GALLAUDET
UNIVERSITY

THE HONORS PROGRAM

The Mystery of FUT2

A population genetics analysis of the secretor gene using the 1000 Genomes Project

An Honors Capstone Submitted in Partial Fulfillment of the Requirements for Graduation with University Honors

By: Brienna Herold

Committee Chair: Derek C. Braun, PhD

Second Reader: Kathleen S. Arnos, PhD

Honors Director: Shirley Shultz Myers, PhD

Honors Coordinator: Geoffrey Whitebread, MA

May 5, 2014

ABSTRACT

Because the ability to resist infectious disease has been a constant factor in the survival of humans over the past few millennia, disease resistance has played a crucial role in shaping the genes of the immune system. One of these genes, FUT2, is suggested to have evolved as an adaptation reinforcing the immune system, although the specific function of its Lewis b and ABO blood group antigens in body fluids and secretions are still unclear. However, nonfunctional alleles are present in populations worldwide at frequencies higher than expected under neutral evolution. To elucidate the role of FUT2 in the immune system, my analysis examined the nucleotide variation at FUT2 using an array of neutrality tests and sequence data from the 1000 Genomes Project. My analysis, like those preceding mine, was inconclusive with respect to the possible role of FUT2 in the immune system. It is possible that FUT2 is undergoing directional selection in Asians or that the gene is undergoing no natural selection at all. Either way, both conclusions have direct implications for public health, but this needs to be investigated further to decide what the implications are.

TABLE OF CONTENTS

INTRODUCTION

In the absence of modern infectious disease control (e.g., hand soap, antibiotics, and vaccines), a common obstacle to passing down genes from parents to children would be a fatal infection that strikes before reproduction (Weatherall et al., 2006). Thus disease resistance has been a constant factor in the evolution of humans over the past few millennia. To survive long enough to pass down their genes, human populations have had to evolve adaptations to resist the pressure of a variety of infectious diseases or else perish. This fact is evident in the genes of the immune system, the body's defense against foreign disease-causing microorganisms. Many of the genes involved in immunity show strong signatures of having been shaped by natural selection, a direct consequence of the genes' success in resisting disease at least long enough for reproduction to occur (Sabeti, 2008; Ford, 2002).

One such gene, FUT2, located on chromosome 19, is suggested to have evolved as an adaptation reinforcing the immune system (D'Adamo and Kelly, 2001; Linden et al., 2008). FUT2 encodes the secretor-type galactoside 2-α-L-fucosyltransferase, an enzyme found in the epithelia of secretory tissues such as the salivary glands, sweat glands, and gastrointestinal tract. This enzyme regulates the expression of the H antigen, a precursor of the Lewis b antigen and the ABO blood group antigens, in body fluids and secretions (Dean, 2005). The biological mechanisms of these antigens in resisting disease are not yet fully understood (Campi et al., 2012), although several correlations to increased disease susceptibility have been linked to their absence. Some examples include Crohn's disease (McGovern et al., 2010), gum disease (Tabasum and Nayak, 2011), oral cancer (Moreno et al., 2009), *Helicobacter pylori* infections (Linden et al., 2008), and reduced levels of healthy intestinal microbial diversity (Wacklin et al., 2011).

If these correlations are the result of FUT2 being part of the immune system, it is then remarkable that the frequencies of several nonfunctional FUT2 alleles (alternative forms of the gene) are much higher than expected for a gene involved in immunity. The most common nonfunctional allele found in Caucasians, se^{428}, comprises 47 percent of the total FUT2 allelic variation in this ethnic group (Liu et al., 1998). Another common nonfunctional allele, $se^{357,385}$, is found in the Japanese at a frequency of 43.6 percent (Narimatsu et al., 1998). Each nonfunctional allele is caused by at least one loss-of-function mutation, and individuals who inherit two copies of a nonfunctional allele cannot express Lewis b or ABO blood group antigens in body fluids or secretions (Dean, 2005). Because these individuals, appropriately called non-secretors, are believed to have diminished immune system efficiency in comparison to secretors, it is unusual that nonfunctional alleles such as se^{428} and $se^{357,385}$ have risen to such high frequencies. The immune system is crucial to an individual's survival, so it is expected that natural selection would have removed nonfunctional FUT2 alleles from the total allelic variation or at least have reduced their frequencies (Khan, 2009).

An explanation for the high frequencies of several nonfunctional FUT2 alleles may be that there is also a relationship between increased infectious disease resistance and the absence of antigens in body fluids and secretions. For example, several recent studies have reported that non-secretors are more resistant than secretors to several respiratory diseases including the norovirus, a virus causing sudden gastroenteritis that is normally only inconvenient today but would have been much more life-threatening before modern infectious disease control (Thorven et al., 2005; Rydell et al., 2011). Also, other recent studies have reported that non-secretors carrying se^{428} alleles show a slower progression of HIV-1 infection compared to secretors (Ali et al., 2000; Kindberg et al., 2006). But again, the biological mechanisms underlying the

correlations between increased disease resistance and the absence of antigens are still unclear (Campi et al., 2012).

To help elucidate the role FUT2 plays in the immune system, researchers are carrying out population genetics analyses in which they use statistical methods called neutrality tests to examine the nucleotides constituting the gene for patterns of variation that are known to be characteristic of natural selection. Investigating the relationships between natural selection and infectious disease resistance can reveal the forces that shaped the present state of the human immune system and improve understanding of gene function (Nielsen et al., 2007). The human β-globin gene is the best example of a gene whose role in the immune system was elucidated through population genetics analyses. For a long time, it was not clear why a mutated β-globin allele, HbS, was so widespread in the African population when individuals who carried two copies of this mutated allele suffered from sickle cell anemia, a major risk factor in premature death. Population genetics analyses revealed that a form of natural selection known as balancing selection was maintaining both the mutated allele and the normal allele at intermediate frequencies in the African population. This discovery in addition to epidemiological observations allowed Haldane (1949) to conclude that the deadly malaria parasite was the main force driving the evolution of the β-globin gene; the mutated allele conferred some resistance to the deadly malaria parasite, and individuals who carried both alleles benefited from this, while individuals who carried two copies of either allele suffered from sickle cell anemia or caught malaria and died.

While population genetics analyses have helped to elucidate the role of the human β-globin gene in the immune system, the current collection of results from reported population genetics analyses of FUT2 are contradictory and reveal nothing conclusive about the gene. Koda,

Soejima, and Kimura (2001), Koda et al. (2001), Soejima et al. (2007), and Ferrer-Admetlla et al. (2009) reported signatures of balancing selection at the gene in several non-Asian populations including Caucasians, European-Africans, Iranians, and Ghanaians. In Asian populations including the Japanese and the Chinese, the signatures that have been reported include those of directional selection (Ferrer-Admetlla et al., 2009) and nothing relevant (Koda et al., 2001). Some studies reported that there are no signatures of natural selection at FUT2 in Africans (Koda et al., 2001) or in the entire global population (Clark et al., 2003). Walsh et al. (2005) reported signatures of natural selection at FUT2 in the YSN (Yoruba Nigerians from Southwest Nigeria) and HCG (Han Chinese trios from Guangxi) populations but did not specify which form of natural selection they detected.

This study attempts to further elucidate the role FUT2 plays in the immune system by building upon previous studies in two ways. First, this study includes in its carefully selected array of neutrality tests a test that has never been used before. This neutrality test is the Hudson-Kreitman-Aguade Test, one of the most powerful neutrality tests currently in use (Zhai et al., 2009). Second, this study analyzes DNA sequence data from the 1000 Genomes Project, a database that has never been used before. Previous studies either sequenced their own data or used data from the International Haplotype Map Project, an older database with similar goals as the 1000 Genomes Project but containing less sequence detail (Buchanan et al., 2012). The advantage of using the 1000 Genomes Project as a database is that it is currently the most complete inventory of human sequence data ever assembled (Buchanan et al., 2012) and is thus the best representation of the human population. With sequence data from the 1000 Genomes Project and an array of neutrality tests including Tajima's D, Fay and Wu's H, the McDonald-Kreitman Test, and the powerful Hudson-Kreitman-Aguade Test, the aim of this population

genetics analysis of FUT2 is to infer whether the gene has undergone natural selection and to interpret the inference with respect to the possible role of FUT2 in the immune system.

Neutrality tests

This study implements an array of neutrality tests which needs an explanation. When carried out on DNA sequence data, neutrality tests quantify the observed patterns of nucleotide variation and infer whether they match the patterns expected under the null hypothesis (the default predicted outcome). Most neutrality tests use the neutral theory of molecular evolution as the null hypothesis. The neutral theory of molecular evolution holds that in a constant-sized population most mutations are neutral with respect to fitness—the organism's ability to survive and reproduce in its environment—and that these mutations' individual fates are governed by random evolution, or genetic drift, rather than non-random evolution, or natural selection (Nachman, 2006). In other words, the null hypothesis predicts that the observed patterns of nucleotide variation should match the patterns that are expected under random evolution. If the observed and expected patterns do not match, the null hypothesis is rejected, leaving the alternative hypotheses of demography and natural selection as explanations for the observed patterns.

Differentiating between demography and natural selection is a major challenge in population genetics, because both processes leave similar effects upon sequence data (Nielsen et al., 2007). Nielsen et al. (2007) suggests two strategies to deal with this challenge. The first strategy involves examining multiple independent loci, or regions, in the genome. This strategy depends on the fact that demography affects the entire genome, whereas natural selection affects only a few loci. If a population has contracted or expanded in size, then all loci should show patterns reflecting this demographic change. If a population has undergone natural selection, then

only a few loci should show patterns appearing as outliers in the overall pattern. The second strategy involves the use of a second species, called an outgroup, to determine whether a mutation is fixed (found in both species) or polymorphic (found within only one of the species). Since fixed mutations date to a time before the two species diverged, their inclusion in the analysis helps to differentiate the signatures of natural selection from the signatures of demography. Most neutrality tests incorporate at least one of the two strategies.

Neutrality tests can be categorized based on the type of data they use. There currently exist three categories of neutrality tests: 1) those that use variation data from within a species, which is also known as intraspecific variation data; 2) those that use variation data from between species, which is also known as interspecific variation data; and 3) those that use both (Nachman, 2006). Neutrality tests within each category differ in the specific variation measures they use and their predictions of how these measures are expected to manifest in DNA sequence data under the null hypothesis. Below I describe the tests that I included in my analysis with the type of data and specific variation measures they use, the predictions they make under the null hypothesis, and how the results may be interpreted with respect to the null hypothesis.

Tajima's D

The most widely used neutrality test based on intraspecific variation is Tajima's D (Zhai, 2009). Tajima's D compares two measures of intraspecific variation: the amount of mutations per nucleotide and the average amount of mutations between all possible sequence pairs. Both measures are expected to be equal under the null hypothesis, resulting in a D value that is statistically indistinguishable from zero. If the D value is significantly positive or negative, the two measures are not equal, and the null hypothesis is rejected. Positive values indicate an excess

of intermediate-frequency mutations relative to neutral expectation, which can result from balancing selection or a recent population contraction. Negative values indicate a deficit of intermediate-frequency mutations, which can result from directional selection, purifying selection, or a recent population expansion (Misawa and Tajima, 1997).

Fay and Wu's H

An improvement over Tajima's D (Zhai, 2009), Fay and Wu's H is based on both intraspecific and interspecific variation. That is, Fay and Wu's H compares the same two measures of variation as Tajima's D, except that these measures have been adjusted to rely on sequence data from not one species but two—human and chimpanzee in this study. The inclusion of an outgroup is one of the two strategies that allows for better differentiation between demography and natural selection. The null hypothesis predicts that polymorphic mutations should be present in frequencies much lower than fixed mutations. An H value close to zero indicates that this prediction is true. Significantly negative H values indicate an excess of high-frequency polymorphic mutations, which can result from directional selection or a recent population expansion (Fay and Wu, 2000).

The McDonald-Kreitman Test

Like Fay and Wu's H, the McDonald-Kreitman Test is also based on both intraspecific and interspecific variation, although it uses different measures of the variation. Fay and Wu's H compares the amount of mutations per nucleotide and the average amount of mutations between all possible sequence pairs, whereas the McDonald-Kreitman Test compares the amounts of replacement mutations (those that change the amino acids constituting the protein) and silent

mutations (those that do not). Under the null hypothesis, the ratio of polymorphic replacement to silent mutations and the ratio of fixed replacement to silent mutations should be equal. The McDonald-Kreitman Test gives an easy-to-interpret neutrality index that compares the two ratios. A neutrality index significantly greater than one indicates an excess of polymorphic replacement mutations relative to neutral expectation, which can result from purifying selection. A neutrality index significantly less than one indicates an excess of fixed replacement mutations relative to neutral expectation, which can result from directional selection (McDonald and Kreitman, 1991).

The Hudson-Kreitman-Aguade Test

Somewhat like the McDonald-Kreitman-Test, the Hudson-Kreitman-Aguade, which is also based on both intraspecific and interspecific variation, compares the ratios of polymorphic mutations to fixed mutations for multiple loci in two species. The null hypothesis predicts that the two ratios should be equal, and a built-in chi-square goodness-of-fit test evaluates this equality. The null hypothesis is usually rejected when the significance level for these ratios is $\chi^2 > 3.0$ and $p < 0.1$.

The Hudson-Kreitman-Aguade Test is one of the most powerful neutrality tests currently in use. Its inclusion of multiple loci and an outgroup gives it a nearly unparalleled power in differentiating between natural selection and demography (Zhai, 2009). The loci must comprise a test locus and at least one control locus that must be neutral (Hudson et al., 1987). I initially chose SEC1, located on chromosome 19, as the control locus, because it is a nonfunctional pseudogene of FUT2 and therefore free of natural selection (Soejima and Koda, 2013). However, when I realized that SEC1, being on the same chromosome as FUT2, might share a correlated

evolutionary history due to possible genetic linkage (Nielsen et al., 2007), I switched the choice of the control locus from SEC1 to GULOP, another pseudogene that is located on a different chromosome.

METHODOLOGY

To carry out my population genetics analysis of FUT2, I followed a procedure in which I prepared DNA sequence data, implemented neutrality tests using the prepared data, and evaluated the results for statistical significance. The sections below describe in detail each step of this procedure.

Sequence data preparation

I used sequence data from FUT2, SEC1, and GULOP, the latter two being the control loci in the Hudson-Kreitman-Aguade Test and the former being the test locus. For these three genes, I obtained human sequence data from the 1000 Genomes Browser v3.0.2 (ncbi.nlm.nih.gov/variation/tools/1000genomes) and chimpanzee sequence data from Ensembl (useast.ensembl.org). In both species, FUT2 and SEC1 sit on chromosome 19, and GULOP sits on chromosome 8. The nucleotide positions of each human FUT2, SEC1, and GULOP sequence are respectively 49,198,336 - 49,210,142; 49,136,919 - 49,189,899; and 27,433,907 - 27,447,872. The nucleotide positions of each chimpanzee FUT2, SEC1, and GULOP sequence are respectively 53,677,338 - 53,690,027; 53,612,189 - 53,668,907; and 23,872,269 - 238,862,259.

Since evolution occurs on the population level (Hamilton, 2009), I grouped human FUT2, SEC1, and GULOP sequences by population. (There was no need to sort the chimpanzee

sequences, because I had downloaded only one sequence per gene.) After this grouping step, I had 14 populations to work with: Americans of African Ancestry in Southwestern USA (ASW); Utah Residents with Northern and Western European ancestry (CEU); Han Chinese in Beijing, China (CHB); Southern Han Chinese (CHS); Colombians from Medellin Colombia (CLM); Finnish in Finland (FIN); British in England and Scotland (GBR); Iberian population in Spain (IBS); Japanese in Tokyo, Japan (JPT); Luhya in Webuye, Kenya (LWK); Mexican Ancestry from Los Angeles, USA (MXL); Puerto Ricans from Puerto Rico (PUR); Toscani in Italia (TSI); and Yoruba in Ibadan, Nigeria (YRI). The sequences totaled 1,704 and belonged to 852 unrelated individuals. The exact number of sequences belonging to each population are shown in Table 4.

To prepare the sequences for analysis by tests requiring sequence data from both human and chimpanzee (Fay and Wu's H, the MK Test, and the HKA Test), I generated a human-chimpanzee multiple sequence alignment for each population using the MAFFT v7 program (Katoh and Standley, 2013).

Neutrality test analyses

I implemented the following neutrality tests on each population: Tajima's D, Fay and Wu's H, the McDonald-Kreitman Test, and the Hudson-Kreitman-Aguade Test. To carry out each test, I used DnaSP v5.10.1 (Ramos-Onsins and Rozas, 2009), a population genetics software package which offers several neutrality tests as part of its array of modules designed to analyze sequence data. The neutrality test modules I used in this study include "Tajima's D," "McDonald-Kreitman Test," and "HKA Test (Direct Mode)." I obtained Fay and Wu's H values from the "Polymorphisms/Divergence Data" module.

The process of running each neutrality test differed. Tajima's D, Fay and Wu's H, and the MK Test analyzed the sequence data directly, but only Tajima's D required no additional input values. Fay and Wu's H and the MK Test required sequence sets to be defined, so that the program could identify which sequences were human or chimpanzee. The MK Test also required a protein-coding region to be defined in FUT2, since the test measures the amount of mutations that change the amino acids constituting the protein (replacement mutations). I used DQ321371.1, an annotated FUT2 sequence deposited in GenBank by Guo et al. (2004), to help me pinpoint the exact nucleotide positions of the protein coding region in each population. The HKA Test, instead of analyzing the sequence data directly, required the following values to be entered into a dialog box: intraspecific variation measures including the number of polymorphisms and the sample size, and interspecific variation measures including the average number of differences and the number of nucleotides. I obtained most of these input values from the "Polymorphism Data" module. I obtained the average number of differences from the "Polymorphism/Divergence Data" module and rounded the value to the nearest integer.

Evaluation of results

When implementing neutrality tests, the critical question is whether the observed results match the results expected under the null hypothesis. I assessed the significance of the observed Tajima's D and Fay and Wu's H values via "The Coalescent" module, which used the coalescent theory to simulate 5,000 neutrally evolving genealogies based on measures obtained from the sequence data. In this analysis, those measures were the sample size and the number of polymorphic mutations. The genealogy simulations shape a probability distribution for the expected D or H value, allowing for the calculation of an easy-to-interpret 95% confidence

interval, which is a range of values that could be true for the expected D or H value. If an observed value falls outside of the confidence interval, it is considered unexpected, and the null hypothesis is rejected. "The Coalescent" module also gives a p-value, which is a simple cut-off beyond which any mismatch between the observed and expected results is considered statistically significant (Davies and Crombie, 2009). Because I set the confidence interval at 95%, a p-value equal to or less than 0.05 was considered significant. The confidence interval and p-values are shown along with the results from Tajima's D and Fay and Wu's H in Tables 2 and 3.

To evaluate the results from the McDonald-Kreitman Test and the Hudson-Kreitman-Aguade Test, I used functions built into the modules for each test. For the McDonald-Kretiman Test, it was the Fisher's exact p-value, which is also significant when it is equal to or less than 0.05. For the Hudson-Kreitman-Aguade Test, it was the chi-square goodness-of-fit value and its accompanying p-value.

RESULTS AND DISCUSSION

To infer whether natural selection has shaped the nucleotide variation at FUT2, I carried out a population genetics analysis on 1,704 human sequences from FUT2, SEC1, and GULOP, in addition to a chimpanzee sequence from each of the three loci. The results for Tajima's D, Fay and Wu's H, the McDonald-Kreitman Test, and the Hudson-Kreitman-Aguade Test are shown in Tables 1, 2, 3, and 4, respectively. Table 5 summarizes the main results from each of the four neutrality tests.

Table 1.

Tajima's D results.

Superpopulation	Population	D	Confidence interval	*p*
ASN	CHB	-1.484	-1.581 — 1.886	0.038
	JPT	-0.365	-1.594 — 2.024	0.420
	CHS	-0.626	-1.623 — 1.919	0.308
EUR	CEU	2.437	-1.659 — 1.888	0.992
	TSI	2.331	-1.566 — 1.863	0.991
	FIN	2.078	-1.531 — 1.796	0.985
	GBR	2.201	-1.562 — 1.866	0.988
	IBS	0.404	-1.767 — 1.712	0.722
AFR	YRI	1.515	-1.574 — 1.879	0.948
	LWK	1.256	-1.556 — 1.940	0.922
	ASW	0.982	-1.597 — 1.853	0.872
AMR	MXL	0.751	-1.608 — 1.840	0.833
	PUR	1.466	-1.614 — 1.785	0.951
	CLM	0.894	-1.605 — 1.842	0.850

Table 2.

Fay and Wu's H results.

Superpopulation	Population	H	Confidence Interval	*p*
ASN	CHB	-29.794	-4.990 — 4.078	0.000
	JPT	-5.430	-6.380 — 2.430	0.035
	CHS	-4.464	-6.119 — 2.735	0.049
EUR	CEU	-0.012	-13.378 — 5.894	0.320
	TSI	-0.046	-14.300 — 6.000	0.326
	FIN	-6.434	-13.868 — 6.007	0.103
	GBR	-1.640	-14.064 — 6.034	0.232

Superpopulation	Population	H	Confidence Interval	*p*
	IBS	-6.769	-14.330 — 7.077	0.109
AFR	YRI	0.247	-22.503 — 9.545	0.333
	LWK	-0.599	-23.565 — 9.957	0.305
	ASW	-4.325	-23.032 — 9.855	0.213
AMR	MXL	-10.698	-16.775 — 7.031	0.062
	PUR	-2.008	-16.904 — 7.480	0.236
	CLM	-9.466	-17.766 — 7.733	0.095

Table 3.

McDonald-Kreitman Test results.

Superpopulation	Population	Neutrality index	Fisher's *p*
ASN	CHB	5.550	0.018
	JPT	5.203	0.041
	CHS	6.061	0.018
EUR	CEU	10.176	0.015
	TSI	5.815	0.014
	FIN	3.180	0.161
	GBR	3.118	0.167
	IBS	2.963	0.239
AFR	YRI	14.417	0.002
	LWK	2.870	0.265
	ASW	5.267	0.027
AMR	MXL	3.975	0.078
	PUR	7.943	0.003
	CLM	4.675	0.005

Table 4.

Hudson-Kreitman-Aguade Test input and output values for each of the 14 populations and each of the three loci.

		Input values				Output values	
		Intraspecific variation		Interspecific variation		HKA Test	
Gene	Population	Sample size	Polymorphisms	Fixed differences	Sites	χ^2	*p*
FUT2	ASW	88	95	21	11798	-	-
	CEU	68	54	17	11805	-	-
	CHB	194	57	3	11802	-	-
	CHS	100	29	3	11805	-	-
	CLM	66	72	16	11799	-	-
	FIN	186	62	15	11804	-	-
	GBR	178	64	17	11804	-	-
	IBS	14	47	15	11805	-	-
	JPT	178	29	2	11805	-	-
	LWK	194	108	22	11799	-	-
	MXL	70	64	14	11804	-	-
	PUR	64	70	18	11804	-	-
	TSI	196	64	17	11804	-	-
	YRI	108	95	22	11799	-	-
SEC1	ASW	88	221	50	52599	0.010	0.922
	CEU	68	130	28	52764	0.097	0.755
	CHB	194	132	32	52670	0.805	0.370
	CHS	100	122	32	52614	0.555	0.456
	CLM	66	172	36	52809	0.000	0.984
	FIN	186	149	28	52693	0.051	0.822
	GBR	178	144	27	52634	0.084	0.772
	IBS	14	117	34	52799	0.001	0.972

		Input values				Output values	
		Intraspecific variation		Interspecific variation		HKA Test	
Gene	Population	Sample size	Polymorphisms	Fixed differences	Sites	χ^2	*p*
	JPT	178	128	29	52615	0.565	0.452
	LWK	194	271	43	52633	0.039	0.844
	MXL	70	153	32	52789	0.001	0.976
	PUR	64	171	33	52769	0.248	0.619
	TSI	196	182	27	52648	0.248	0.619
	YRI	108	236	45	52692	0.037	0.847
GULOP	ASW	88	100	13	13959	0.114	0.736
	CEU	68	50	9	13960	0.178	0.673
	CHB	194	60	11	13960	0.332	0.564
	CHS	100	54	10	13960	0.100	0.752
	CLM	66	58	11	13960	0.012	0.913
	FIN	186	56	8	13960	0.159	0.690
	GBR	178	60	10	13957	0.150	0.699
	IBS	14	26	10	13960	0.018	0.894
	JPT	178	59	11	13960	0.264	0.608
	LWK	194	114	14	13959	0.121	0.728
	MXL	70	60	9	13960	0.056	0.812
	PUR	64	60	10	13959	0.091	0.763
	TSI	196	68	10	13957	0.227	0.634
	YRI	108	97	13	13956	0.139	0.710

Table 5.

A summary table containing the main results from Tajima's D, Fay and Wu's H, the McDonald-Kreitman Test, and the Hudson-Kreitman-Aguade Test.

Superpopulation	Population	D	H	MK neutrality index	HKA χ^2	
					SEC1	GULOP
ASN	CHB	-1.484	-29.794	5.550	0.805	0.332
	JPT	-0.365	-5.430	5.203	0.565	0.264
	CHS	-0.626	-4.464	6.061	0.555	0.100
EUR	CEU	2.437	-0.012	10.176	0.097	0.178
	TSI	2.331	-0.046	5.815	0.248	0.227
	FIN	2.078	-6.434	3.180	0.051	0.159
	GBR	2.201	-1.640	3.118	0.084	0.150
	IBS	0.404	-6.769	2.963	0.001	0.018
AFR	YRI	1.515	0.247	14.417	0.037	0.139
	LWK	1.256	-0.599	2.870	0.039	0.121
	ASW	0.982	-4.325	5.267	0.010	0.114
AMR	MXL	0.751	-10.698	3.975	0.001	0.056
	PUR	1.466	-2.008	7.943	0.248	0.091
	CLM	0.894	-9.466	4.675	0.000	0.012

As Tables 1 and 5 show, the results from Tajima's D indicate that the null hypothesis is rejected at FUT2 in only one population: CHB (Han Chinese in Beijing, China). The D value in this population, -1.484, is significantly negative as determined by 5,000 coalescent simulations with a 95% confidence interval, which gave a p-value of less than 0.05. A significantly negative D value results from the presence of fewer intermediate-frequency mutations than expected under random evolution, and it can be explained by directional selection or a recent population expansion.

JPT (Japanese in Tokyo, Japan) and CHS (Southern Han Chinese), the other two populations in the ASN (Asian) superpopulation, may be undergoing a similar evolutionary or demographic change at FUT2. All of the three populations comprising the ASN superpopulation show a trend toward negative D values that is not shared by any other superpopulation in this study. In fact, CHB, JPT, and CHS are the only populations in this study that show a negative D value. The p-values for JPT and CHS, while not significant, are much closer to significance than in any of the other 11 populations.

The results from Fay and Wu's H, shown in Tables 2 and 5, also indicate that the null hypothesis is rejected at FUT2 in the ASN superpopulation. The H values for JPT and CHS, -5.40 and -4.464 respectively, are significantly negative ($p < 0.05$). For CHB, the 95% confidence interval of Fay and Wu's H lies between -4.990 and 4.078. Consequently, the observed H value, -29.794, for CHB is very unlikely under the null hypothesis; indeed, the probability of 5,000 coalescent simulations obtaining expected values equal to this observed H value is 0.000. Significantly negative H values indicate an excess of high-frequency polymorphic mutations relative to neutral expectations, which can be explained by directional selection.

This inference that the ASN population is undergoing directional selection is not supported by the results from the McDonald-Kreitman Test, but there is an issue with the presence of outliers. As shown in Tables 3 and 5, the neutrality indexes for CHB, JPT, and CHS are significantly greater than one (5.550, 5.203, and 6.601 respectively, with $p < 0.05$). A neutrality index significantly greater than one indicates an excess of polymorphic replacement mutations relative to neutral expectation, which can result from purifying selection, not directional selection. However, because the neutrality indexes for the YRI (Yoruba in Ibadan, Nigeria), ASW (Americans of African Ancestry in Southwestern USA), PUR (Puerto Ricans

from Puerto Rico), and CLM (Colombians from Medellin, Colombia) populations are also significant, it is important to note that with neutrality tests which focus on summary statistics such as D, H, and the neutrality index, it is not unusual to encounter incorrect rejections of the null hypothesis, since a summary statistic must throw away some of the information in order to condense nucleotide variation patterns into a single value (Kreitman, 2000). To reduce the chances of an incorrect rejection of the null hypothesis, the next implementation of the McDonald-Kreitman Test should include the modified version of the neutrality index suggested by Stoletzki and Eyre-Walker (2011).

None of the results from the Hudson-Kreitman-Aguade Test, shown in Tables 4 and 5, indicate that the null hypothesis is rejected in any of the 14 populations. The results are all not significant ($X2 < 3.0$ and $p > 0.1$), which indicate the ratios of polymorphic to fixed mutations for FUT2 and GULOP are statistically equal. In other words, the Hudson-Kreitman-Aguade Test shows no signature of natural selection at FUT2.

There are three possible explanations for the absence of significant results. The first explanation is that there was an error in implementing the test. There are three solutions to this problem. First, I can run the test again using the rhesus macaque as the outgroup instead of the chimpanzee. The reason for this change is because neither very similar nor very different sequences contain much information. Greater divergence time results in more fixed mutations and improves power to detect selection (Zhai, 2009). Second, I can search the literature for a locus that has been repeatedly proven to be neutral, since it is possible GULOP, despite it being a pseudogene without function, is undergoing natural selection too. Third, I can use more than one control locus and average the results. The more loci, the more statistical power the test has (Zhai, 2009).

The second explanation for the absence of significant results is that natural selection did not shape the nucleotide variation at FUT2, and that the signatures detected by Tajima's D and Fay and Wu's H are that of demography. This is a plausible explanation, because the Hudson-Kreitman-Aguade Test incorporates both of the two strategies that allow for better differentiation between natural selection and demography, meaning that significant results indicate only the action of natural selection. Moreover, East Asians are known to have undergone a recent rapid population expansion.

If the second explanation is true, this raises the question of whether FUT2 even has an important role in the immune system. It is possible that the presence or absence of Lewis b and ABO blood group antigens in body fluids or secretions was never an advantage or a disadvantage to fitness. With no impact upon fitness, no selective force would have had a reason to shape the nucleotide variation at the gene.

The third explanation is that the null hypothesis itself needs revising. The null hypothesis is based on the neutral theory of molecular evolution, which holds that random and non-random evolution act in isolation. This is not true of natural populations (Andrews, 2010), and there are discussions ongoing about building a better neutral theory. Some have come up with a theory called the nearly neutral theory. Some are coming up with something else altogether. The process is ongoing. The more sequence data, the better population geneticists will be able to form a better null hypothesis.

CONCLUSION

My analysis, like those preceding mine, was inconclusive with respect to the possible role of FUT2 in the immune system. Despite these inconclusive results, two major conclusions can still be drawn from the data presented in this analysis: It is possible that FUT2 is undergoing directional selection in the ASN superpopulation. If so, this would indicate that FUT2 is an adaptation to some kind of force, likely some kind of infectious disease. It is also possible that FUT2 is undergoing no evolution at all in any of the 14 populations. If so, this would indicate that FUT2 may not even have an important role in the immune system. Both conclusions have direct implications for public health, because investigating the links between natural selection and infectious disease resistance can reveal the forces that shaped the present state of the human immune system. These conclusions, in addition to previous inconclusive studies, emphasize the need to include a better model of demography in the null hypothesis in population genetics.

BIBLIOGRAPHY

Ali S, Niang MAF, N'doye I, Critchlow CW, Hawes SE, Hill AVS, Kiviat NB. Secretor polymorphism and human immunodeficiency virus infection in Senegalese women. J Infect Dis. 2000;181(2):737-739.

Altshuler DM, Gibbs RA, Peltonen L, et al. International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. Nature. 2010;467(7311):52-58.

Andrews, CA. Natural selection, genetic drift, and gene flow do not act in isolation in natural populations. Natural Education Knowledge. 2010;3(10):5.

Buchanan CC, Torstenson ES, Bush WS, Ritchie MD. A comparison of cataloged variation between International HapMap Consortium and 1000 Genomes Project data. J Am Med Inform Assoc. 2012;19:289-294.

Campi C, Escovich L, Borras SG, Racca L, Racca A, Cotorruelo C, Biondi C. Expression of the gene encoding secretor type galactoside 2 α fucosyltransferase (FUT2) and ABH antigens in patients with oral lesions. Med Oral Patol Oral Cir Bucal 2012 Jan;17(1):e63-8.

Clark AG, Glanowski S, Nielsen R, Thomas PD, Kejariwal A, Todd MA, Tanenbaum DM, Civello D, Lu F, Murphy B, et al. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. Science. 2003;302:1960-1963.

D'Adamo PJ, Kelly GS. Metabolic and Immunologic Consequences of ABH Secretor and Lewis Subtype Status. Alternative Medicine Review. 2001;6(4):390-405.

Davies Huw TO; Crombie, Iain K. What are confidence intervals and p-values? In: What is…? series. UK: Hayward Medical Communications, Hayward Group Ltd; 2009.

Dean, Laura. Chapter 5: The ABO Blood Group. In: Blood Groups and Red Cell Antigens. Bethesda (MD): National Center for Biotechnology Information (US); 2005.

Fay JC, Wu C. Hitchhiking under positive Darwinian selection. Genetics. 2000 Jul;155:1405-1413.

Ferrer-Admetlla A, Bosch E, Sikora M, Marques-Bonet T, Ramirez-Soriano A, Muntasell A, Navarro A, Lazarus R, Calafell F, Bertranpetit J, et al. Balancing selection is the main force shaping the evolution of innate immunity genes. J Immunol. 2008;181:1315-1322.

Ferrer-Admetlla A, Sikora M, Laayouni H, Esteve A, Roubinet F, Blancher A, Calafell F, Bertranpetit J, Casals F. A natural history of FUT2 polymorphism in humans. Mol Biol Evol. 2009;26(9):1993-2003.

Ford MJ. Applications of selective neutrality tests to molecular ecology. Molecular Ecology. 2002;11:1245-1262.

Guo ZH, Xiang D, Zhu ZY, Wang JL, Zhang JM, Liu X, Shen W, Chen HP. Analysis on FUT1 and FUT2 gene of 10 para-Bombay individuals in China. Zhonghua Yi Xue Yi Chuan Xue Za Zhi 2004 Oct;21(5):417-421.

Haldane, JBS. The rate of mutation of human genes. Hereditas. 1949 Dec;35(S1):267-273.

Hudson RR, Kreitman M, Aguade M. A test of molecular evolution based on nucleotide data. Genetics. 1987 May;116(1):153-159.

Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013 Apr;30(4):772-780.

Khan R. Natural selection of a human gene: FUT2 [Internet]. New York: Seed Media Group. 2009 June 3. Cited 2014 April 11. Available from: http://scienceblogs.com/gnxp/2009/06/03/natural-selection-of-a-human-g/.

Kindberg E, Hejdeman B, Bratt G, Wahren B, Lindblom B, Hinkula J, Svensson L. A nonsense mutation (428G—>A) in the fucosyltransferase FUT2 gene affects the progression of HIV-1 infection. AIDS. 2006;20:685-689.

Koda Y, Soejima M, Kimura H. The polymorphisms of fucosyltransferases. Leg Med (Tokyo). 2001 Mar;3(1):2-14.

Koda Y, Tachida H, Pang Hao, Liu Y, Soejima M, Ghaderi AA, Takenaka O, Kimura H. Contrasting patterns of polymorphisms at the ABO-secretor gene (FUT2) and plasma α(1,3)fucosyltransferase gene (FUT6) in human populations. Genetics. 2001 Jun;158(2):747-756.

Linden S, Mahdavi J, Semino-Mora C, Olsen C, Carlstedt I, Boren T, Dubois A. Role of ABO secretor status in mucosal innate immunity and H. pylori infection. PLoS Pathog. 2008 Jan;4(1):e2.

Liu Y, Koda Y, Soejima M, Pang H, Schlaphoff T, du Toit ED, Kimura H. Extensive polymorphism of the FUT2 gene in an African (Xhosa) population of South Africa. Hum Genet. 1998;103:204-210.

McDonald JH, Kreitman M. Adaptive protein evolution at the Adh locus in Drosophila. Nature. 1991 Jun 20;351(6328):652-654.

McGovern DPB, Jones MR, Taylor KD, Marciante K, Yan X, Dubinsky M, Ippoliti A, Vasiliauskas E, Berel D, Derkowski C, et al. Fucosyltransferase 2 (FUT2) non-secretor status is associated with Crohn's disease. Human Molecular Genetics. 2010 June;19(17):3468-3476.

Misawa K, Tajima F. Estimation of the amount of DNA polymorphism when the neutral mutation rate varies among sites. Genetics. 1997 Dec;147(4):1959-1964.

Moreno A, Campi C, Escovich L, Borras SG, Racca L, Racca A, Cotorruelo C, Biondi C. Analysis of the FUT2 gene and Secretor status in patients with oral lesions. Immunologia. 2009;28(3):131-134.

Nachman, Michael W. Chapter 7: Detecting selection at the molecular level. In: Fox, Charles W.; Wolf, Jason B., editors. Evolutionary Genetics: Concepts and Case Studies. New York (NY): Oxford University Press; 2006.

Narimatsu H, Iwasaki H, Nakayama F, Ikehara Y, Kudo T, Nishihara S, Sugano K, Okura H, Fujita S, Hirohashi S. Lewis and Secretor gene dosages affect CA19-9 and DU-PAN-2 serum levels in normal individuals and colorectal cancer patients. Cancer Res. 1998;58:512-518.

Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG. Recent and ongoing selection in the human genome. Nat Rev Genet. 2007 Nov 8;8(11):857-868.

Ramirez-Soriano A, Ramos-Onsins SE, Rozas J, Calafell F, Navarro A. Statistical power analysis of neutrality tests under demographic expansions, contractions and bottlenecks with recombination. Genetics. 2008 May;179(1):555-567.

Rydell GE. Susceptibility to winter vomiting disease: a sweet matter. Rev Med Virol. 2011 Nov;21(6):370-382.

Sabeti P. Natural Selection: Uncovering Mechanisms of Evolutionary Adaptation to Infectious Disease. Nature Education. 2008;1(1):13.

Soejima M, Koda Y. TaqMan real-time polymerase chain reaction for detection of SEC1-FUT2 hybrid alleles: Identification of novel hybrid allele. Clinica Chimica Acta. 2013;415:59-62.

Soejima M, Pang H, Koda Y. Genetic variation of FUT2 in a Ghanaian population: identification of four novel mutations and inference of balancing selection. Ann Hematol. 2007 Mar;86(3):199-204.

Tabasum ST, Nayak RP. Salivary blood group antigens and microbial flora. Int J Dent Hygiene. 2011;9:117-121.

Thorven M. A homozygous nonsense mutation (428G—>A) in the human secretor (FUT2) gene provides resistance to symptomatic norovirus (GGII) infections. J Virol. 2005 Dec;79(24):15351-15355.

Wacklin P, Makivuokko H, Alakulppi N, Nikkila J, Tenkanen H, Rabina J, Partanen J, Aranko K, Matto J. Secretor genotype (FUT2 gene) is strongly associated with the composition of bifidobacteria in the human intestine. PLoS ONE. 2011;6(5):e20113. doi:10.1371/journal.pone.0020113.

Walsh EC, Sabeti P, Hutcheson HB, Fry B, Schaffner SF, de Bakker PIW, Varilly P, Palma AA, Roy J, Cooper R, et al. Searching for signals of evolutionary selection in 168 genes related to immune function. Hum Genet. 2006;119:92-102.

Weatherall, David; Greenwood, Brian; Leng Chee, Heng, et al. Chapter 5: Science and Technology for Disease Control: Past, Present, and Future. In: Jamison, Dean T.; Breman, Joel G.; Measham, Anthony R., et al, editors. Disease Control Priorities in Developing Countries. 2nd ed. Washington (DC): World Bank; 2006.

Zhai W, Nielsen R, Slatkin M. An investigation of the statistical power of neutrality tests based on comparative and population genetic data. Mol Biol Evol. 2009;26(2):273-283.