3-2-2014

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

Aaron Sams
Cornell University, as2847@cornell.edu

John Hawks
University of Wisconsin-Madison

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

Aaron Sams[1]* and John Hawks[2]

[1]Cornell University

[2]University of Wisconsin-Madison

*Correspondence to: Aaron Sams, Department of Biological Statistics and Computational Biology, Cornell University, 102B Weill Hall, Ithaca, NY 14853. E-mail: as2847@cornell.edu.

Key words: Celiac disease, gene networks, complex traits, natural selection.

short title: Celiac Disease and Evolution of Disease in Humans

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.

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), and during the early twentieth century some workers noted the value of a starch free diet for prevention of CD (Haas 1924). Dicke and

colleagues (1950, 1953) observed the detrimental effect of wheat flour (but not wheat starch) on CD patients experimentally confirming a role for diet as a driver of CD. 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. Only since the 1950s have clinicians recognized that a gluten-free diet is an effective treatment (Barker and Liu 2008). Before that time, diagnosed and undiagnosed CD reduced the health and fitness of sufferers, with juvenile cases leading to malnutrition or death, and adult cases causing wasting, lack of adequate nutrition, greater susceptibility to infection, and direct reductions in fertility (Corrao et al. 2001; Soni and Badawy 2010). Despite these apparent costs, CD is common today (>1%) in several populations with long histories of wheat agriculture, including Europe, the Near East, and North Africa (Catassi et al. 1996; Corrao et al. 1996; Mäki et al. 2003; Rubio-Tapia et al. 2012). CD occurs at even higher frequencies in some parts of these regions, reaching 5–6% among the Saharawi of northern Africa (Catassi et al. 1999, 2001). There is little evidence that the present incidence of CD in these populations can be explained by shifts in the environment, such as the "hygiene effect" of lower pathogen and parasite loads that has been suggested to explain increases in allergies and asthma. Both historical and archaeological evidence show that CD was present in these populations before industrialization. Recent

diet shifts across much of this region have likely decreased the role of dietary gluten, due to the adoption of maize, rice and other staple alternatives to wheat, and more widespread use of refined flours with low gluten content.

The paradox of high CD incidence in these populations despite the evident fitness costs seems to require some evolutionary explanation. Similarly, common human disorders that have a fitness cost in one or more populations, such as variegate porphyria, sickle-cell anemia, and cystic fibrosis, have been explained either as a result of selective balances or founder effects. In contrast to such examples, in which a single locus is implicated, CD is genetically complex. The additive variance underlying CD risk can so far be attributed to more than 40 genetic loci identified in genome-wide association or case-control studies. The strongest of these associations are in the Human Leukocyte Antigen (HLA) region of the genome, which much evidence shows to be subject to recent natural selection (Albrechtsen et al. 2010). But the larger network of more than 40 other risk loci presents a potentially much more complex story than the well-known single-gene examples of founder effects and balancing selection.

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.

Presentation, Pathophysiology, and Genetics of CD

Clinical Presentation and Treatment. The clinical presentation of CD is highly variable and depends on age (Fasano 2005; Barker and Liu 2008). Classical symptoms include abdominal pain and distension, diarrhea, malnutrition, and failure to thrive within the first few years of life (Barker and Liu 2008). CD is often recognized based on its characteristic damage to the mucosal lining of the small intestine. The scoring system developed by Marsh (1992) 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. Recent diagnoses have been assessed using the presence of tissue transglutaminase (TGase) autoantibodies, such as immunoglobulin A (IgA), and then confirmed after intestinal biopsy by response to treatment with a gluten-free diet (GFD) (Farrell and Kelly 2002).

The complications of CD go far beyond the classical symptoms into other aspects of biology and health. CD can cause vitamin (D and K) and micronutrient (calcium, iron, folate) deficiencies, with prolonged cases often leading to conditions such as rickets, osteoporosis, and anemia (Farrell and Kelly 2002; Fasano 2005; Green and Cellier 2007). Linear enamel hypoplasias on the teeth, a marker of nutritional stress, 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 et al. 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). As noted above, complications from CD may also include reduced fertility and increased mortality (Corrao et al. 2001; Soni and Badawy 2010).

Celiac disease is unusually common among certain autoimmune disease cohorts. Several researchers have noted a high occurrence of CD in patients diagnosed with type 1 diabetes mellitus (T1DM) (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 conditions share with CD an association with risk haplotypes HLA-DQ2 and HLA-DQ8 (Ch'ng et al. 2007; Pham Short et al. 2012). In addition to T1DM, shared (non-HLA) genetic risk factors with CD have been noted for Crohn's disease (CrD) (Festen et al. 2011) and 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 interpretation of CD as an evolutionary paradox. Prior to the recent discovery of gluten as the

primary environmental driver of 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 wheat gluten and related proteins are clearly the primary environmental trigger of CD, diet alone is not a sufficient explanation for the occurrence of the disease (Green and Cellier 2007; Kagnoff 2007). The genetic factors that influence pathogenesis of CD became evident and have been confirmed by various family and twin studies of CD (Risch 1987; Petronzelli et al. 1997; Bevan et al. 1999; Greco 2002; Nistico 2006). This work has securely established HLA genes as the largest genetic components of CD risk. 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 findings supported 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 comprising HLA-DQ. 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 other 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. Not all CD patients respond to the same sets of peptides (Koning 2012), which renders difficult the development of non-immunogenic grains (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 (IELs) 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 IELs 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. Although HLA comprises only 40% of the genetic variance of CD risk, no compelling non-HLA loci were shown to be associated with CD prior to the advent of genome-wide association studies (GWAS) (Kumar et al. 2012). The first GWA 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 identified a single risk locus in an LD block harboring IL2-IL21. Two follow up studies on the same sample (Hunt et al. 2008; Trynka et al. 2009) identified an additional nine non-HLA 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 13 risk loci to the list of CD associated genomic regions. Importantly,

this study also demonstrated that nearly 50% of CD risk SNPs are in close proximity to expression quantitative trait loci (eQTLs), suggesting a likely role for changes in levels of gene expression in CD. As mentioned above, additional GWA studies have identified risk loci shared between CD and other autoimmune conditions such as T1DM, RA, CrD, and ulcerative colitis (UC) (Zhernakova et al. 2007, 2011; Smyth et al. 2008; Barrett et al. 2009; Festen et al. 2009, 2011; Cotsapas et al. 2011; Gutierrez-Achury et al. 2011). These studies have illuminated the tremendously pleiotropic nature of loci underlying autoimmune and other immune related conditions. For a comprehensive review of risk allele sharing among autoimmune diseases, see Ricaño-Ponce and Wijmenga (2013).

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 12 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, for which alleles of

different SNPs in the same genomic region contribute independently to risk across CD cases. This brought the total number of non-HLA CD associated loci to 39 with 57 independently associated SNPs and 29 signals confined to single genes.

Genetic Architecture and Heritability. 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 narrow-sense heritability (h^2) of CD at between 57% and 87%, with variation in the h^2 estimate due to the assumed population prevalence (1/1000 and 1/91, respectively). (Nistico 2006). This estimate was not age specific. Rather, mean age at diagnosis of twins in this study ranged from 0.5 to 57 years. However, 50% of twins were diagnosed before age three. Further, this study did not explicitly control for the presence of gluten in the diet. The fine-mapping study by Trynka and colleagues (2011) has brought the total percentage of that genetic variance currently explained to ~53.7%, 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). Still, widespread support for a major role of low-frequency variants remains unsupported. The dense genotyping Immunochip analysis by Trynka and colleagues (2011) identified lower frequency variants (< 5%) at four loci (CD28-CTLA4-ICOS, PTPN2, RGS1, SOCS1-PRM1-PRM2), all of low effect size (OR < 1.7). An exome sequencing study of a family of six segregating for CD failed to identify any rare causal variants (Szperl et al. 2011). Finally, in a recent high-coverage exome sequencing study of 20 risk loci with overlapping associations to six autoimmune diseases in 25 individuals, Hunt and colleagues (2013) estimate that rare variants contribute to less than 3% of the heritability explained by common variants at the risk loci included in the study. Though rare-variants in coding regions appear to contribute little to disease risk, exploration of non-coding regions may prove more fruitful. The statistical factors making identification of rare causal variants with larger effect sizes also 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 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.

Environmental Risk Factors. 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.

While ingestion of gluten and related proteins are well established as the driver of CD, the fact that such a small percentage of those with increased genetic

risk actually develop the condition suggests that other environmental triggers are likely involved in the pathogenesis of CD.

Much focus has been placed on understanding the relationship(s) between the human gut microbiota and disease pathogenesis (Fujimura et al. 2010). Recently, study of the human gut microbiome has been applied to CD pathogenesis (Sellito et al. 2012; Cheng et al. 2013; Sjöberg et al. 2013). Almost a decade ago, Forsberg and colleagues (2004) noted the presence of rod-shaped bacteria in CD patients, but not control, suggesting a potential role for bacterial infection in breaking gluten tolerance in children. More recently, a prospective study by Sellitto and colleagues (2012) tested the hypothesis that the entire gut microbial ecosystem is involved in the switch from gluten tolerance to autoimmune response in genetically susceptible individuals. The authors tracked the changes in gut microbiome composition from birth to 24 months in 34 infants with HLA-DQ2 and/or HLA-DQ8 risk genotypes and known to be first-degree relatives of biopsy proven CD patients. The infants were randomly assigned to two groups; one fed a gluten-free diet from 6-12 months, the other a gluten-included diet. All infants resumed normal feeding at 12 months. The sample included 13 infants from each group that completed all protocol, 8 infants from each group were chosen randomly for final analysis. The researchers show a significant increase in anti-gliadin antibodies (AGA), specifically immunoglobulin G (IgG) in the early exposure group compared to the late

exposure group after normalizing by time of exposure. Additionally, they show that the gut microbiota of at-risk infants were lacking in members of the phylum Bacteroidetes and had higher abundance of the phylum Fermicutes, a composition directly opposed to that of non at-risk infants. An earlier study by Palmer and colleagues (2007) illustrated that in healthy infants, the microbial ecosystem reaches a profile similar to that of adults within the first year of life. In contrast, Sellitto and colleagues (2012) find that the gut microbiota of infants at risk for CD maintain immature gut microbiota, characterized by low levels of Bacteroidetes and high levels of Firmicutes, throughout the first two years of life. The authors show that delayed exposure to gluten had an overall positive effect on prolonging gluten tolerance and delaying onset of CD autoimmunity. They hypothesize that the introduction of gluten in an immature gut microbiome could trigger the development of autoimmunity.

In contrast, a slightly later study by Cheng and colleagues (2013) found no significant differences in the overall composition and diversity of gut microbiota (including Bacteroidetes) between a sample of 10 children with newly diagnosed CD (aged 3–14) and 9 healthy controls (aged 4–16). Thus, the differences observed by Sellitto and colleagues (2012) may not persist into later childhood. Cheng and colleagues (2013), however, did identify a sub-population profile of eight genus-like bacteria that differed significantly between the CD cases and healthy controls.

Others have noted a relationship between gut bacteria and CD development (Forsberg et al. 2004; Ou et al. 2009; Sanz et al. 2011; Sjöberg et al. 2013). Most recently, Sjöberg and colleagues (2013) demonstrate a relationship between IL-17A, a proinflammatory cytokine involved in antibacterial defense, gut microbiome composition, and CD pathogenesis. The authors suggest that both gluten and CD associated bacteria, such as the rod-shaped bacterium *Lachnoanaerobaculum umeaense* provoke IL-17A production in the intestinal mucosa of CD patients. They argue that the magnitude of the IL-17A reaction to gluten is largely influenced by the makeup of the gut microbiota, particularly the amount of CD associated bacteria present. In line with the results of Sellitto and colleagues (2012) mentioned above, these authors posit that conditions leading to disturbances in the composition of the gut microbiota early in life can lead to long-term dysbiosis with surpluses of CD associated bacteria. These abnormal microbiota compositions may influence the amount of IL-17A responses to gluten and constitute a major risk factor for contraction of CD in genetically at risk children.

While investigation of the human gut microbiota has revealed a potential role for variation in gut associated bacteria in CD development, the connection between genes and environmental risk factors have provided no clear understanding of the breakdown of gluten tolerance leading to downstream tissue damage.

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 Risk

The CD Paradox. 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). Human populations consume the cereals that can induce CD in widely varying amounts (Abadie et al. 2011). CD is more common among populations with European and Near Eastern ancestry, even though these are precisely the populations with the longest history of wheat, rye, and barley agriculture (Tresset and Vigne 2011; Zeder 2011).

Although the analogy is not entirely apt, we can contrast the pattern of CD with the global pattern of lactose intolerance (lactase persistence), in which the populations with the longest history of dairying and milk consumption are those with the greatest adaptation to lactose digestion.

The population incidence of CD increases with age, but 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.

2. CD risk alleles may have been increased in frequency by positive selection due to pleiotropic effects on other, non-CD phenotypes.

3. The genetic architecture of CD may be such that most loci contributing to CD risk have only minor effects on risk, meaning that selection against CD would have miniscule to nonexistent effects on any single risk locus (Stranger et al. 2011).

4. 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 the advantage of heterozygous individuals in resisting malaria. Recognition of this evolutionary history was facilitated by the geographic association of hemoglobinopathies with historic malaria incidence, and the clear fitness costs to homozygotes in both the presence and absence of malaria.

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. In fact, many HLA Class II alleles pre-date

the divergence of all hominids (Humans & Apes) (Bergström and Gyllensten 1995). In fact, Bergström and Gyllensten (1995) found that DQB1 and DQA1 in hominids possess some of the oldest polymorphism of all class II genes according to the ratio of nonsynonymous to synonymous polymorphism at non-antigen recognition site (ARS) codons. While the DQ loci appear to possess some of the oldest allelic variation of all class II genes, they also maintain a relatively small number of alleles indicating long-term balancing selection in hominids.

This pattern holds within the human species as well. In the largest sequence study of the HLA complex in humans to date, Buhler and Sanchez-Mazas (2011) analyzed 2,062 DNA sequences representing seven HLA genes (A, B, Cw, DRB1, DQA1, DQB1, DPB1) in a sample of 23,500 individuals from 200 worldwide populations. The researchers found that DQA1 and DQB1 have the highest percentage of significant Tajima's D values across human populations. DQA1 in particular has significantly positive values of Tajima's D across nearly every European sample in the study, suggesting the maintenance of allelic variation due to some type of balancing or diversifying selection. Relative to most other HLA loci, DQB1 alleles are less diverse in pairwise comparisons of populations within regions. Buhler and Sanchez-Mazas (2011) suggest that this lack of allelic diversity may reflect purifying selection on DQB1 due to structural constraints on the formation of HLA-DQ heterodimers, but this hypothesis has yet to be formally tested.

An earlier study by Solberg and colleagues (2008) found a similar result after examining the Ewens-Watterson test of homozygosity in 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. The primary causal CD heterodimer (DQ2) is present at substantial frequencies (>1%) in many populations worldwide but reaches frequencies as high as 25% in populations of North Africa and Europe where CD is most common (Traherne 2008; Gonzalez-Galarza et al. 2011). The relatively high frequency of this heterodimer in Europeans may be, in part, due to recent expansion of the COX long-range HLA haplotype approximately 25,000 years ago (Smith et al. 2006; Traherne 2008).

Some researchers have argued that the cause of balancing selection in HLA is overdominance (Takahata and Nei 1990; Takahata et al. 1992) within environments with particular pathogen communities. 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 may in part explain the occurrence of CD in European and North African populations today.

Under this hypothesis, CD is a fitness-reducing side effect of selection associated with pathogens. 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 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. 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) each contributes to CD risk only conditioned upon the presence of the major HLA risk alleles, and (b) each makes only a small contribution to CD risk. 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.

GWAS have provided nearly all the links between non-HLA loci and 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 strictly 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.

In contrast to functional integration, *evolutionary integration* provides an alternative way of describing genetic systems. When different genes respond together to a single environmental pressure, independent or semi-distinct

functional pathways may come to exhibit patterns of similarity, either in genetic variation or differentiation. Evolutionary integration 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, 2010b, 2011).

To the extent that a trait involves related functional networks of genes, both the environment and genetic architecture are necessary to drive evolutionary integration. 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 is similar to an exponential model of effect size, in which the non-HLA loci make increasingly tiny contributions to the additive variance in risk. 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) (Figure 1). In this model, we expect that HLA should bear the majority of the negative fitness effects of CD. Indeed, any response to selection associated with CD will affect HLA genetic variation 40% / 14% = 2.86 times more than all of the other background loci 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 et al. 2002). However, Sams and Hawks (2013) found that this haplotype is present in the genome of Ötzi, the Tyrolean Iceman, which dates to 5,300 years ago (Keller et al. 2012). The apparent antiquity of the allele does not necessarily conflict with evidence for recent strong selection, if a pre-existing rare allele was selected within the last 2,000 years, but the apparent contradiction may also be explained by the wide variance in estimates of time of onset of selection. A broader comparison of genes involved in the CD risk network showed a low heterozygosity for these loci in the Ötzi genome, suggesting that many of the risk loci most tightly associated with CD risk today are not present in this ancient individual. Along these lines, Fumagalli and colleagues (2011) show, after correcting for demography, that many non-HLA genes associated with autoimmune disease are significantly correlated with pathogenic diversity. These genes include nine previously associated with CD risk, further supporting a role for local adaptation to pathogens in shaping CD risk.

While the past 10,000 years was a time of significant selection pressures on the immune system, the present risk pattern of CD may also reflect earlier events. 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 (Sams and Hawks 2013). Genes contributing to CD risk stand out against the overall genome-wide pattern of population structure, specifically when comparing genetic samples from Europe against East Asia. Within continents, African populations show relatively high differentiation of the CD risk network compared to the genome-wide pattern. Two samples from China also diverged more highly than the genome-wide expectation. These results point to the possibility of selection affecting some of the genes that contribute to CD risk today, at the time of differentiation of European and Asian populations, and again within Africa. Although as noted above some loci show evidence based on iHS for recent selection in Europe, this does not drive differentiation of CD risk genes as a group relative to the rest of the genome (Sams and Hawks 2013). This observation weakens the hypothesis that the broad network of CD risk loci has experienced directional selection within the last 10,000 years, although 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), CD risk loci may have been

selected early enough in the history of this demographic event to have spread to all present European populations

These population comparisons point not to evolutionary integration of the CD risk network, but instead to a heterogeneity of environmental influences in past populations. The CD background risk network was affected by selection at different times and in different populations, with no uniformity or consistency. Further analysis of this pattern with explicit reference to effect sizes of CD-associated loci confirms the lack of systematic integration, as the signal of selection at a CD-associated locus does not predict the effect-size of the locus on present risk (Figure 2). Hence, although the joint association network underlying CD risk in European populations today has not experienced the same selection dynamics as the rest of the genome, these genes have responded to different environments, not a single Neolithic pulse of adaptation. With respect to the background of risk outside of the HLA system, CD is not an evolutionary tradeoff.

Conclusions about CD Evolution. Celiac disease 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.

But this mechanism leaves much unexplained. 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 became a staple food among the Saharawi only within the last century, 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.

Non-HLA risk loci for CD show a much more complex pattern of interactions between ancient environments and present-day risk. Some of these loci do appear to have responded to selection on a recent timescale, but most show no evidence of selection individually. While some parts of the CD risk joint-association network are connected by protein–protein interactions (Figure 1), the CD risk network as a whole 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. These were influenced by different environmental pressures, at different times in human evolutionary history. Celiac disease is a side effect of evolution, apparently in multiple ancient environments. Its persistence despite the consumption of wheat in ancient human populations may be attributed both to balancing selection on HLA class II alleles and the low additive variance contributed by many other loci associated with CD risk. 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; Cotsapas et al. 2011; Gutierrez-Achury et al. 2011; Ahn et al. 2012; Cifuentes et al. 2012). Celiac disease may provide a model for understanding the relations of the present risk of autoimmune conditions to past human environments.

Literature Cited

Abadie, V., L. M. Sollid, L. B. Barreiro et al. 2011. Integration of genetic and immunological insights into a model of celiac disease pathogenesis. *Annu. Rev. Immunol.* 29:493–525.

Ahn, R., Y. Ding, J. Murray et al. 2012. Association analysis of the extended MHC region in celiac disease implicates multiple independent susceptibility

loci. *PLoS ONE* 7:e36926.

Albrechtsen, A., I. Moltke, and R. Nielsen. 2010. Natural selection and the distribution of identity-by-descent in the human genome. *Genetics* 186:295–308.

Alvey, C., C. M. Anderson, and M. Freeman. 1957. Wheat gluten and coeliac disease. *Arch. Dis. Child.* 32:434–437.

Anderson, C. M., J. M. French, H. G. Sammons et al. 1952. Coeliac disease: Gastrointestinal studies and the effect of dietary wheat flour. *Lancet* 1:836–842.

Armelagos, G. J., and K. N. Harper. 2005. Genomics at the origins of agriculture, part two. *Evol. Anthropol.* 14:109–121.

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

Barker, J., and E. Liu. 2008. Celiac disease: Pathophysiology, clinical manifestations, and associated autoimmune conditions. *Adv. Pediatr.* 55:349–365.

Barreiro, L. B., and L. I. S. Quintana-Murci. 2010. From evolutionary genetics to human immunology: How selection shapes host defence genes. *Nat. Rev.*

Genet. 11:17–30.

Barrett, J. C., D. G. Clayton, P. Concannon et al. 2009. Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat. Genet.* 41:703–707.

Bergström, T., and U. Gyllensten. 1995. Evolution of MHC class II polymorphism: The rise and fall of class II gene function in primates. *Immunol. Rev.* 143:13–31.

Bevan, S., S. Popat, C. P. Braegger et al. 1999. Contribution of the MHC region to the familial risk of coeliac disease. *J. Med. Genet.* 36:687–690.

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

Botstein, D., J. M. Cherry, M. Ashburner et al. 2000. Gene ontology: Tool for the unification of biology. *Nat. Genet.* 25:25–29.

Bucci, P., F. Carile, A. Sangianantoni et al. 2007. Oral aphthous ulcers and dental enamel defects in children with coeliac disease. *Acta. Paediatr.* 95:203–207.

Buhler, S., and A. Sanchez-Mazas. 2011. HLA DNA sequence variation among human populations: Molecular signatures of demographic and selective events. *PLoS ONE* 6:e14643.

Campisi, G., C. Di Liberto, G. Iacono et al. 2007. Oral pathology in untreated coelic disease. *Aliment. Pharmacol. Thera.* 26:1,529–1,536.

Catassi, C., M. Doloretta Macis, I. M. Ratsch et al. 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., E. Fabiani, I. M. Ratsch et al. 1996. The coeliac iceberg in Italy. A multicentre antigliadin antibodies screening for coeliac disease in school-age subjects. *Acta. Paediatr.* 85:29–35.

Catassi, C., I. Ratsch, L. Gandolfi et al. 1999. Why is coeliac disease endemic in the people of the Sahara? *Lancet* 354:647–648.

Ch'ng, C. L., M. K. Jones, and J. G. C. Kingham. 2007. Celiac disease and autoimmune thyroid disease. *Clin. Med. Res.* 5:184–192.

Cheng, J., M. Kalliomäki, H. G. H. J. Heilig et al. 2013. Duodenal microbiota composition and mucosal homeostasis in pediatric celiac disease. *BMC Gastroenterol.* 13:113.

Cheng, J., T. Malahias, P. Brar et al. 2010. The association between celiac disease, dental enamel defects, and aphthous ulcers in a United States cohort. *J. Clin. Gastroenterol.* 44:191–194.

Cifuentes, R. A., D. Restrepo-Montoya, and J.-M. Anaya. 2012. The autoimmune tautology: An in silico approach. *Autoimmune Dis.* 2012:1–10.

Corrao, G., G. R. Corazza, V. Bagnardi et al. 2001. Mortality in patients with coeliac disease and their relatives: A cohort study. *Lancet* 358:356–361.

Corrao, G., P. Usai, G. Galatola et al. 1996. Estimating the incidence of coeliac disease with capture-recapture methods within four geographic areas in Italy. *J. Epidemiol. Community Health* 50:299–305.

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

Cotsapas, C., B. F. Voight, E. Rossin et al. 2011. Pervasive sharing of genetic effects in autoimmune disease. *PLoS Genet.* 7:e1002254

DeMarchi, M., I. Borelli, E. Olivetti et al. 1979. Two HLA-D and DR alleles are associated with coeliac disease. *Tissue Antigens* 14:309–316.

Dicke, W. K. 1950. *Coeliakie: Een onderzoek naar de nadelige invloed van sommige graansoorten op de lijder aan coeliakie.* Thesis, Rijksuniversiteit te Utrecht.

Dicke, W. K., H. A. Weijers, and J. H. V. D. Kamer. 1953. Coeliac disease: The presence in wheat of a factor having a deleterious effect in cases of coeliac

disease. *Acta Paediatr.* 42:34–42.

Dieterich, W., T. Ehnis, M. Bauer et al. 1997. Identification of tissue transglutaminase as the autoantigen of celiac disease. *Nat. Med.* 3:797–801.

Dowd, B., and J. Walker-Smith. 1974. Samuel Gee, Aretaeus, and the coeliac affection. *Br. Med. J.* 2:45.

Dubois, P. C. A., G. Trynka, L. Franke et al. 2010. Multiple common variants for celiac disease influencing immune gene expression. *Nat. Genet.* 42:295–302.

Falchuk, Z. M., G. N. Rogentine, and W. Strober. 1972. Predominance of histocompatibility antigen HL-A8 in patients with gluten-sensitive enteropathy. *J. Clin. Invest.* 51:1,602–1,605.

Farrell, R. J., and C. P. Kelly. 2002. Celiac sprue. *N. Engl. J. Med.* 346:180–188.

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

Festen, E. A. M., P. Goyette, R. Scott et al. 2009. Genetic variants in the region harbouring IL2/IL21 associated with ulcerative colitis. *Gut* 58:799–804.

Festen, E. A. M., P. Goyette, T. Green 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 Genet.* 7:e1001283.

Fleckenstein, B. 2004. Molecular characterization of covalent complexes between tissue transglutaminase and gliadin peptides. *J. Biol. Chem.* 279:17,607–17,616.

Fleckenstein, B., Ø. Molberg, S.-W. Qiao et al. 2002. Gliadin T cell epitope selection by tissue transglutaminase in celiac disease. *J. Biol. Chem.* 277:34,109–34,116.

Forsberg, G., A. Fahlgren, P. Hörstedt et al. 2004. Presence of bacteria and innate immunity of intestinal epithelium in childhood celiac disease. *Am. J. Gastroenterol.* 49:170–176.

Franceschini, A., D. Szklarczyk, S. Frankild et al. 2013. STRING v9.1: Protein–protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.* 41(Database issue):D808–815.

Frenette, P. S., and G. F. Atweh. 2007. Sickle cell disease: Old discoveries, new concepts, and future promise. *J. Clin. Invest.* 117:850–858.

Fu, Q., P. Rudan, S. Pääbo et al. 2012. Complete mitochondrial genomes reveal Neolithic expansion into Europe. *PLoS ONE* 7:e32473.

Fujimura, K. E., N. A. Slusher, M. A. Cabana, and S. A. Lynch. 2010. Role of the gut microbiota in defining human health. *Expert Rev. Anti Infect. Ther.* 8:435–454.

Fumagalli, M., M. Sironi, U. Pozzoli et al. 2011. Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. *PLoS Genet.* 7:e1002355.

Gamba, C., E. Fernández, M. Tirado 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., R. Pratesi, J. Cordoba et al. 2000. Prevalence of celiac disease among blood donors in Brazil. *Am. J. Gastroenterol.* 95:689–692.

Gonzalez-Galarza, F. F., S. Christmas, D. Middleton, and A. R. Jones. 2011. Allele frequency net: A database and online repository for immune gene frequencies in worldwide populations. *Nucleic Acid Res.* 39:D913–D919.

Gorlov, I. P., O. Y. Gorlova, S. R. Sunyaev et al. 2008. Shifting paradigm of association studies: Value of rare single-nucleotide polymorphisms. *Am. J. Hum. Genet.* 82:100–112.

Greco, L. 2002. The first large population based twin study of coeliac disease.

Gut 50:624–628.

Greco, D., M. Pisciotta, F. Gambina et al. 2012. Celiac disease in subjects with type 1 diabetes mellitus: a prevalence study in western Sicily (Italy). *Endocrine* 43:108–111.

Green, P., and C. Cellier. 2007. Celiac disease. *N. Engl. J. Med.* 357:1,731–1,743.

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

Haak, W., O. Balanovsky, J. J. Sanchez et al. 2010. Ancient DNA from European early neolithic farmers reveals their near eastern affinities. *PLoS Biol.* 8:e1000536.

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

Hancock, A. M., G. Alkorta-Aranburu, D. B. Witonsky et al. 2010a. Adaptations to new environments in humans: The role of subtle allele frequency shifts. *Philos. Trans. R. Soc. Lond. B. Biol. Sci*. 365:2459–2468.

Hancock, A. M., D. B. Witonsky, G. Alkorta-Aranburu et al. 2011. Adaptations to climate-mediated selective pressures in humans. *PLoS Genet.* 7:e1001375.

Hancock, A. M., D. B. Witonsky, E. Ehler et al. 2010b. Colloquium paper: Human adaptations to diet, subsistence, and ecoregion are due to subtle shifts in allele frequency. *Proc. Nat. Acad. Sci. USA* 107:8,924–8,930.

Hawks, J., E. T. Wang, G. M. Cochran et al. 2007. Recent acceleration of human adaptive evolution. *Proc. Nat. Acad. Sci. USA* 104:20,753–20,758.

Hunt, K. A., V. Mistry, N. A. Bockett et al. 2013. Negligible impact of rare autoimmune-locus coding-region variants on missing heritability. *Nature* 498:232–235.

Hunt, K. A., A. Zhernakova, G. Turner et al. 2008. Newly identified genetic risk variants for celiac disease related to the immune response. *Nat. Genet.* 40:395–402.

Hüe, S., J. J. Mention, R. C. Monteiro et al. 2004. A direct role for NKG2D/MICA interaction in villous atrophy during celiac disease. *Immunity* 21:367–377.

Kagnoff, M. F. 1984. Possible role for a human adenovirus in the pathogenesis of celiac disease. *J. Exp. Med.* 160:1,544–1,557.

Kagnoff, M. F. 2007. Celiac disease: Pathogenesis of a model immunogenetic disease. *J. Clin. Invest.* 117:41–49.

Karell, K., A. S. Louka, S. J. Moodie et al. 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. *Hum. Immunol.* 64:469–477.

Karinen, H., P. Kärkkäinen, J. Pihlajamäki et al. 2006. Gene dose effect of the DQB1*0201allele contributes to severity of coeliac disease. *Scand. J. Gastroenterol.* 41:191–199.

Keller, A., A. Graefen, M. Ball 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., H. Özkan, C. Pozzi et al. 2009. Domestication of the Triticeae in the fertile crescent. In *Genetics and Genomics of the Triticeae,* Vol. 7., C. Feuillet and G. J. Muehlbauer, eds. London; New York: Springer, 81–119.

Koning, F. 2012. Celiac disease: Quantity matters. *Semin. Immunopathol*. Online Preprint 34:541–549.

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

Larsen, C. S. 2006. The agricultural revolution as environmental catastrophe: Implications for health and lifestyle in the Holocene. *Quat. Int*. 150:12–20.

Lawlor, D., J. Zemmour, P. Ennis et al. 1990. Evolution of class-I MHC genes and proteins: From natural selection to thymic selection. *Annu. Rev. Immunol.* 8:23–63.

Losowsky, M. S. 2008. A history of coeliac disease. *Dig. Dis.* 26:112–120.

Ma, L., A. Brautbar, E. Boerwinkle et al. 2012. Knowledge-driven analysis identifies a gene–gene interaction affecting high-density lipoprotein cholesterol levels in multi-ethnic populations. *PLoS Genet.* 8:e1002714.

Margaritte-Jeannin, P., M. C. Babron, M. Bourgey 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, M. N. 1992. Mucosal pathology in gluten sensitivity. In *Coeliac Disease*, M. N. Marsh, ed. Oxford, UK: Blackwell Scientific Publishing, 136–191.

Mäki, M., K. Mustalahti, and J. Kokkonen. 2003. Prevalence of celiac disease among children in Finland. *N. Engl. J. Med.* 348:2,517–2,524.

Mention, J. J., M. Ben Ahmed, B. Bègue et al. 2003. Interleukin 15: A key to disrupted intraepithelial lymphocyte homeostasis and lymphomagenesis in

celiac disease. *Gastroenterol*. 125:730–745.

Meresse, B., Z. Chen, C. Ciszewski 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* 21:357I366.

Molberg, Ø., S. N. McAdam, R. Körner et al. 1998. Tissue transglutaminase selectively modifies gliadin peptides that are recognized by gut-derived T cells in celiac disease. *Nat. Med.* 4:713–717.

Murray, J. A., S. B. Moore, C. T. van Dyke et al. 2007. HLA DQ gene dosage and risk and severity of celiac disease. *Clin. Gastroenterol. and Hepatol.* 5:1,406–1,412.

Nilsen, E. M., F. L. Jahnsen, K. Lundin et al. 1998. Gluten induces an intestinal cytokine response strongly dominated by interferon gamma in patients with celiac disease. *Gastroenterol.* 115:551–563.

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

Ou, G., M. Hedberg, P. Hörstedt et al. 2009. Proximal small intestinal microbiota and identification of rod-shaped bacteria associated with childhood celiac disease. *Am. J. Gastroenterol.* 104:3,058–3,067.

Palmer, C., E. M. Bik, D. B. Digiulio et al. 2007. Development of the human infant intestinal microbiota. *PLoS Biol.* 5:e177.

Petronzelli, F., M. Bonamico, and P. Ferrante. 1997. Genetic contribution of the HLA region to the familial clustering of coeliac disease. *Ann. Hum. Genet.* 61:307–317.

Pham Short, A., K. C. Donaghue, G. Ambler et al. 2012. Coeliac disease in Type 1 diabetes from 1990 to 2009: higher incidence in young children after longer diabetes duration. *Diabet. Med.* 29:e286–e289.

Pickrell, J. K., G. Coop, J. Novembre et al. 2009. Signals of recent positive selection in a worldwide sample of human populations. *Genome Res.* 19:826–837.

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

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

Rajapakse, I., M. D. Perlman, P. J. Martin et al. 2012. Multivariate detection of gene–gene interactions. *Genet. Epidemiol.* 36:622–630.

Rao, S., M. Yuan, X. Zuo et al. 2011. A novel evolution-based method for detecting gene-gene interactions. *PLoS ONE* 6:e26435.

Ricaño-Ponce, I., and C. Wijmenga. 2013. Mapping of immune-mediated disease genes. *Annu. Rev. Genomics Hum. Genet.* 14:11.1–11.29.

Riddle, M. S., J. A. Murray, and C. K. Porter. 2012. The incidence and risk of celiac disease in a healthy US adult population. *Am. J. Gastroenterol.* 107:1,248–1,255. Advance online: DOI: 10.1146/annurev-genom-091212-153450

Rioux, J., P. Goyette, T. Vyse et al. 2009. Mapping of multiple susceptibility variants within the MHC region for 7 immune-mediated diseases. *Proc. Nat. Acad. Sci. USA* 106:18,680–18,685.

Risch, N. 1987. Assessing the role of HLA-linked and unlinked determinants of disease. *Am. J. Hum. Genet.* 40:1–14.

Rubio-Tapia, A., J. F. Ludvigsson, T. L. Brantner et al. 2012. The prevalence of celiac disease in the United States. *Am. J. Gastroenterol.* 107:1,538–1,544.

Sabeti, P. C., D. E. Reich, J. M. Higgins et al. 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419:832–837.

Sams, A., and J. Hawks. 2013. Patterns of population differentiation and natural selection on the celiac disease background risk network. *PLoS ONE* 8:e70564.

Sanz, Y., G. De Palma, M. Laparra. 2011. Unraveling the ties between celiac disease and intestinal microbiota. *Int. Rev. Immunol.* 30:207–218.

Sánchez-Quinto, F., H. Schroeder, O. Ramirez et al. 2012. Genomic affinities of two 7,000-year-old Iberian hunter-gatherers. *Curr. Biol.* 22:R631–R633.

Sellito, M., G. Bai, G. Serena et al. 2012. Proof of concept of microbiome-metabolome analysis and delayed gluten exposure on celiac disease autoimmunity in genetically at-risk infants. *PLoS ONE* 7:e33387.

Shan, L., Ø. Molberg, I. Parrot et al. 2002. Structural basis for gluten intolerance in celiac sprue. *Science* 297:2,275–2,279.

Simoons, F. J. 1981. Celiac disease as a geographic problem. In *Food, Nutrition, and Evolution: Food as an Environmental Factor in the Genesis of Human Variability*, N. Kretchmer and D. N. Walcher, eds. New York: Masson, 179–199.

Sjöberg, V., O. Sandström, M. Hedberg et al. 2013. Intestinal T-cell responses in celiac disease: Impact of celiac disease associated bacteria. *PLoS ONE* 8:e53414.

Skoglund, P., H. Malmström, M. Raghavan et al. 2012. Origins and genetic legacy of neolithic farmers and hunter-gatherers in Europe. *Science* 336:466–469.

Smith, W. P., Q. Vu, S. S. Li et al. 2006. Toward understanding MHC disease associations: partial resequencing of 46 distinct HLA haplotypes. *Genomics* 87:561.

Smyth, D. J., V. Plagnol, N. M. Walker et al. 2008. Shared and distinct genetic variants in type 1 diabetes and celiac disease. *N. Engl. J. Med.* 359:2,767–2,777.

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

Solberg, O. D., S. J. Mack, A. K. Lancaster et al. 2008. Balancing selection and heterogeneity across the classical human leukocyte antigen loci: A meta-analytic review of 497 population studies. *Hum. Immunol.* 69:443–464.

Sollid, L. M. 1989. Evidence for a primary association of celiac disease to a particular HLA-DQ alpha/beta heterodimer. *J. Exp. Med.* 169:345–350.

Sollid, L. M., and E. Thorsby. 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.

Soni, S., and S. Badawy. 2010. Celiac disease and its effect on human reproduction. *J. Reprod. Med.* 55:3–8.

Spurkland, A., L. M. Sollid, K. S. Rønningen et al. 1990. Susceptibility to develop celiac disease is primarily associated with HLA-DQ alleles. *Hum. Immunol.* 29:157–165.

Stene, L. C., M. C. Honeyman, E. J. Hoffenberg et al. 2006. Rotavirus infection frequency and risk of celiac disease autoimmunity in early childhood: A longitudinal study. *Am. J. Gastroenterol.* 101:2333–2340.

Stranger, B. E., E. A. Stahl, and T. Raj. 2011. Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics* 187:367–383.

Subramanian, A., P. Tamayo, V. K. Mootha et al. 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Nat. Acad. Sci. USA* 102:15,545–15,550.

Szperl, A. M., I. Ricaño-Ponce, J. K. Li et al. 2011. Exome sequencing in a family segregating for celiac disease. *Clin. Genet.* 80:138–147.

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

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

Townsend, A., and H. Bodmer. 1989. Antigen recognition by class I-restricted T lymphocytes. *Annu. Rev. Immunol.* 7:601–624.

Traherne, J. A. 2008. Human MHC architecture and evolution: Implications for disease association studies. *Int. J. Immunogenet.* 35:179–192.

Tresset, A., and J. D. Vigne. 2011. Last hunter-gatherers and first farmers of Europe. *C. R. Biol.* 334:182–189.

Trynka, G., K. A. Hunt, N. A. Bockett et al. 2011. Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nat. Genet.* 43:1,193–1,201.

Trynka, G., A. Zhernakova, J. Romanos et al. 2009. Coeliac disease-associated risk variants in TNFAIP3 and REL implicate altered NF- B signalling. *Gut* 58:1,078–1,083.

Van De Wal, Y., Y. Kooy, P. Van Veelen et al. 1998. Cutting edge: Selective

deamidation by tissue transglutaminase strongly enhances gliadin-specific T cell reactivity. *J. Immunol.* 161:1585–1588.

Van Heel, D. A., L. Franke, K. A. Hunt et al. 2007. A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21. *Nat. Genet*. 39:827–829.

Willcox, G. 2012. The beginnings of cereal cultivation and domestication in Southwest Asia. In *A Companion to the Archaeology of the Ancient Near East: Volume I,* D. T. Potts, ed. Chichester, West Sussex; Malden, MA: Blackwell Publishing Ltd., 163–180.

Zeder, M. A. 2011. The origins of agriculture in the Near East. *Curr. Anthropol.* 52:S221–S235.

Zelnik, N., A. Pacht, R. Obeid et al. 2004. Range of neurologic disorders in patients with celiac disease. *Pediatrics* 113:1672–1676.

Zhernakova, A., B. Z. Alizadeh, M. Bevova 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. *Am. J. Hum. Genet.* 81:1,284–1,288.

Zhernakova, A., C. C. Elbers, B. Ferwerda et al. 2010. Evolutionary and

functional analysis of celiac risk loci reveals SH2B3 as a protective factor against bacterial infection. *Am. J. Hum. Genet*. 86:970–977.

Zhernakova, A., E. A. Stahl, G. Trynka et al. 2011. Meta-analysis of genome-wide association studies in celiac disease and rheumatoid arthritis identifies fourteen non-HLA shared loci. *PLoS Genet.* 7:e1002004.

Zuk, O., E. Hechter, S. R. Sunyaev et al. 2012. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc. Nat. Acad. Sci. USA* 109:1,193–1,198.

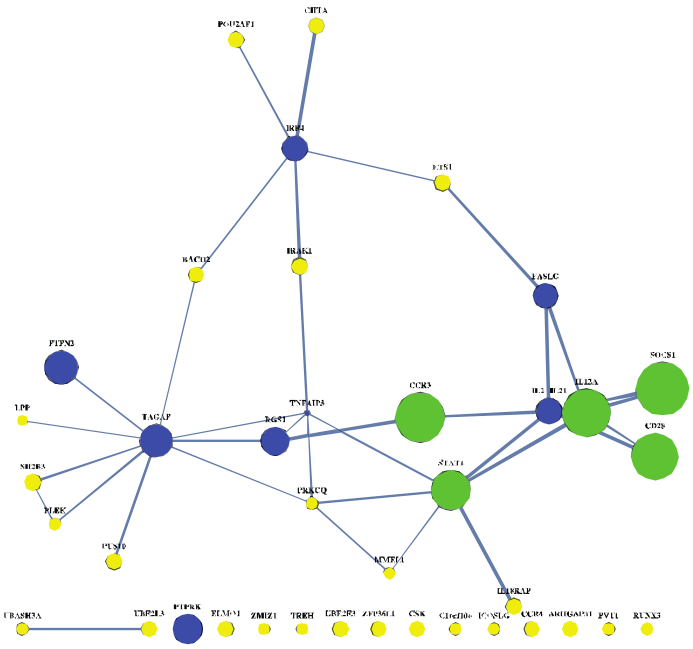

Figure 1. Protein–protein interactions and effect-sizes among non-HLA celiac disease risk loci. Protein–protein interactions among non-HLA CD risk loci via the STRING (v.9.1) database (Franceschini et al. 2013), which is a database of known and predicted protein interactions derived from multiple contexts. Node colors reflect the number of independent associations to CD at the locus centered on each gene (Yellow = 1; Blue = 2; Green = 3). Relative node sizes reflect the total effect size of each locus, with independent signals at each locus summed. Edge thicknesses reflect the confidence in each interaction according to the STRING database.

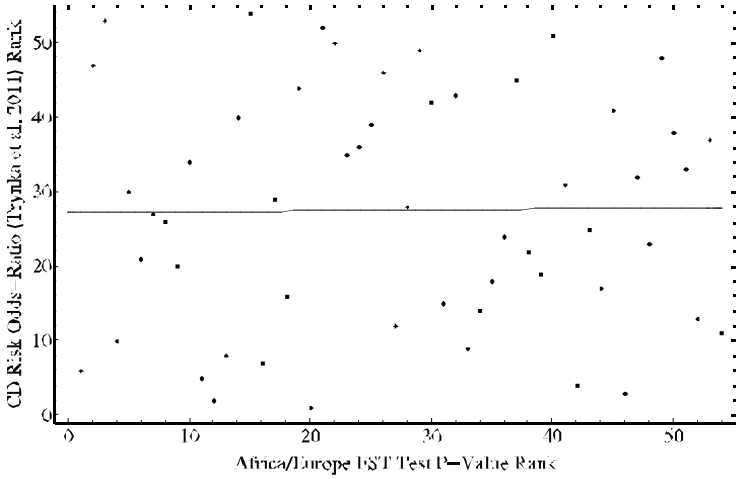

Figure 2. Africa/Europe CD risk odds-ratio / F_{ST} test p-value rank order correlation. Rank order correlation between Africa/Europe genome-wide F_{ST} test p-values and CD risk locus odds-ratios reported by Trynka and colleagues (2011) demonstrating the lack of correlation between these two variables.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