an allele with a true age of 200 generations and find estimates of time across a range greater than the mean -- from less than half to more than 50% higher.

We worked to evaluate the effect of sample size on the estimates of allele age generated by the moment estimator. In our procedure, we assumed that sampling from the

population is random and therefore frequencies are estimated consistent with a binomial distribution. We assumed values for *t*, the true age of the allele, *y*, the true initial frequency of the background allele linked to our focal allele, *c*, the recombination rate, and *p*, the frequency of the focal allele. The moment estimator of allele age is derived from the function that predicts the decline in frequency of the linked haplotype over time, namely:

$$x_t - y = (1 - c)^t (1 - y)$$

We used this equation to derive the expected x_t in the population given *t*, *c*, and *y*. Then, we derived the expected binomial sampling probability of every discrete value corresponding to possible observed x_t and *y* in a sample of 200 chromosomes with 200*p* copies of the focal allele. Each of these combinations of observed x_t and *y* would correspond to a single discrete estimate of t, as related by the moment estimator. The joint binomial sampling probabilities of x_t and *y* therefore provide a probability density for estimated t, given the assumed values of true *t*, *y*, *c* and *p*. We generated this probability density across a wide range of values for these parameters, to evaluate the sample space in which estimated allele ages are likely to be substantially in error.

In particular, we were interested in parameter values where high true ages nevertheless have a substantial probability of yielding low estimated ages. In these cases in particular, our 5,300-year-old human specimen might carry alleles that are truly old, but that look based on LD to be young. To the extent possible, we hoped to evaluate how much of the observation of "young" alleles in the Iceman genome could be attributed to

sampling error. We presume that such cases that cannot be attributed to sampling must instead be explained by other processes, including stochastic variation in gene genealogies and recombination.

Iceman genome data

We obtained aligned genome reads from the Tyrolean Iceman (Ötzi) from the European Short Read Archive. At the time of our download, the Ötzi data were provided as three BAM files; we used samtools (http://samtools.org) to merge these into a single dataset and extracted the SNP sites for each allele with a valid age estimation. In order to align the Ötzi reads (hg18) to the HapMap/1000 Genomes data (hg19) we first had to perform a genome build liftover using the UCSC Genome Browser Batch Coordinate Conversion utility (http://genome.ucsc.edu/cgi-bin/hgLiftOver).

The resulting dataset included the positions, frequency, age estimation, derived allele, and Ötzi reads for every available SNP in the HapMap recombination map. This file was used to produce a final output file, which scored each site for the presence or absence of the derived allele in the Ötzi genome. This final set of files was filtered to include only sites for which age was estimated by the weighted average of at least (5) nearby SNPs.

REFERENCES

Abadie V, Sollid LM, Barreiro LB, and Jabri B. 2011. Integration of Genetic and Immunological Insights into a Model of Celiac Disease Pathogenesis. Annual Review of Immunology 29:493–525.

Altshuler D, Brooks LD, Chakravarti A, Collins FS, Daly MJ, and Donnelly P. 2005. A haplotype map of the human genome. Nature 437:1299–1320.

Ammerman AJ, and Cavalli-Sforza LL. 1984. The Neolithic transition and the genetics of populations in Europe. Princeton, NJ: Princeton University Press.

Balaresque P, Bowden GR, Adams SM, Leung H-Y, King TE, Rosser ZH, Goodwin J, Moisan J-P, Richard C, et al. 2010. A Predominantly Neolithic Origin for European Paternal Lineages. Plos Biol 8:e1000285.

Barreiro LB, and Quintana-Murci LIS. 2010. From evolutionary genetics to human immunology: how selection shapes host defence genes. Nat Rev Genet 11:17–30.

Barreiro LB, Laval G, Quach H, Patin E, and Quintana-Murci L. 2008. Natural selection has driven population differentiation in modern humans. Nature Genetics 40:340–345.

Belle EMS, Landry PA, and Barbujani G. 2006. Origins and evolution of the Europeans' genome: evidence from multiple microsatellite loci. Proceedings of the Royal Society B: Biological Sciences 273:1595–1602.

Bersaglieri T, Sabeti P, Patterson N, Vanderploeg T, Schaffner S, Drake J, Rhodes M, Reich D, and Hirschhorn J. 2004. Genetic signatures of strong recent positive selection at the lactase gene. The American Journal of Human Genetics 74:1111–1120.

Bollongino R, Edwards CJ, Alt KW, Burger J, and Bradley D. 2006. Early history of European domestic cattle as revealed by ancient DNA. Biology Letters 2:155–159.

Bramanti B, Thomas MG, Haak W, Unterlaender M, Jores P, Tambets K, Antanaitis-Jacobs I, Haidle MN, Jankauskas R, et al. 2009. Genetic Discontinuity Between Local Hunter-Gatherers and Central Europe's First Farmers. Science 326:137–140.

Bramanti B. 2008. Ancient DNA: genetic analysis of aDNA from sixteen skeletons of the Vedrovice. Anthropologie 46:153–160.

Burger J, Kirchner M, Bramanti B, Haak W, and Thomas MG. 2007. Absence of the lactase-persistence-associated allele in early Neolithic Europeans. Proceedings of the National Academy of Sciences 104:3736–3741.

Childe VG. 1958. The dawn of European civilization. 6th ed. Knopf.

Conrad DF, Jakobsson M, Coop G, Wen X, Wall JD, Rosenberg NA, and Pritchard JK. 2006. A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. Nature Genetics 38:1251–1260.

Durbin RM, Altshuler DL, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, Collins FS, La Vega De FM, et al. 2010. A map of human genome variation from population-scale sequencing. Nature 467:1061–1073.

Enattah N, Sahi T, Savilahti E, and Terwilliger J. 2002. Identification of a variant associated with adult-type hypolactasia. Nature 30:233–237.

Enattah N, Trudeau A, Pimenoff V, and Maiuri L. 2007. Evidence of still-ongoing

convergence evolution of the lactase persistence T-13910 alleles in humans. The American Journal of Human Genetics 81:615–625.

Enattah NS, Jensen TGK, Nielsen M, Lewinski R, Kuokkanen M, Rasinperä H, El-Shanti H, Seo JK, Alifrangis M, et al. 2008. Independent introduction of two lactase-persistence alleles into human populations reflects different history of adaptation to milk culture. The American Journal of Human Genetics 82:57–72.

Fledel-Alon A, Leffler EM, Guan Y, Stephens M, Coop G, and Przeworski M. 2011. Variation in Human Recombination Rates and Its Genetic Determinants. PLoS ONE 6:e20321.

Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, et al. 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851–861.

Fu Q, Rudan P, Pääbo S, and Krause J. 2012a. Complete Mitochondrial Genomes Reveal Neolithic Expansion into Europe. PLoS ONE 7:e32473.

Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, Gabriel S, Altshuler D, Shendure J, et al. 2012b. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature. Online Preprint.

Gerbault P, Moret C, Currat M, and Sanchez-Mazas A. 2009. Impact of selection and demography on the diffusion of lactase persistence. PLoS ONE 4:e6369

Griffiths RC, and Tavaré S. 1998. The age of a mutation in a general coalescent tree. Communications in Statistics. Stochastic Models 14:273–295.

Griffiths RC, and Tavaré S. 1999. The ages of mutations in gene trees. Annals of Applied Probability:567–590.

Grossman SR, Shylakhter I, Karlsson EK, Byrne EH, Morales S, Frieden G, Hostetter E, Angelino E, Garber M, et al. 2010. A composite of multiple signals distinguishes causal variants in regions of positive selection. Science 327:883–886.

Guba Z, Hadadi É, Major Á, Furka T, Juhász E, Koós J, Nagy K, and Zeke T. 2011. HVS-I polymorphism screening of ancient human mitochondrial DNA provides evidence for N9a discontinuity and East Asian haplogroups in the Neolithic Hungary. Journal of Human Genetics 56:784–796.

Haak W, Balanovsky O, Sanchez JJ, Koshel S, Zaporozhchenko V, Adler CJ, Sarkissian Der CSI, Brandt G, Schwarz C, et al. 2010. Ancient DNA from European Early Neolithic Farmers Reveals Their Near Eastern Affinities. Plos Biol 8:e1000536.

Haak W, Brandt G, Jong HND, Meyer C, Ganslmeier R, Heyd V, Hawkesworth C, Pike AWG, Meller H, et al. 2008. Ancient DNA, Strontium isotopes, and osteological analyses shed light on social and kinship organization of the Later Stone Age.

Proceedings of the National Academy of Sciences 105:18226–18231.

Haak W, Forster P, Bramanti B, Matsumura S, Brandt G, Tänzer M, Villems R, Renfrew C, Gronenborn D, et al. 2005. Ancient DNA from the first European farmers in 7500-year-old Neolithic sites. Science 310:1016–1018.

Hawks J, Wang ET, Cochran GM, Harpending HC, and Moyzis RK. 2007. Recent acceleration of human adaptive evolution. Proceedings of the National Academy of Sciences 104:20753–20758.

Hinch AG, Tandon A, Patterson N, Song Y, Rohland N, Palmer CD, Chen GK, Wang K, Buxbaum SG, et al. 2011. The landscape of recombination in African Americans. Nature 476:170-175.

Holden C, and Mace R. 1997. Phylogenetic analysis of the evolution of lactose digestion in adults. Human Biology 69:605–628.

Keinan A, and Clark AG. 2012. Recent Explosive Human Population Growth Has Resulted in an Excess of Rare Genetic Variants. Science 336:740–743.

Keller A, Graefen A, Ball M, Matzas M, Boisguerin V, Maixner F, Leidinger P, Backes C, Khairat R, et al. 2012. New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing. Nat Comms 3:698.

Kelley JM, Hughes LB, Feng R, Liu N, Padilla MA, Vaughan LK, and Bridges SL. 2008. Evaluating linkage disequilibrium and recombination provides a haplotype-tagging SNP panel of the major histocompatibility complex in African Americans. Genes Immun 9:271–273.

Kim Y, and Nielsen R. 2004. Linkage Disequilibrium as a Signature of Selective Sweeps. Genetics. 167:1513-1524.

Kimura M, and Ohta T. 1973. The age of a neutral mutation persisting in a finite population. Genetics 75:199–212.

Kong A, Thorleifsson G, Gudbjartsson DF, Masson G, Sigurdsson A, Jonasdottir A, Bragi Walters G, Jonasdottir A, Gylfason A, et al. 2010. Fine-scale recombination rate differences between sexes, populations and individuals. Nature 467:1099-1103.

Kuokkanen M, Enattah NS, Oksanen A, Savilahti E, Orpana A, and Järvelä I. 2003. Transcriptional regulation of the lactase-phlorizin hydrolase gene by polymorphisms associated with adult-type hypolactasia. Gut 52:647–652.

Lee EJ, Makarewicz C, Renneberg R, Harder M, Krause-Kyora B, Müller S, Ostritz S, Fehren-Schmitz L, Schreiber S, et al. 2012. Emerging genetic patterns of the european neolithic: Perspectives from a late neolithic bell beaker burial site in Germany. Am J Phys Anthropol 148:571–579.

Macpherson JM, González J, Witten DM, Davis JC, Rosenberg NA, Hirsh AE, and Petrov DA. 2008. Nonadaptive explanations for signatures of partial selective sweeps in Drosophila. Molecular Biology and Evolution 25:1025–1042.

Malmström H, Linderholm A, Lidén K, Storå J, Molnar P, Holmlund G, Jakobsson M, and Götherström A. 2010. High frequency of lactose intolerance in a prehistoric hunter-gatherer population in northern Europe. BMC Evolutionary Biology 10:89.

Nikitin AG, Newton JR, and Potekhina ID. 2012. Mitochondrial haplogroup C in ancient mitochondrial DNA from Ukraine extends the presence of East Eurasian genetic lineages in Neolithic Central and Eastern Europe. Journal of Human Genetics 57:610–612.

Novembre J, and Stephens M. 2008. Interpreting principal component analyses of spatial population genetic variation. Nature Genetics 40:646–649.

Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, Absher D, Srinivasan BS, Barsh GS, Myers RM, et al. 2009. Signals of recent positive selection in a worldwide sample of human populations. Genome Research 19:826–837.

Plantinga TS, Alonso S, Izagirre N, Hervella M, Fregel R, van der Meer JW, Netea MG, and la Rúa de C. 2012. Low prevalence of lactase persistence in Neolithic South-West Europe. European Journal of Human Genetics 20:778–782.

Price TD. 2000. Europe's First Farmers. New York: Cambridge University Press.

Puit I, Sajantila A, Simanainen J, Georgiev O, Schaffner W, and Pääbo S. 1994. Mitochondrial DNA sequences from Switzerland reveal striking homogeneity of European populations. Biological Chemistry Hoppe-Seyler 375:837–840.

Richards M, Macaulay V, Hickey E, Vega E, Sykes B, Guida V, Rengo C, Sellitto D, Cruciani F, et al. 2000. Tracing European Founder Lineages in the Near Eastern mtDNA Pool. American Journal of Human Genetics 67:1251.

Risch N, de Leon D, Ozelius L, Kramer P, Almasyz L, Singer B, Fahn S, Breakefield X, and Bressman S. 1995. Genetic analysis of idiopathic torsion dystonia in Ashkenazi Jews and their recent descent from a small founder population. Nature Genetics 9:l52–159.

Rosser ZH, Zerjal T, Matthew E Hurles, Adojaan M, Alavantic D, Amorim AN, Amos W, Armenteros M, Arroyo E, et al. 2000. Y-Chromosomal Diversity in Europe Is Clinal and Influenced Primarily by Geography, Rather than by Language. American Journal of Human Genetics 67:1526.

Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, Mccarroll SA, et al. 2007. Genome-wide detection and characterization of positive selection in human populations. Nature 449:913–918.

Sabeti, P.C., Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, et al. 2002. Detecting recent positive selection in the human genome from haplotype structure. Nature 419:832–837.

Sánchez-Quinto F, Schroeder H, Ramirez O, Avila-Arcos MC, Pybus M, Olalde I, Velazquez A, Marcos MEP, Encinas JMV, et al. 2012. Genomic Affinities of Two 7,000-Year-Old Iberian Hunter-Gatherers. Current Biology 22:R631–R633.

Semino O, Passarino G, Oefner PJ, Lin AA, and Arbuzova S. 2000. The Genetic Legacy of Paleolithic Homo sapiens sapiens in Extant Europeans: A Y Chromosome Perspective. Science 290:1155-1159.

Serre JL, Simon-Bouy B, Mornet E, Jaume-Roig B, Balassopoulou A, Schwartz M, Taillandier A, Bou J, and Bou A. 1990. Studies of RFLP closely linked to the cystic fibrosis locus throughout Europe lead to new considerations in populations genetics. Hum Genet 84:449-454.

Simoni L, Calafell F, Pettener D, Bertranpetit J, and Barbujani G. 2000. Geographic patterns of mtDNA diversity in Europe. The American Journal of Human Genetics 66:262–278.

Skoglund P, Malmström H, Raghavan M, Storå J, Hall P, Willerslev E, Gilbert MTP, Götherström A, and Jakobsson M. 2012. Origins and Genetic Legacy of Neolithic Farmers and Hunter-Gatherers in Europe. Science 336:466–469.

Slatkin M, and Rannala B. 2000. Estimating allele age. Annu. Rev. Genom. Human Genet. 1:225–249.

Sokal RR, and Menozzi P. 1982. Spatial autocorrelations of HLA frequencies in Europe support demic diffusion of early farmers. American Naturalist 119:1–17.

Sokal RR, Oden NL, and Wilson C. 1991. Genetic evidence for the spread of agriculture in Europe by demic diffusion. Nature 351:143–145.

Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, Powell K, Mortensen HM, Hirbo JB, et al. 2006. Convergent adaptation of human lactase persistence in Africa and Europe. Nature Genetics 39:31–40.

Voight BF, Kudaravalli S, Wen X, and Pritchard JK. 2006. A map of recent positive selection in the human genome. Plos Biol 4:e72.

Wang E, Kodama G, Baldi P, and Moyzis R. 2006. Global landscape of recent inferred Darwinian selection for Homo sapiens. Proceedings of the National Academy of Sciences 103:135-140.

Wegmann D, Kessner DE, Veeramah KR, Mathias RA, Nicolae DL, Yanek LR, Sun YV, Torgerson DG, Rafaels N, et al. 2011. Recombination rates in admixed individuals identified by ancestry-based inference. Nature Genetics 43:847–853.

Zhernakova A, Elbers CC, Ferwerda B, Romanos J, Trynka G, Dubois PC, de Kovel CGF, Franke L, Oosting M, et al. 2010. Evolutionary and Functional Analysis of Celiac Risk Loci Reveals SH2B3 as a Protective Factor against Bacterial Infection. The American Journal of Human Genetics 86:970–977.

Zvelebil M, and Dolukhanov P. 1991. The transition to farming in Eastern and Northern Europe. J World Prehist 5:233–278.

Figure 1. Distributions of allele ages

Histograms representing the distribution of estimated allele ages with no filter (blue) and filtered to include only alleles with derived allele frequency between 5% and 50%.

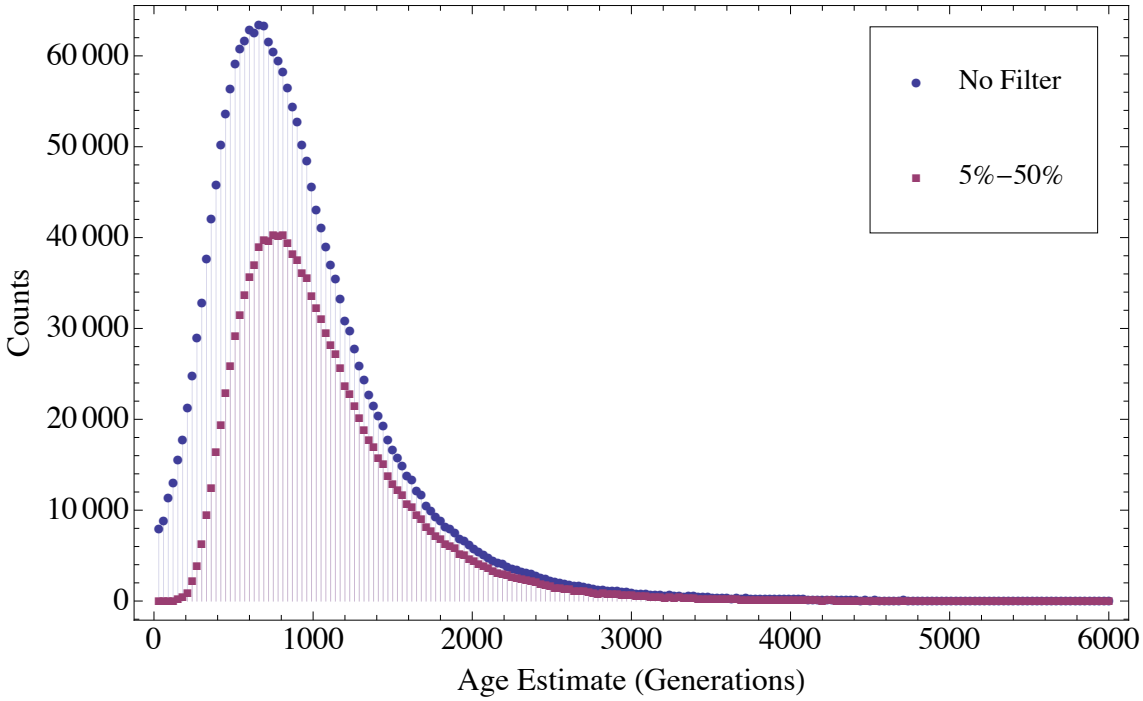


Figure 2. Effect of sampling bias from allele frequency

Plot of age estimation probabilities for all possible allele count combinations given a sample size of 200 haplotypes, true allele age of 200 generations, recombination rate 0.003, and *y* equal to 0.2. The difference between the two plots is in the value of *p* (the focal allele frequency). Variance in sampling probability is higher in the case where *p* is extremely low.

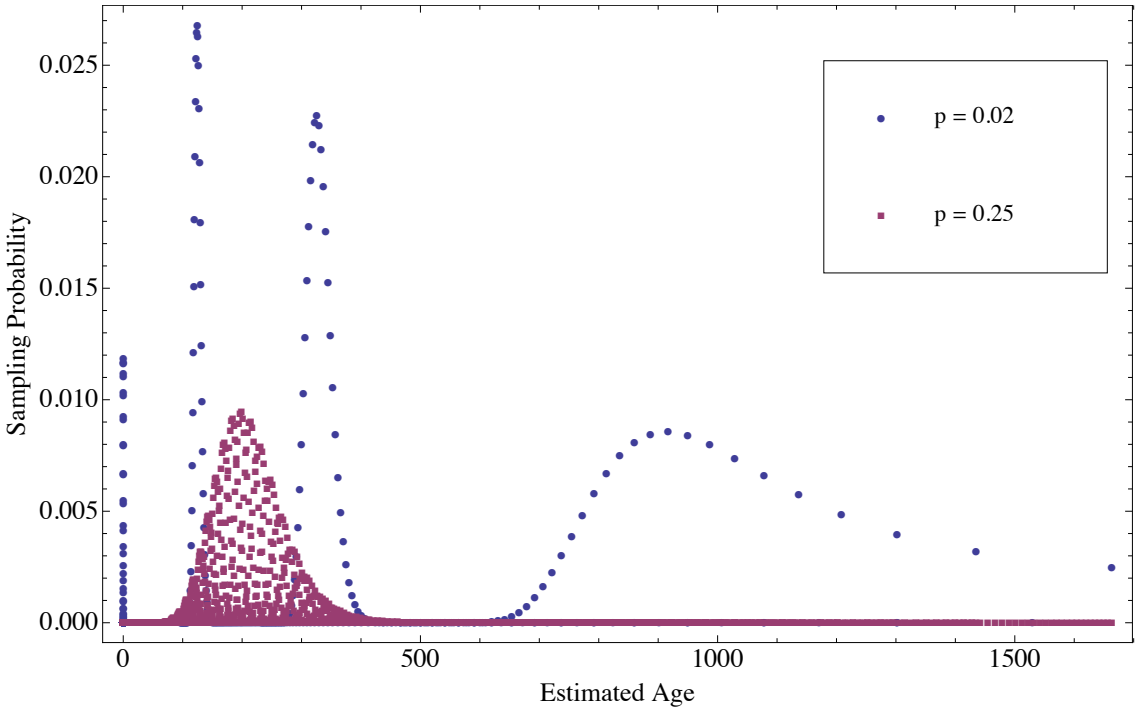


Figure 3. Effect of sampling bias from *y*

Plot of age estimation probabilities for all possible allele count combinations given a sample size of 200 haplotypes, true allele age of 200 generations, recombination rate 0.003, and focal allele frequency 0.25. The difference between the two plots is in the value of *y* (the proportion of haplotypes carrying the non-focal allele and the allele linked to the focal allele). Variance in sampling probability is higher in the case where *y* is higher (and therefore linked allele frequency is also higher).

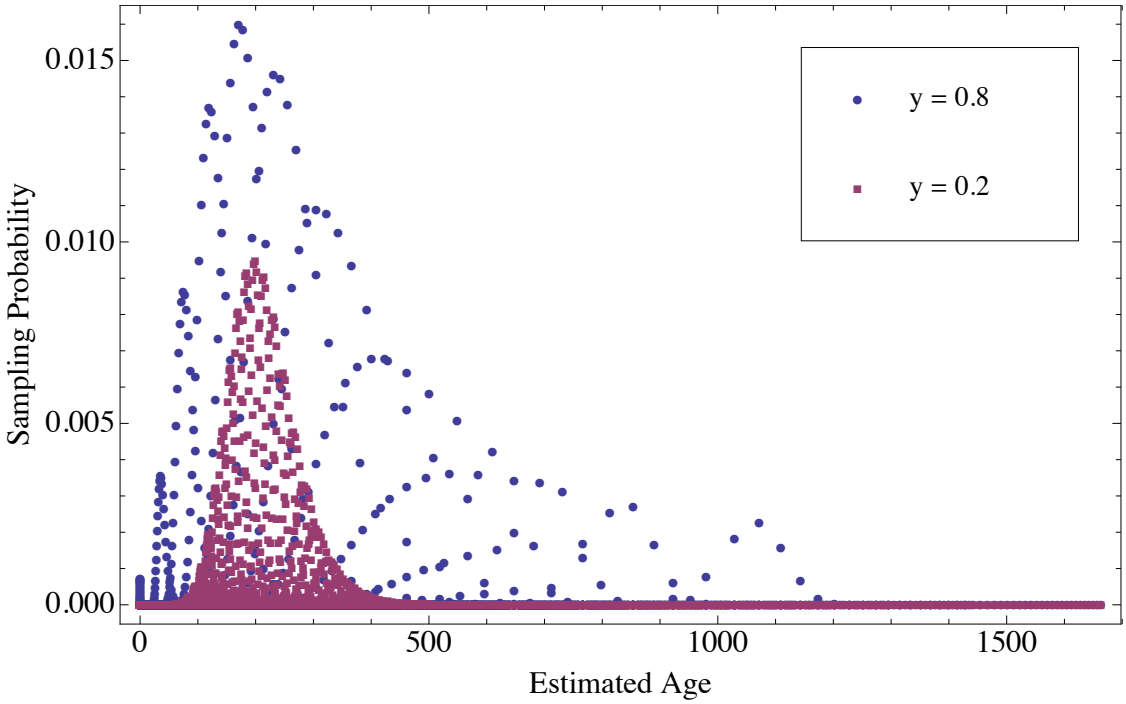


Table 1. Division of alleles by age and presence in Ötzi

Age estimates calculated for every SNP in dataset based on weighted average of all SNPs within 0.001-0.005 recombination units (see Methods). Filtering of high and low frequency classes removes a large proportion of sampling bias.

	No frequency filter	5% - 50%
Present in Ötzi t < 200 / t > 200	78617 / 933398 (0.0842)	510 / 500178 (0.00102)
Absent from Ötzi t < 200 / t > 200	9705/1187744 (0.00817)	858 / 826525 (0.00104)
χ^2 contingency test p-values	$p < 1 \times 10^{-100}$	$p = 0.749$

Table 2. Sampling bias on age estimation when linked allele frequency is high and when *y* is high.

Age estimation where linked allele frequency (q) is relatively low results in a smaller proportion of cases in which young alleles are found in Ötzi's genome. These values consist of every available SNP in the HapMap dataset. Estimates were calculated based on a single randomly chosen SNP between 0.001 and 0.005 recombination units from the focal SNP. Filtering on y alone demonstrates that high values of y on average produce a heavy bias towards young alleles as indicated by the abundance of young alleles in both categories, the excess of young alleles present in Ötzi ensures that these are mis-estimations.

	$0.2 \leq q \leq 0.25$	$0.75 \leq q \leq 0.8$	$0.2 \leq y \leq 0.25$	$0.75 \leq y \leq 0.8$
Present in Ötzi t < 200 / t > 200	86 / 166885 (0.000515)	10076 / 532789 (0.0189)	40790 / 365060 (0.112)	36344 / 15755 (2.31)
Absent from Ötzi t < 200 / t > 200	179 / 684067 (0.000262)	1802 / 608100 (0.00296)	7207 / 295102 (0.0244)	1046 / 724 (1.44)
χ^2 contingency test p-values	$p < 1 \times 10^{-5}$	$p < 1 \times 10^{-100}$	$p < 1 \times 10^{-100}$	$p < 1 \times 10^{-20}$

Chapter V

GENERAL DISCUSSION

Key insights

The results from this study suggest that the genetic loci currently associated with celiac disease (CD) risk have experienced different evolutionary dynamics than the genome-wide average. To reiterate our methods: If this set of loci had evolved under the same pattern of demographic history and selection as the rest of the genome, then comparisons among populations ought to show the same pattern of between-population differentiation for these loci as for the genome as a whole. For some inter-population comparisons, the CD risk loci do show a pattern that is not significantly different from randomly chosen loci across the genome. In these cases, we cannot conclude that selection had no effect on the CD risk loci, but we can say that it did not affect these loci in a different way from randomly chosen sets of loci across the genome. For some cross-continental comparisons, we find that the CD risk loci show a significantly greater degree of differentiation than for randomly chosen sets of loci. In these cases, we can conclude that the dynamics of the CD risk loci during the evolution of these populations were different from the genome as a whole. The genome reflects demography, widespread purifying selection on most coding regions, and occasional balancing and directional selection on a fraction of genes.

For CD risk loci, we can conclude that this set of genes likely underwent a different pattern of selection leading to greater differentiation of some populations.

The signal of greater differentiation involves comparisons that include Europe, which suggests that natural selection was acting on several of the CD risk loci during the time period that present Europeans were differentiating from other continental populations. Directional selection on standing variants during the evolutionary history of this population may explain the pattern of differentiation of these genes. Alternatively, a relaxation of selection relative to the rest of the genome might explain the high differentiation of this set of loci. If selection were relaxed during the evolution of Europeans, then we should expect that differentiation of this set of loci within Europe would also be high relative to the rest of the genome. We observe instead that differentiation within Europe was low for this set of loci. Likewise, if directional selection on this set of loci were uniform across the evolution of European populations, or if directional selection were more intense in recent times than in the initial differentiation of European populations from other continental populations, we might expect that the differentiation would have continued within recent times, affecting the pattern of within-Europe differentiation. Again, we find that the present differences between two European populations (TSI from Tuscany and CEU with origins in northern Europe) are slight relative to the genome as a whole. Therefore, we cannot conclude that the Neolithic transition was the major driver of selection on these loci.

Despite our conclusion that Holocene selection within Europe cannot explain the overall pattern of differentiation, nevertheless some of the loci are consistent with recent positive selection within Europe. This is where our consideration of the network of loci

must become more complex. For example, analysis of population differentiation within Europe revealed that CD associated loci are devoid of genetic differentiation between samples from northern (CEU) and southern (TSI) Europe. This result might be interpreted to reflect a lack of selection since the founding and differentiation of these two populations. However, demographic factors might explain the lack of differentiation across European populations. For example, if a significant portion of the ancestry of present Europeans is found among early Neolithic Near Eastern agriculturalists as has been predicted from archaeological and genomic data (Sokal et al., 1991; Haak et al., 2010; Fu et al., 2012; Gamba et al., 2012; Pinhasi, 2012; Sánchez-Quinto et al., 2012; Skoglund et al., 2012). Given the genetic uniformity of European samples in our analysis, CD risk loci may have been selected early enough in the history of this demographic event to have spread to all present European populations, contrary to previous estimates (Zhernakova et al., 2010). Our investigation of CD associated GWAS SNPs in the Iceman genome revealed some evidence of frequency shifts at these loci during the past 5,000 years.

The strongest signal of recent selection on the CD background network from previous studies is the rs3184504 SNP in the SH2B3 gene, which was estimated to have undergone a selective sweep in Europe within the past 1,700 to 1,200 years (Zhernakova et al., 2010). However, based on the fact that we found this risk SNP in the 5,300 Tyrolean Iceman genome, we used the full Iceman genome to investigate whether underestimation of allele ages from intra-allelic variability is a more common problem. We found in Chapter 4 that sampling error does not likely account for all cases of allele age underestimation (as represented by the presence of an allele in the Iceman genome

that is estimated to be younger than 200 generations). This suggests that some existing cases of putatively recently selected alleles may not have been selected as recently as previously thought. With respect to CD, we can conclude with some certainty that this is the case for the SH2B3 risk allele.

An argument has been repeatedly made that recent increases in rates of autoimmune diseases in Westernized populations are the result of a lack of infectious agents, particularly parasitic infections, that in the past may have mediated autoimmune responses(Dunne, 2005). This argument, related to the "hygiene hypothesis" posited for increased rates of allergies and asthma in Westernized populations, has been supported recently for a few autoimmune diseases, particularly irritable bowel disease (Olszak et al., 2012) (see (Dunne, 2005) for a comprehensive review). However, no convincing evidence has been presented to support this argument for CD. In fact, some have argued a vital role for viral infections in the pathogenesis of CD (Kagnoff, 1984; 2007), although some recent research does not support this argument (Lebwohl et al., 2012). Investigating this further is beyond the scope of the current project. However, future directions discussed below should help to contribute to this question. For example, current knowledge about the distribution of CD risk is biased towards European populations. This lack of knowledge about the global distribution and rates of CD risk prevent us from understanding if CD occurrence has actually increased as a result of post-industrial hygienic environments. Additionally, we are currently ill-equipped to understand the extent to which CD pathogenesis is related to the broader evolutionary picture of CD risk alleles. For example, this analysis revealed variation between African populations. It is unclear if variation between African populations in CD risk loci might

contribute to functional variation in autoimmune risk because little is known about rates of autoimmune diseases, including CD, in these populations. Increased insights into the deeper evolutionary history of these loci will provide more information about the evolution of complex autoimmune diseases and will additionally facilitate a more complete understanding of the conflicting roles of diet and immunity in shaping several autoimmune diseases such as CD and type 1 diabetes. Significantly higher occurrence of CD in populations of north Africa, such as the Saharawi (Catassi et al., 1999; 2001; Alarida et al., 2010), compared to Europe, and the shorter history of gluten based cereal consumption in these populations suggest a potential role for balancing selection between diet and immunity in European populations to mediate celiac risk. More information about the current and past global distribution of CD is necessary to address such questions. Priority should be placed at first on ascertaining the global distribution of CD risk, particularly in African populations. Several sub-Saharan African populations, such as the San, Mbuti, and Hadza maintain foraging lifestyles. Others have long histories of agricultural/horticultural lifestyles. An initial goal of this project was to understand if, as has been predicted for Europe (Barreiro and Quintana-Murci, 2010; Zhernakova et al., 2010; Abadie et al., 2011), agricultural lifestyles (including larger population sizes and increased sedentism) are associated with evolution in immune-related regions of the genome, and whether this evolution might lead to population scale differences in autoimmune risk. However, a lack of information about population scale autoimmune risk and whole genome sequence data in sub-Saharan African populations prevented such investigation. For Europe, where CD risk is well known, no comparative foraging populations remain, therefore increasingly large samples of ancient DNA are key to

understanding the role of agricultural lifestyles in shaping CD risk. Our current results implying that at least some CD loci have experience recent frequency shifts suggest that future analysis using larger samples of ancient DNA will be promising.

Future approaches

The idea that the past distribution and occurrence of CD can be ascertained using ancient DNA was recently bolstered by skeletal and DNA evidence from a roughly 2000 year-old female from the archaeological site of Cosa, near Tuscany, Italy. This skeleton includes classic paleopathological signs of infection and malnutrition leading to death prior to age 20 and also contains the most common HLA-DQ risk haplotype (DQ2.5) for CD (Gasbarrini et al., 2012). Given large enough ancient samples that include skeletal and DNA data, an analysis analogous to present day case/control studies could be performed to study the past distribution of CD risk. The enrichment of known CD genetic risk markers in archaeological samples presenting symptoms of CD on the skeleton (early death associated with malnourishment, stunted growth, osteoporosis, etc…) would provide a means of ascertaining the past distribution of CD risk. Furthermore, with large enough samples from both Neolithic (agricultural) and Mesolithic (pre-agricultural) samples, the recent evolution of CD risk, including the role of agriculture and demography in shaping this evolution, can be clarified. We expect that future observations will confirm our conclusion that the genetic network underlying CD risk has evolved piecemeal rather than cohesively due to a single period of past selection and that such a model is applicable to a wider array of common immune-mediated conditions.

REFERENCES

Abadie V, Sollid LM, Barreiro LB, and Jabri B. 2011. Integration of Genetic and Immunological Insights into a Model of Celiac Disease Pathogenesis. Annual Review of Immunology 29:493–525.

Alarida K, Harown J, Di Pierro M, and Drago S. 2010. HLA-DQ2 and-DQ8 genotypes in celiac and healthy Libyan children. Digestive and Liver Disease 42:425-427.

Barreiro LB, and Quintana-Murci LIS. 2010. From evolutionary genetics to human immunology: how selection shapes host defence genes. Nat Rev Genet 11:17–30.

Catassi C, Doloretta Macis M, Ratsch IM, De Virgiliis S, and Cucca F. 2001. The distribution of DQ genes in the Saharawi population provides only a partial explanation for the high celiac disease prevalence. Tissue Antigens 58:402–406.

Catassi C, Ratsch I, Gandolfi L, Pratesi R, and Fabiani E. 1999. Why is coeliac disease endemic in the people of the Sahara? The Lancet 354:647–648.

Dunne D. 2005. A worm's eye view of the immune system: consequences for evolution of human autoimmune disease. Nature Reviews Immunology 5:420-426.

Fu Q, Rudan P, Pääbo S, and Krause J. 2012. Complete Mitochondrial Genomes Reveal Neolithic Expansion into Europe. PLoS ONE 7:e32473.

Gamba C, Fernández E, Tirado M, Deguilloux M, Pemonge M, Utrilla P, Edo M, Molist M, Rasteiro R, et al. 2012. Ancient DNA from an Early Neolithic Iberian population supports a pioneer colonization by first farmers. Mol Ecol 21:45-56.

Gasbarrini G, Rickards O, and Martínez-Labarga C. 2012. Origin of celiac disease: How old are predisposing haplotypes? World Journal of Gastroenterology 18:5300–5304.

Haak W, Balanovsky O, Sanchez JJ, Koshel S, Zaporozhchenko V, Adler CJ, Sarkissian Der CSI, Brandt G, Schwarz C, et al. 2010. Ancient DNA from European Early Neolithic Farmers Reveals Their Near Eastern Affinities. Plos Biol 8:e1000536.

Kagnoff MF. 1984. Possible role for a human adenovirus in the pathogenesis of celiac disease. Journal of Experimental Medicine 160:1544–1557.

Kagnoff MF. 2007. Celiac disease: pathogenesis of a model immunogenetic disease. Journal of Clinical Investigation 117:41–49.

Lebwohl B, Green P, Murray JA, and Ludvigsson JF. 2012. Season of birth in a nationwide cohort of coeliac disease patients. Archives of Disease in Childhood online preprint.

Olszak T, An D, Zeissig S, Vera MP, Richter J, Franke A, Glickman JN, Siebert R, Baron RM, et al. 2012. Microbial Exposure During Early Life Has Persistent Effects on

Natural Killer T Cell Function. Science 336:489–493.

Pinhasi R. 2012. A Craniometric Perspective on the Transition to Agriculture in Europe. Human Biology 84:45-66.

Sánchez-Quinto F, Schroeder H, Ramirez O, Avila-Arcos MC, Pybus M, Olalde I, Velazquez A, Marcos MEP, Encinas JMV, et al. 2012. Genomic Affinities of Two 7,000-Year-Old Iberian Hunter-Gatherers. Current Biology 22:R631–R633.

Skoglund P, Malmström H, Raghavan M, Storå J, Hall P, Willerslev E, Gilbert MTP, Götherström A, and Jakobsson M. 2012. Origins and Genetic Legacy of Neolithic Farmers and Hunter-Gatherers in Europe. Science 336:466–469.

Sokal RR, Oden NL, and Wilson C. 1991. Genetic evidence for the spread of agriculture in Europe by demic diffusion. Nature 351:143–145.

Zhernakova A, Elbers CC, Ferwerda B, Romanos J, Trynka G, Dubois PC, de Kovel CGF, Franke L, Oosting M, et al. 2010. Evolutionary and Functional Analysis of Celiac Risk Loci Reveals SH2B3 as a Protective Factor against Bacterial Infection. The American Journal of Human Genetics 86:970–977.

APPENDIX

Part1. Original CD associated regions and 50kb intervals used in Ch. II.

Regions from Trynka et al. 2011 Table 2: Positions of highly correlated CD variants (hg18).				Results from this study.		
Chr.	**hg18 Start**	**hg18 Stop**	**Nearby Protein Coding Genes**	**Chr.**	**hg19 Start**	**hg19 End**
chr1	2510162	2710035	C1orf93,MMEL1,TTC34	Partial Coordinate Liftover Failure		
chr1	25162321	25177139	0-10 kb 5' and first exon of RUNX3	1	25272143	25322143
chr1	170940206	170948695	35-43 kb 5' of FASLG	1	172652828	172702828
chr1	171129607	171131275	Intergenic region between FASLG and TNFSF18	1	172838818	172888818
chr1	190779182	single	32 kb 5' RGS1	1	192487559	192537559
chr1	190786488	190811722	0-24 kb 5' of and the first exon of RGS1	1	192507482	192557482
chr1	199148015	single	Ninth intron of c1orf106	1	200856392	200906392
chr2	61040333	61058360	Exons 5-11 of PUS10	2	61170843	61220843
chr2	68493221	68499064	Intergenic region between PLEK and FBXO48	2	68617639	68667639
chr2	102338297	102459513	IL18R1 and IL18RAP	2	102957473	103007473
				2	103007473	103057473
				2	103057473	103107473
chr2	181708291	181803246	Intergenic region between UBE2E3 and ITGA4	2	181972524	182022524
				2	182022524	182072524
				2	182072524	182122524
chr2	191621279	191643278	Exons 6-14 of STAT4	2	191899034	191949034
chr2	191656882	single	Intron 3 of STAT4	2	191923637	191973637
chr2	191681808	single	Intron 3 of STAT4	2	191948563	191998563
chr2	204158521	204168206	111-121 kb 5' of CD28	2	204430119	204480119
chr2	204318641	204320303	Intergenic region between CD28 and CTLA4	2	204688462	204738462
chr2	204470572	204478299	Intergenic region between CTLA4 and ICOS	2	204741191	204791191
chr3	33012725	33012756	Intergenic region between CCR4 and GLB1	3	33012737	33062737
chr3	46162711	46180690	38-55 kb 3' of CCR1	3	46171697	46221697
chr3	46321275	46377631	Intergenic region between CCR3 and CCR2	3	46349449	46399449
chr3	46458634	46480319	Exons 2-13 of LTF	3	46469473	46519473
chr3	120601187	120605968	Intron 10 of ARHGAP31	3	119095888	119145888
chr3	161106253	single	Intergenic region between SCHIP1 and IL12A	3	159598529	159648529
chr3	161112778	161147744	Intergenic region between SCHIP1 and IL12A	3	159622566	159672566
chr3	161136316	161168494	Intergenic region between SCHIP1 and IL12A	3	159644711	159694711

chr3	189587750	189602595	Intron 2 of LPP	3	188087479	188137479
chr4	123257527	123770564	Multiple genes (KIAA1109, ADAD1, IL2 and IL21) (combined 2 signals for this analysis)	4	123045809	123095809
				4	123095809	123145809
				4	123145809	123195809
				4	123195809	123245809
				4	123245809	123295809
				4	123295809	123345809
				4	123345809	123395809
				4	123395809	123445809
				4	123445809	123495809
				4	123495809	123545809
chr6	341321	single	Intron 4 of IRF4	6	371321	421321
chr6	353079	355417	3' UTR of IRF4	6	384248	434248
chr6	90866360	90875874	Intron 2 of BACH2	6	90789396	90839396
chr6	128307943	128339304	PTPRK exons 28-30 in the 3' UTR to 24kb 3'	6	128256931	128306931
chr6	128332892	128335255	Last exon of PTPRK in the 3' UTR	6	128267381	128317381
chr6	138000928	138048197	Intergenic region between OLIG3 and TNFAIP3	6	137957870	138007870
chr6	138015797	138043754	Intergenic region between OLIG3 and TNFAIP3	6	137963083	138013083
chr6	159385965	159390046	4 kb 5' and 5' UTR of TAGAP	6	159443018	159493018
chr6	159418255	single	32 kb 5' of TAGAP	6	159473267	159523267
chr7	37366994	37404402	Intron 1 of ELMO1	7	37394173	37444173
chr8	129333242	129345888	151-163 kb 3' of PVT1	8	129248883	129298883
chr10	6430198	single	Intergenic region between PFKFB3 and PRKCQ	10	6365192	6415192
chr10	80728033	single	Intron 14 of ZMIZ1	10	81033027	81083027
chr11	110682429	110815769	*Not high density genotyped, POU2AF1, C11orf93	11	111218889	111268889
chr11	118080536	118085075	Intergenic region between TREH and DDX6	11	118552596	118602596
chr11	127886184	127901948	5 kb 5' and the first exon of ETS1	11	128363856	128413856
chr12	110368991	110494139	5' UTR and exons 1-3 of SH2B3, exons 2-25 and the 3' UTR of ATXN2	12	111871182	111921182
				12	111921182	111971182
				12	111971182	112021182
chr14	68329159	68341722	1 kb 5' of and the first exon of ZFP36L1	14	69240688	69290688
chr15	72397784	73270664	*Not high density genotyped, CLK3, CSK and	15	75022171	75072171

			multiple other genes			
chr16	10834038	10903351	*Not high density genotyped, CIITA	16	10961194	11011194
chr16	11254549	11268703	11 kb 5' of, 1kb 3' of and all of SOCS1	16	11329125	11379125
chr16	11281298	single	Intergenic region between PRM1 and PRM2	16	11348797	11398797
chr16	11292457	single	10 kb 5' of PRM1	16	11359956	11409956
chr18	12811903	12870206	Exons 2-5 of PTPN2	18	12826055	12876055
chr18	12847758	single	Intron 2 of PTPN2	18	12832758	12882758
chr21	42728136	single	Intron 9 of UBASH3A	21	43830067	43880067
chr21	44446245	44453549	18-25 kb 3' of ICOSLG	21	45600469	45650469
chr22	20250903	20313260	UBE2L3, YDJC	22	21927081	21977081

(Bold values in Tables 2-7 indicate Ratios > 0.01 or P-Values < 0.05)

Part 2. Within Africa Fst Count-Ratio Results

Results from this study.			**Africa (LWK x YRI)**			
Chr.	**hg19 Start**	**hg19 End**	**# SNPs**	**# > 99%**	**Ratio**	**P-Value**
Partial Coordinate Liftover Failure						
1	25272143	25322143	424	1	0.0024	0.5285
1	172652828	172702828	336	29	**0.0863**	**0.0112**
1	172838818	172888818	320	6	**0.0188**	0.1578
1	192487559	192537559	318	10	**0.0314**	0.0835
1	192507482	192557482	390	2	0.0051	0.3897
1	200856392	200906392	345	5	**0.0145**	0.2021
2	61170843	61220843	237	6	**0.0253**	0.1103
2	68617639	68667639	550	61	**0.1109**	**0.0064**
2	102957473	103007473	425	12	**0.0282**	0.0981
2	103007473	103057473	382	0	0.0000	1.0000
2	103057473	103107473	459	8	**0.0174**	0.1705
2	181972524	182022524	326	0	0.0000	1.0000
2	182022524	182072524	335	0	0.0000	1.0000
2	182072524	182122524	256	0	0.0000	1.0000
2	191899034	191949034	304	2	0.0066	0.3447
2	191923637	191973637	319	2	0.0063	0.3526
2	191948563	191998563	335	2	0.0060	0.3630
2	204430119	204480119	405	1	0.0025	0.5155
2	204688462	204738462	330	0	0.0000	1.0000
2	204741191	204791191	339	0	0.0000	1.0000
3	33012737	33062737	446	5	**0.0112**	0.2493
3	46171697	46221697	321	27	**0.0841**	**0.0116**
3	46349449	46399449	419	16	**0.0382**	0.0627
3	46469473	46519473	402	45	**0.1119**	**0.0062**

3	119095888	119145888	403	0	0.0000	1.0000
3	159598529	159648529	334	20	**0.0599**	**0.0259**
3	159622566	159672566	426	33	**0.0775**	**0.0135**
3	159644711	159694711	440	34	**0.0773**	**0.0135**
3	188087479	188137479	424	6	**0.0142**	0.2059
4	123045809	123095809	344	2	0.0058	0.3674
4	123095809	123145809	282	0	0.0000	1.0000
4	123145809	123195809	276	2	0.0072	0.3278
4	123195809	123245809	256	1	0.0039	0.4388
4	123245809	123295809	226	0	0.0000	1.0000
4	123295809	123345809	309	0	0.0000	1.0000
4	123345809	123395809	313	0	0.0000	1.0000
4	123395809	123445809	322	0	0.0000	1.0000
4	123445809	123495809	308	0	0.0000	1.0000
4	123495809	123545809	347	0	0.0000	1.0000
6	371321	421321	448	5	**0.0112**	0.2497
6	384248	434248	423	5	**0.0118**	0.2398
6	90789396	90839396	324	1	0.0031	0.4690
6	128256931	128306931	366	12	**0.0328**	0.0790
6	128267381	128317381	376	10	**0.0266**	0.1046
6	137957870	138007870	443	0	0.0000	1.0000
6	137963083	138013083	426	0	0.0000	1.0000
6	159443018	159493018	400	0	0.0000	1.0000
6	159473267	159523267	417	1	0.0024	0.5233
7	37394173	37444173	430	13	**0.0302**	0.0877
8	129248883	129298883	540	3	0.0056	0.3761
10	6365192	6415192	468	12	**0.0256**	0.1084
10	81033027	81083027	370	0	0.0000	1.0000
11	111218889	111268889	400	8	**0.0200**	0.1493
11	118552596	118602596	489	0	0.0000	1.0000
11	128363856	128413856	294	2	0.0068	0.3386
12	111871182	111921182	221	0	0.0000	1.0000
12	111921182	111971182	199	0	0.0000	1.0000
12	111971182	112021182	243	2	0.0082	0.3035
14	69240688	69290688	353	1	0.0028	0.4828
15	75022171	75072171	324	1	0.0031	0.4690
16	10961194	11011194	485	0	0.0000	1.0000
16	11329125	11379125	470	9	**0.0191**	0.1548
16	11348797	11398797	530	9	**0.0170**	0.1751
16	11359956	11409956	549	2	0.0036	0.4485
18	12826055	12876055	438	0	0.0000	1.0000

18	12832758	12882758	452	0	0.0000	1.0000
21	43830067	43880067	554	20	**0.0361**	0.0675
21	45600469	45650469	605	9	**0.0149**	0.1978
22	21927081	21977081	364	1	0.0027	0.4891
	Totals		25770	464	**0.0180**	
	Totals (no repeat SNPs)		22923	403	**0.0176**	

Part 3. Within East Asia Fst Count-Ratio Results

Results from this study.			**East Asia (CHB x CHS)**			
Chr.	**hg19 Start**	**hg19 End**	**# SNPs**	**# > 99%**	**Ratio**	**P-Value**
Partial Coordinate Liftover Failure						
1	25272143	25322143	223	0	0.0000	1.0000
1	172652828	172702828	184	0	0.0000	1.0000
1	172838818	172888818	187	0	0.0000	1.0000
1	192487559	192537559	154	0	0.0000	1.0000
1	192507482	192557482	193	0	0.0000	1.0000
1	200856392	200906392	217	1	0.0046	0.2718
2	61170843	61220843	119	2	**0.0168**	0.1115
2	68617639	68667639	343	0	0.0000	1.0000
2	102957473	103007473	296	0	0.0000	1.0000
2	103007473	103057473	248	0	0.0000	1.0000
2	103057473	103107473	308	0	0.0000	1.0000
2	181972524	182022524	216	1	0.0046	0.2709
2	182022524	182072524	213	0	0.0000	1.0000
2	182072524	182122524	162	0	0.0000	1.0000
2	191899034	191949034	155	0	0.0000	1.0000
2	191923637	191973637	161	1	0.0062	0.2156
2	191948563	191998563	179	1	0.0056	0.2321
2	204430119	204480119	205	1	0.0049	0.2595
2	204688462	204738462	140	0	0.0000	1.0000
2	204741191	204791191	172	0	0.0000	1.0000
3	33012737	33062737	248	5	**0.0202**	0.0987
3	46171697	46221697	177	0	0.0000	1.0000
3	46349449	46399449	237	0	0.0000	1.0000
3	46469473	46519473	252	0	0.0000	1.0000
3	119095888	119145888	262	0	0.0000	1.0000
3	159598529	159648529	168	0	0.0000	1.0000
3	159622566	159672566	204	0	0.0000	1.0000
3	159644711	159694711	221	0	0.0000	1.0000
3	188087479	188137479	194	0	0.0000	1.0000
4	123045809	123095809	120	3	**0.0250**	0.0827

4	123095809	123145809	103	6	**0.0583**	**0.0359**
4	123145809	123195809	93	3	0.0323	0.0672
4	123195809	123245809	78	1	0.0128	0.1326
4	123245809	123295809	92	2	**0.0217**	0.0933
4	123295809	123345809	128	3	**0.0234**	0.0869
4	123345809	123395809	131	4	**0.0305**	0.0702
4	123395809	123445809	119	2	**0.0168**	0.1115
4	123445809	123495809	153	3	**0.0196**	0.1006
4	123495809	123545809	175	1	0.0057	0.2287
6	371321	421321	253	0	0.0000	1.0000
6	384248	434248	234	4	**0.0171**	0.1107
6	90789396	90839396	191	0	0.0000	1.0000
6	128256931	128306931	180	0	0.0000	1.0000
6	128267381	128317381	179	0	0.0000	1.0000
6	137957870	138007870	302	0	0.0000	1.0000
6	137963083	138013083	285	0	0.0000	1.0000
6	159443018	159493018	199	1	0.0050	0.2527
6	159473267	159523267	174	18	**0.1034**	**0.0151**
7	37394173	37444173	250	1	0.0040	0.3029
8	129248883	129298883	265	0	0.0000	1.0000
10	6365192	6415192	187	2	**0.0107**	0.1513
10	81033027	81083027	244	0	0.0000	1.0000
11	111218889	111268889	171	0	0.0000	1.0000
11	118552596	118602596	386	1	0.0026	0.3511
11	128363856	128413856	149	3	**0.0201**	0.0988
12	111871182	111921182	88	0	0.0000	1.0000
12	111921182	111971182	115	0	0.0000	1.0000
12	111971182	112021182	102	0	0.0000	1.0000
14	69240688	69290688	241	30	**0.1245**	**0.0113**
15	75022171	75072171	214	0	0.0000	1.0000
16	10961194	11011194	214	0	0.0000	1.0000
16	11329125	11379125	236	1	0.0042	0.2908
16	11348797	11398797	274	1	0.0036	0.3183
16	11359956	11409956	287	1	0.0035	0.3255
18	12826055	12876055	202	1	0.0050	0.2560
18	12832758	12882758	215	1	0.0047	0.2696
21	43830067	43880067	244	0	0.0000	1.0000
21	45600469	45650469	330	36	**0.1091**	**0.0136**
22	21927081	21977081	148	1	0.0068	0.2046
	Totals		13789	142	**0.0103**	
	Totals (no repeat SNPs)		12129	138	**0.0114**	

Part 4. Within Europe Fst Count-Ratio Results

Results from this study.			Europe (CEU x TSI)			
Chr.	**hg19 Start**	**hg19 End**	**# SNPs**	**# > 99%**	**Ratio**	**P-Value**
Partial Coordinate Liftover Failure						
1	25272143	25322143	288	0	0.0000	1.0000
1	172652828	172702828	179	0	0.0000	1.0000
1	172838818	172888818	184	0	0.0000	1.0000
1	192487559	192537559	192	0	0.0000	1.0000
1	192507482	192557482	236	0	0.0000	1.0000
1	200856392	200906392	247	0	0.0000	1.0000
2	61170843	61220843	143	0	0.0000	1.0000
2	68617639	68667639	323	0	0.0000	1.0000
2	102957473	103007473	316	0	0.0000	1.0000
2	103007473	103057473	253	0	0.0000	1.0000
2	103057473	103107473	321	0	0.0000	1.0000
2	181972524	182022524	212	0	0.0000	1.0000
2	182022524	182072524	231	0	0.0000	1.0000
2	182072524	182122524	161	0	0.0000	1.0000
2	191899034	191949034	168	0	0.0000	1.0000
2	191923637	191973637	163	1	0.0061	0.1977
2	191948563	191998563	184	1	0.0054	0.2098
2	204430119	204480119	241	5	**0.0207**	0.0985
2	204688462	204738462	184	0	0.0000	1.0000
2	204741191	204791191	174	0	0.0000	1.0000
3	33012737	33062737	270	0	0.0000	1.0000
3	46171697	46221697	229	0	0.0000	1.0000
3	46349449	46399449	273	0	0.0000	1.0000
3	46469473	46519473	277	0	0.0000	1.0000
3	119095888	119145888	257	0	0.0000	1.0000
3	159598529	159648529	170	0	0.0000	1.0000
3	159622566	159672566	228	0	0.0000	1.0000
3	159644711	159694711	272	0	0.0000	1.0000
3	188087479	188137479	204	0	0.0000	1.0000
4	123045809	123095809	216	0	0.0000	1.0000
4	123095809	123145809	168	2	0.0119	0.1381
4	123145809	123195809	149	2	0.0134	0.1281
4	123195809	123245809	149	0	0.0000	1.0000
4	123245809	123295809	153	0	0.0000	1.0000
4	123295809	123345809	175	0	0.0000	1.0000
4	123345809	123395809	199	2	**0.0101**	0.1525

4	123395809	123445809	210	1	0.0048	0.2260
4	123445809	123495809	184	0	0.0000	1.0000
4	123495809	123545809	352	2	0.0057	0.2052
6	371321	421321	352	2	0.0057	0.2052
6	384248	434248	310	2	0.0065	0.1931
6	90789396	90839396	185	0	0.0000	1.0000
6	128256931	128306931	206	0	0.0000	1.0000
6	128267381	128317381	191	0	0.0000	1.0000
6	137957870	138007870	311	0	0.0000	1.0000
6	137963083	138013083	290	0	0.0000	1.0000
6	159443018	159493018	241	0	0.0000	1.0000
6	159473267	159523267	205	0	0.0000	1.0000
7	37394173	37444173	258	0	0.0000	1.0000
8	129248883	129298883	227	0	0.0000	1.0000
10	6365192	6415192	209	0	0.0000	1.0000
10	81033027	81083027	272	8	**0.0294**	0.0775
11	111218889	111268889	225	1	0.0044	0.2371
11	118552596	118602596	344	2	0.0058	0.2030
11	128363856	128413856	164	0	0.0000	1.0000
12	111871182	111921182	97	0	0.0000	1.0000
12	111921182	111971182	91	0	0.0000	1.0000
12	111971182	112021182	103	0	0.0000	1.0000
14	69240688	69290688	233	0	0.0000	1.0000
15	75022171	75072171	232	8	**0.0345**	0.0693
16	10961194	11011194	233	0	0.0000	1.0000
16	11329125	11379125	288	1	0.0035	0.2799
16	11348797	11398797	306	1	0.0033	0.2873
16	11359956	11409956	321	0	0.0000	1.0000
18	12826055	12876055	248	0	0.0000	1.0000
18	12832758	12882758	264	0	0.0000	1.0000
21	43830067	43880067	324	0	0.0000	1.0000
21	45600469	45650469	376	0	0.0000	1.0000
22	21927081	21977081	175	0	0.0000	1.0000
	Totals		15816	41	0.0026	
	Totals (no repeat SNPs)		13770	35	0.0025	

Part 5. Between Africa and East Asia Fst Count-Ratio Results

Results from this study.			**Africa (LWK,YRI) x East Asia (CHB,CHS,JPT)**			
Chr.	**hg19 Start**	**hg19 End**	**# SNPs**	**# > 99%**	**Ratio**	**P-Value**
Partial Coordinate Liftover Failure						
1	25272143	25322143	567	2	0.0035	0.4157

1	172652828	172702828	507	23	**0.0454**	0.0506
1	172838818	172888818	451	40	**0.0887**	**0.0105**
1	192487559	192537559	431	0	0.0000	1.0000
1	192507482	192557482	526	3	0.0057	0.3514
1	200856392	200906392	502	1	0.0020	0.4787
2	61170843	61220843	371	12	**0.0323**	0.0903
2	68617639	68667639	707	16	**0.0226**	0.1397
2	102957473	103007473	565	72	**0.1274**	**0.0035**
2	103007473	103057473	522	48	**0.0920**	**0.0095**
2	103057473	103107473	625	7	**0.0112**	0.2484
2	181972524	182022524	480	0	0.0000	1.0000
2	182022524	182072524	495	0	0.0000	1.0000
2	182072524	182122524	424	1	0.0024	0.4588
2	191899034	191949034	450	7	**0.0156**	0.2004
2	191923637	191973637	473	3	0.0063	0.3359
2	191948563	191998563	493	1	0.0020	0.4759
2	204430119	204480119	539	0	0.0000	1.0000
2	204688462	204738462	449	0	0.0000	1.0000
2	204741191	204791191	465	0	0.0000	1.0000
3	33012737	33062737	580	12	**0.0207**	0.1527
3	46171697	46221697	471	0	0.0000	1.0000
3	46349449	46399449	607	1	0.0016	0.5238
3	46469473	46519473	552	1	0.0018	0.5004
3	119095888	119145888	561	0	0.0000	1.0000
3	159598529	159648529	482	4	0.0083	0.2945
3	159622566	159672566	590	0	0.0000	1.0000
3	159644711	159694711	589	0	0.0000	1.0000
3	188087479	188137479	583	3	0.0051	0.3660
4	123045809	123095809	474	1	0.0021	0.4696
4	123095809	123145809	390	1	0.0026	0.4530
4	123145809	123195809	390	0	0.0000	1.0000
4	123195809	123245809	347	0	0.0000	1.0000
4	123245809	123295809	353	0	0.0000	1.0000
4	123295809	123345809	439	3	0.0068	0.3247
4	123345809	123395809	469	3	0.0064	0.3354
4	123395809	123445809	459	2	0.0044	0.3883
4	123445809	123495809	478	2	0.0042	0.3928
4	123495809	123545809	488	4	0.0082	0.2969
6	371321	421321	619	0	0.0000	1.0000
6	384248	434248	567	1	0.0018	0.5081
6	90789396	90839396	473	5	**0.0106**	0.2568

6	128256931	128306931	542	0	0.0000	1.0000
6	128267381	128317381	538	0	0.0000	1.0000
6	137957870	138007870	637	1	0.0016	0.5343
6	137963083	138013083	619	1	0.0016	0.5281
6	159443018	159493018	535	7	**0.0131**	0.2250
6	159473267	159523267	563	5	0.0089	0.2835
7	37394173	37444173	580	0	0.0000	1.0000
8	129248883	129298883	719	1	0.0014	0.2835
10	6365192	6415192	620	13	**0.0210**	0.1502
10	81033027	81083027	546	2	0.0037	0.4105
11	111218889	111268889	519	0	0.0000	1.0000
11	118552596	118602596	630	0	0.0000	1.0000
11	128363856	128413856	441	0	0.0000	1.0000
12	111871182	111921182	305	49	**0.1607**	**0.0018**
12	111921182	111971182	320	39	**0.1219**	**0.0040**
12	111971182	112021182	334	43	**0.1287**	**0.0034**
14	69240688	69290688	498	2	0.0040	0.3981
15	75022171	75072171	443	0	0.0000	1.0000
16	10961194	11011194	653	4	0.0061	0.3409
16	11329125	11379125	639	11	**0.0172**	0.1851
16	11348797	11398797	708	13	**0.0184**	0.1733
16	11359956	11409956	713	17	**0.0238**	0.1320
18	12826055	12876055	576	0	0.0000	1.0000
18	12832758	12882758	596	0	0.0000	1.0000
21	43830067	43880067	746	2	0.0027	0.4495
21	45600469	45650469	749	7	0.0093	0.2765
22	21927081	21977081	494	0	0.0000	1.0000
	Totals		36266	496	**0.0137**	
	Totals (no repeat SNPs)		31819	470	**0.0148**	

Part 6. Within Africa Fst Count-Ratio Results

Results from this study.			**Africa (LWK,YRI) x Europe (CEU,FIN,GBR,IBS,TSI)**			
Chr.	**hg19 Start**	**hg19 End**	**# SNPs**	**# > 99%**	**Ratio**	**P-Value**
Partial Coordinate Liftover Failure						
1	25272143	25322143	568	8	**0.0141**	0.1325
1	172652828	172702828	472	3	0.0064	0.2385
1	172838818	172888818	445	5	**0.0112**	0.1597
1	192487559	192537559	457	0	0.0000	1.0000
1	192507482	192557482	524	6	**0.0115**	0.1577
1	200856392	200906392	494	1	0.0020	0.3870
2	61170843	61220843	381	4	**0.0105**	0.1697

2	68617639	68667639	707	3	0.0042	0.2933
2	102957473	103007473	571	0	0.0000	1.0000
2	103007473	103057473	525	0	0.0000	1.0000
2	103057473	103107473	638	0	0.0000	1.0000
2	181972524	182022524	479	0	0.0000	1.0000
2	182022524	182072524	511	0	0.0000	1.0000
2	182072524	182122524	427	0	0.0000	1.0000
2	191899034	191949034	454	19	**0.0419**	**0.0285**
2	191923637	191973637	446	13	**0.0291**	0.0574
2	191948563	191998563	467	4	0.0086	0.1989
2	204430119	204480119	542	0	0.0000	1.0000
2	204688462	204738462	453	0	0.0000	1.0000
2	204741191	204791191	444	0	0.0000	1.0000
3	33012737	33062737	565	2	0.0035	0.3190
3	46171697	46221697	471	5	**0.0106**	0.1682
3	46349449	46399449	602	1	0.0017	0.4371
3	46469473	46519473	556	5	0.0090	0.1932
3	119095888	119145888	544	8	**0.0147**	0.1275
3	159598529	159648529	463	0	0.0000	1.0000
3	159622566	159672566	567	0	0.0000	1.0000
3	159644711	159694711	591	0	0.0000	1.0000
3	188087479	188137479	569	14	**0.0246**	0.0714
4	123045809	123095809	500	0	0.0000	1.0000
4	123095809	123145809	395	0	0.0000	1.0000
4	123145809	123195809	402	0	0.0000	1.0000
4	123195809	123245809	354	0	0.0000	1.0000
4	123245809	123295809	375	0	0.0000	1.0000
4	123295809	123345809	450	10	**0.0222**	0.0801
4	123345809	123395809	469	13	**0.0277**	0.0609
4	123395809	123445809	494	17	**0.0344**	**0.0416**
4	123445809	123495809	451	16	**0.0355**	**0.0388**
4	123495809	123545809	480	15	**0.0313**	0.0513
6	371321	421321	627	7	**0.0112**	0.1611
6	384248	434248	586	7	**0.0119**	0.1528
6	90789396	90839396	459	1	0.0022	0.3752
6	128256931	128306931	517	0	0.0000	1.0000
6	128267381	128317381	511	0	0.0000	1.0000
6	137957870	138007870	637	0	0.0000	1.0000
6	137963083	138013083	624	0	0.0000	1.0000
6	159443018	159493018	535	0	0.0000	1.0000
6	159473267	159523267	553	1	0.0018	0.4148

7	37394173	37444173	588	1	0.0017	0.4315
8	129248883	129298883	702	1	0.0014	0.4652
10	6365192	6415192	618	11	**0.0178**	0.1029
10	81033027	81083027	536	0	0.0000	1.0000
11	111218889	111268889	537	0	0.0000	1.0000
11	118552596	118602596	586	0	0.0000	1.0000
11	128363856	128413856	440	0	0.0000	1.0000
12	111871182	111921182	324	56	**0.1728**	**0.0005**
12	111921182	111971182	318	46	**0.1447**	**0.0007**
12	111971182	112021182	337	50	**0.1484**	**0.0005**
14	69240688	69290688	486	1	0.0021	0.3838
15	75022171	75072171	462	11	**0.0238**	0.0735
16	10961194	11011194	639	12	**0.0188**	0.0975
16	11329125	11379125	638	0	0.0000	1.0000
16	11348797	11398797	703	0	0.0000	1.0000
16	11359956	11409956	716	0	0.0000	1.0000
18	12826055	12876055	573	9	**0.0157**	0.1186
18	12832758	12882758	594	10	**0.0168**	0.1101
21	43830067	43880067	749	3	0.0040	0.3005
21	45600469	45650469	752	7	0.0093	0.1873
22	21927081	21977081	511	21	**0.0411**	**0.0295**
	Totals		36161	427	**0.0118**	
	Totals (no repeat SNPs)		31782	399	**0.0126**	

Part 7. Within Africa Fst Count-Ratio Results

Results from this study.			**East Asia (CHB,CHS,JPT) x Europe (CEU,FIN,GBR,IBS,TSI)**			
Chr.	**hg19 Start**	**hg19 End**	**# SNPs**	**# > 99%**	**Ratio**	**P-Value**
Partial Coordinate Liftover Failure						
1	25272143	25322143	421	0	0.0000	1.0000
1	172652828	172702828	346	12	**0.0347**	0.0618
1	172838818	172888818	325	65	**0.2000**	0.0009
1	192487559	192537559	319	0	0.0000	1.0000
1	192507482	192557482	375	0	0.0000	1.0000
1	200856392	200906392	374	2	0.0053	0.1935
2	61170843	61220843	258	9	**0.0349**	0.0615
2	68617639	68667639	491	0	0.0000	1.0000
2	102957473	103007473	450	0	0.0000	1.0000
2	103007473	103057473	401	0	0.0000	1.0000
2	103057473	103107473	498	1	0.0020	0.2811
2	181972524	182022524	367	0	0.0000	1.0000
2	182022524	182072524	373	0	0.0000	1.0000

2	182072524	182122524	309	0	0.0000	1.0000
2	191899034	191949034	293	0	0.0000	1.0000
2	191923637	191973637	284	0	0.0000	1.0000
2	191948563	191998563	321	0	0.0000	1.0000
2	204430119	204480119	365	6	**0.0164**	0.1167
2	204688462	204738462	334	0	0.0000	1.0000
2	204741191	204791191	335	1	0.0030	0.2322
3	33012737	33062737	383	15	**0.0392**	0.0531
3	46171697	46221697	343	0	0.0000	1.0000
3	46349449	46399449	417	0	0.0000	1.0000
3	46469473	46519473	406	0	0.0000	1.0000
3	119095888	119145888	432	0	0.0000	1.0000
3	159598529	159648529	345	5	**0.0145**	0.1257
3	159622566	159672566	411	0	0.0000	1.0000
3	159644711	159694711	440	0	0.0000	1.0000
3	188087479	188137479	347	0	0.0000	1.0000
4	123045809	123095809	324	0	0.0000	1.0000
4	123095809	123145809	275	0	0.0000	1.0000
4	123145809	123195809	268	0	0.0000	1.0000
4	123195809	123245809	237	0	0.0000	1.0000
4	123245809	123295809	262	0	0.0000	1.0000
4	123295809	123345809	324	0	0.0000	1.0000
4	123345809	123395809	330	0	0.0000	1.0000
4	123395809	123445809	348	0	0.0000	1.0000
4	123445809	123495809	343	0	0.0000	1.0000
4	123495809	123545809	327	0	0.0000	1.0000
6	371321	421321	499	0	0.0000	1.0000
6	384248	434248	456	0	0.0000	1.0000
6	90789396	90839396	347	18	**0.0519**	**0.0343**
6	128256931	128306931	372	7	**0.0188**	0.1073
6	128267381	128317381	352	5	**0.0142**	0.1265
6	137957870	138007870	474	0	0.0000	1.0000
6	137963083	138013083	457	0	0.0000	1.0000
6	159443018	159493018	370	9	**0.0243**	0.0897
6	159473267	159523267	352	21	**0.0597**	**0.0269**
7	37394173	37444173	403	19	**0.0471**	**0.0403**
8	129248883	129298883	424	6	**0.0142**	0.1267
10	6365192	6415192	341	4	**0.0117**	0.1394
10	81033027	81083027	403	0	0.0000	1.0000
11	111218889	111268889	389	0	0.0000	1.0000
11	118552596	118602596	528	0	0.0000	1.0000

11	128363856	128413856	303	0	0.0000	1.0000
12	111871182	111921182	200	21	**0.1050**	**0.0086**
12	111921182	111971182	226	23	**0.1018**	**0.0090**
12	111971182	112021182	209	21	**0.1005**	**0.0094**
14	69240688	69290688	382	0	0.0000	1.0000
15	75022171	75072171	362	7	**0.0193**	0.1058
16	10961194	11011194	393	28	**0.0712**	**0.0206**
16	11329125	11379125	440	23	**0.0523**	**0.0337**
16	11348797	11398797	477	24	**0.0503**	**0.0359**
16	11359956	11409956	475	32	**0.0674**	**0.0218**
18	12826055	12876055	365	0	0.0000	1.0000
18	12832758	12882758	389	0	0.0000	1.0000
21	43830067	43880067	498	1	0.0020	0.2811
21	45600469	45650469	536	0	0.0000	1.0000
22	21927081	21977081	318	0	0.0000	1.0000
	Totals		25541	385	**0.0151**	
	Totals (no repeat SNPs)		22494	347	**0.0154**	