Evidence of Purifying Selection in Mammals

John Long

MIT PRIMES

Final paper

2013

Abstract:

The Human Genome Project completed in 2003 gave us a reference genome for the human species. Before the project was completed, it was believed that the primary function of DNA was to code for protein. However, it was discovered that only 2% of the genome consists of regions that code for proteins. The remaining regions of the genome are either functional regions that regulate the coding regions or junk DNA regions that do nothing. The distinction between these two types of regions is not completely clear. Evidence of purifying selection, the decrease in frequency of deleterious mutations, is likely a sign that a region is functional. The goal of this project was to find evidence of purifying selection in newly acquired regions in the human genome that are hypothesized to be functional. The mean Derived Allele Frequency of the featured regions was compared to that of control regions to determine the likelihood of selection.

I. Introduction

The Human Genome Project was formalized in 1987 with the goal of furthering the understanding of the Human Genome (*1*). Research began in 1990, and the project was projected to take 15 years (*2*). However, in 2003, 2 years earlier than planned, a working draft of the genome was published, and the project was announced complete.

The project sequenced the complete human genome for the first time, including all 21 autosomes and 2 sex chromosomes. Nevertheless, it still left much to be discovered in terms of the biochemical mechanisms behind the functions of our genetic sequences.

Genome

The human genome can generally be divided into three categories: genes, regulatory regions, and junk regions.

Genes are regions of the genome that code for protein. On a textbook level, DNA is transcribed to mRNA, which is similar to DNA except that it has the base Uracil in place of Thymine (see Figure 1). The mRNA is processed and then translated into a protein on the surface of a ribosome. The specific proteins that are synthesized then influence the phenotype that is expressed.

However, genes only make up only 2% of the genome, and the processes of transcription and translation must be regulated in some way. Regulatory regions are sequences in the genome that can help in the process of gene regulation.

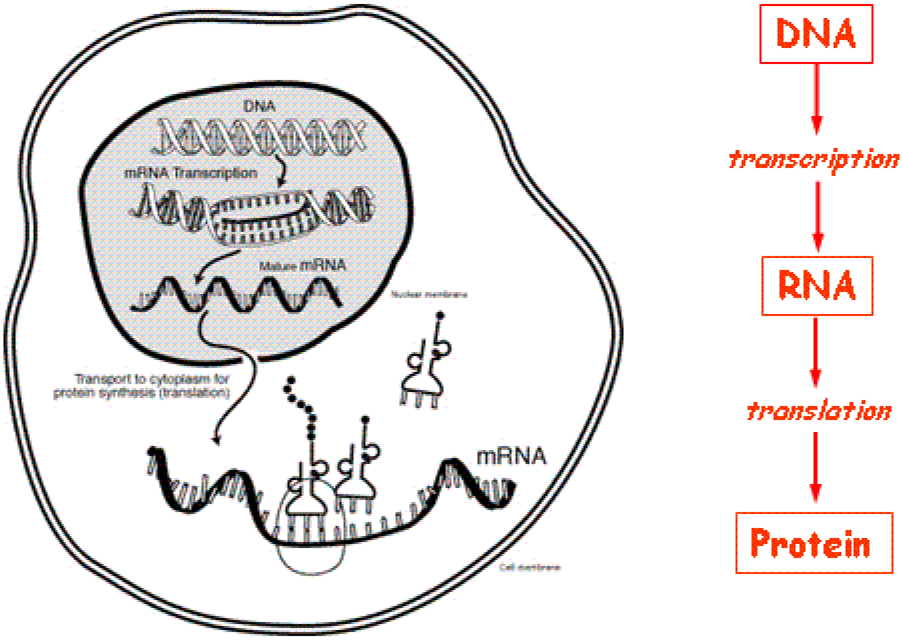

Figure 1: Transcription and translation in a cell (*3*).

One of the textbook examples of a regulatory region in organisms is the Lac Operon present in E. coli (see Figure 2). The Lac Operon regulates the transcription of Lactase, an enzyme used to break up Lactose. Normally, a repressor protein is translated, which binds to the region where the genes for coding Lactase are located. This prevents the transcription factor from binding and beginning translation of Lactase. However, when Lactose is present, it binds to the repressor protein and changes its shape. The repressor protein is no longer completely bound to the Lactase coding sites, which allows the transcription factor to bind, causing Lactase to be expressed. The Lactase then breaks down the Lactose, decreasing the Lactose levels. As these levels decrease, there is fewer Lactose molecules present to bind to the repressor proteins thus the repressor proteins are free to bind with the Lactase gene’s Promoter region once again. This is particularly remarkable, since it is an example of an efficient self-regulating system.

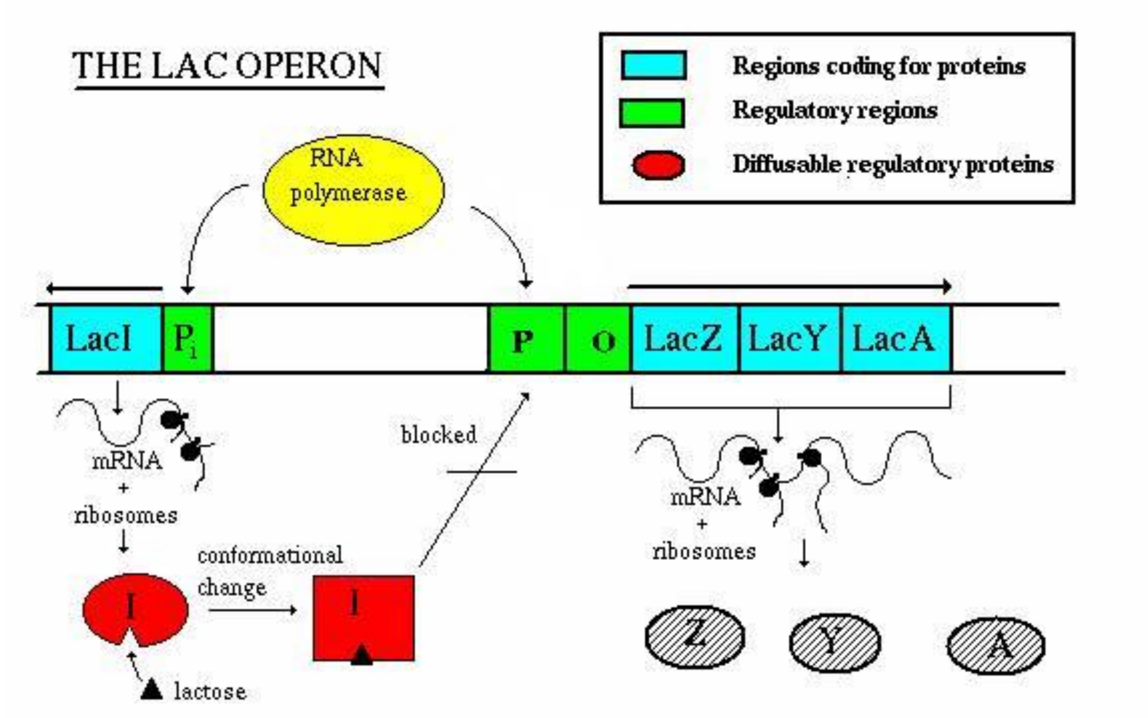

Figure 2: Lactose regulation by the Lac Operon (*4*).

Junk regions, in contrast to the previously described regions, neither code for protein nor serve any regulatory purpose. These regions are neither harmful nor beneficial, so there is no reason for them to be selected against. They are passed under neutral selection over evolutionary time.

Due to a distinct evolutionary signature for genes, we have come a long way in identifying gene regions in the genome. However, both junk regions and regulatory regions can be difficult to identify. Also, the exact process by which regulatory regions regulate genes is often complicated not always known. Luckily, in the field of computational biology, we can utilize genome-wide analysis techniques to computationally identify regions theorized to have some purpose, and experimentalists can later verify these findings. In this report, we describe a computational method by which we can provide further evidence that certain groups of annotated functional regions are not only functional but also evolutionarily preserved. This method can then be generalized to provide evidence of human-specific selection.

Selection

From a genetic standpoint, natural selection acts on populations to reduce the prevalence of deleterious alleles and increase the frequency of beneficial alleles. For example, the allele or alleles leading to the giraffes having an enlarged neck likely enabled those giraffes to reach leaves on tall trees that were out of reach for other giraffes. They likely survived at a higher rate as a result, while many shorter-necked giraffes may have died out, due to competition for food, and the longer necked giraffes were able to pass on their genetic material. In a human- specific example, the allele responsible for sickle-cell anemia also contributes to immunity for the Malaria virus. As a result of a single point mutation, a slightly defective protein is created, which makes it hard for the virus to latch onto blood cells.

Two types of natural selection are positive selection and purifying selection. Generally, when people think of natural selection and “survival of the fittest,” they think of positive selection, which involves the increase in frequency of those alleles that are beneficial to a population. However, positive selection is generally not as common as negative selection, or purifying selection, which is the decrease in frequency of deleterious alleles. Purifying selection is likely more common than positive selection because after years of evolution in a population, a delicate equilibrium has been reached, resulting in complex organisms. Therefore, in such complicated organisms, a random mutation is more likely harmful than beneficial. Purifying selection plays an important role in maintaining the equilibrium.

Significance of Selection

Evidence of natural selection would suggest that a point or region of the genome is important. If a region is not important, then there will be no advantage of having one variant over another. The choice of alleles is random, and over time, the alleles would be equally distributed in frequency. However, if a region is important, then it is likely that only certain variants of the region would provide the desired function. This signal would manifest itself in an unequal distribution frequency of certain variants.

Measuring Selection

Selective pressure can be approximated at a specific locus by using the allele frequency. At a specific locus, usually there will only be two possible alleles that show up in the population at an observable frequency. The ancestral allele is then defined as the allele present in the ancestral population and the derived allele is defined as the second allele, the allele different from the ancestral allele. In the case of humans, the ancestral allele was given by the Chimpanzee genome and already annotated in the 1000 genomes data.

The average derived allele frequency (DAF) of common single nucleotide polymorphisms (SNPs) in humans is an appropriate measure of selection for regions of the genome for two reasons. Firstly, the majority of the human genome is identical across all humans (5). Secondly, the few genetic differences can take the form of SNPs, deletions, or insertions, but due to the biological phenomenon of linkage disequilibrium, we know that the frequency of nearby variants is tightly linked. Therefore, the frequency of any variant in the genome can be approximated by averaging the frequencies of the common SNPs that fall nearby. Therefore, we will estimate and aggregate the selective pressure on a region by averaging the DAFs of SNPs in that region.

Coverage

In the process of sequencing whole genomes, the result for each nucleotide position must be obtained more than once for accuracy, as sequencing errors happen at a non-negligible frequency. Therefore, many short reads that overlap each other are obtained and pieced together to generate the complete sequence.

The read depth or coverage gives a measure of confidence in the resulting sequence by providing the average number of times a base pair was covered over all the reads. It can be calculated by the formula:

coverage=(number of reads)*(average length of read)/(total length of region sequenced)

Conventionally, coverage is calculated on a per-chromosome basis, since chromosomes are naturally separated units of the genomes. The average coverage of a region can be calculated simply by taking the average of the coverage scores of each SNP falling in the region.

Rare SNPs, which are defined as SNPs with low DAF, are unlikely to appear when the coverage is low. Therefore, we control for coverage such that regions with equal coverage are eventually compared.

Regions

This project will look for evidence of purifying selection in exonic regions vs random intergenic regions to test the general pipeline. These regions were downloaded from the gencode site. Regulatory regions, such as miRNA target regions, will then be investigated for evidence of purifying selection, although the pipeline is general enough to accommodate any type of regions. We are especially interested in miRNA target regions and other regulatory regions because evidence of selection in these regions would confirm their importance.

A micro RNA (miRNA) is a short non-coding RNA molecule that aids in gene regulation. It may bind to regions of the mRNA to prevent translation from occurring (*6*). The miRNAs usually bind to target regions in the 3'UTRs following the genes in which they regulate.

II. Methods

These methods were adapted from the methods described in the journal paper Evidence of Abundant Purifying Selection in Humans for Recently Acquired Regulatory Functions (*7)*. The general method, which can be adapted to testing other regions, is described as follows. A file containing regions is first downloaded. Next, randomized regions of the same length as the control regions are generated. Then, SNPs are overlapped with these regions to generate a measurement of coverage and DAF for each region. Finally, the plot of the regions is analyzed and evaluated for signs of selection.

To validate the pipeline, exonic regions were compared with intergenic regions. It has been previously shown that exonic regions are functional and exhibit signs of selective pressure; therefore, we expect them to have a lower mean DAF than comparable random intergenic regions. The pipeline will be discussed in more depth using exonic regions as an example of regions given as input to the pipeline.

Table 1: VCF file first five lines (*8*).

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
1	10291	.	C	T	7981	Truth Sensitivity Tranche 99.94to 100.00	AC=301;AF=0.3195;AN=942;BaseQRankSum=5.549;CalledBy=UM;DP=11556;DS;Dels=0.00;FS=110.558;HRun=0;HaplotypeScore=4.4214;InbreedingCoeff=0.0257;MQ=14.84;MQ0=7941;MQRankSum=-2.209;QD=2.58;ReadPosRankSum=0.639;SB=-3277.82;VQSLOD=-18.5766;pop=ALL
1	10303	.	C	T	477.62	Truth Sensitivity Tranche 99.94to 100.00	AC=52;AF=0.0570;AN=912;BaseQRankSum=-5.482;CalledBy=UM;DP=9636;DS;Dels=0.00;FS=13.179;HRun=0;HaplotypeScore=4.7756;InbreedingCoeff=-0.0594;MQ=17.24;MQ0=5844;MQRankSum=2.401;QD=0.71;ReadPosRankSum=2.903;SB=-412.62;VQSLOD=-17.3319;pop=ALL
1	10309	.	C	T	380.57	Truth Sensitivity Tranche 99.94to 100.00	AC=49;AF=0.0503;AN=974;BaseQRankSum=-7.661;CalledBy=UM;DP=9655;DS;Dels=0.00;FS=0.892;HRun=0;HaplotypeScore=4.9360;InbreedingCoeff=0.0973;MQ=18.11;MQ0=5494;MQRankSum=2.293;QD=0.45;ReadPosRankSum=2.613;SB=-308.12;VQSLOD=-17.4737;pop=ALL
1	10315	.	C	T	992.9	Truth Sensitivity Tranche 99.94to 100.00	AC=103;AF=0.08759;AN=1176;BaseQRankSum=-6.800;CalledBy=UM;DP=10311;DS;Dels=0.00;FS=1.029;HRun=0;HaplotypeScore=4.6195;InbreedingCoeff=-0.0567;MQ=19.02;MQ0=5227;MQRankSum=1.588;QD=0.66;ReadPosRankSum=8.991;SB=-538.39;VQSLOD=-14.5919;pop=ALL
1	10457	.	A	C	222.14	Truth Sensitivity Tranche 99.94to 100.00	AC=20;AF=0.01312;AN=1524;BaseQRankSum=2.411;CalledBy=NCBI;DP=13055;DS;Dels=0.00;FS=32.763;HRun=0;HaplotypeScore=2.2778;InbreedingCoeff=0.0801;MQ=22.83;MQ0=6854;MQRankSum=2.240;QD=0.58;ReadPosRankSum=-0.375;SB=-107.65;VQSLOD=-6.3002;pop=ALL

A file in VCF format containing variants was downloaded from the 1000 genomes project (*8*) (see Table 1). Only the CHROM, POS, REF, ALT, and INFO fields were needed. CHROM denotes the chromosome number, POS the position, REF the reference allele, ALT the alternate allele. Information for DP, AN, and AF could be found in the INFO field.

Table 2: Output from VCFtools:

CHROM	POS	REF	ALT	AA	AF	AN	DP
1	10291	C	T	?	0.3195	942	11556
1	10303	C	T	?	0.057	912	9636
1	10309	C	T	?	0.0503	974	9655
1	10315	C	T	?	0.0875	1176	10311
1	10457	A	C	?	0.0131	1524	13055
1	10469	C	G	?	0.0397	1232	3320
1	10492	C	T	?	0.1044	1216	3035
1	10575	C	G	?	0.003	334	741
1	10583	G	A	?	0.1979	960	2337

The information was parsed using VCFtools (*9*) (see Table 2). The majority of variants were SNPs. Variants such as insertions and deletions were removed. Next, lines containing more than one reference allele or more than one alternate allele were removed. Only SNPs remained after conducting these processes. When the ancestral allele was not given in the VCF file, the allele that had a higher frequency was defined as the ancestral allele. Most often the reference allele corresponded with the ancestral allele. The coverage of each SNP was also calculated by dividing the total depth across all samples (DP) by the total number of samples (AN). Coverage =DP/AN

The SNP file with all the necessary information was converted to bed format. The chr tag was added to all chromosome numbers and the position. Next, the chromosome start coordinate was defined as the SNP position minus one and the end coordinate was defined as the given coordinate. In this way, the file was converted from the 1-based system used in VCF files to the 0-based system used in bed files. Then the file was sorted by the chromosome and start position columns. The chromosome column was sorted alphabetically and the position column was sorted numerically using the unix sort command.

Region files were downloaded from publicly available databases. Two files were downloaded. One file contained feature regions and the other file contained regions from which the control regions would be generated. For example, if a file containing exons was given as one input file, another file containing intergenic regions could be used as the second input file. If the files were not already in bed format, they were parsed and converted into bed format. Coordinates in the downloaded files were converted to the 0-based system by subtracting one from the start position if they used a 1-based system. Duplicate regions, which were defined as regions with the same chromosome number, start position, and end position, were deleted.

Often, the background region file should consist of regions which contain all the input regions. For example, one may use 3'UTRs or the entire genome as the background for miRNA targets. This makes it easier to generate random regions because then it is guaranteed that each input region is smaller than at least one region in the set of background regions, and it is always possible to generate a random region. To make sure the feature regions were indeed contained within the background regions they were intersected using bedintersect.

Random intergenic regions were then created based on the feature regions. We generated a random region of the same length as an input region, and if the random region was located in the intergenic regions defined by the gencode genes (10), it was added to the growing random region file. Otherwise, the region was dropped and a new random region would be continually generated until it fit in the set of background regions. This process was continued until we had identified matching intergenic regions for all the feature regions. Both files, the file containing feature regions and the file containing control regions, were then overlapped with the modified VCF file containing SNPs from the 1000 genomes project using bedtools (*11*). A mean DAF and mean coverage was calculated for each region by averaging the DAFs and coverages of the SNPs falling in the region. The feature and control regions together were then sorted by coverage. The bottom 5% and top 5% of regions were dropped.

The regions were then separated back into the feature vs random groups and binned by interval. The range of coverage was split into 20 equal intervals, and the regions in each bin were averaged by coverage and DAF. The resulting 20 points were then plotted DAF on the y-axis vs. coverage on the x-axis with error bars showing the standard error of the mean DAF in the bin. Standard error of the mean was calculated as the standard deviation of the bin divided by the square root of the number of points in the bin.

IV. Results

Exons vs. Intergenic Regions

Exonic Regions

Bin	Coverage	DAF	stderr
1	1.97666	0.0492288	0.001519184
2	2.04481	0.0475484	0.001333624
3	2.11226	0.0462967	0.001153468
4	2.17909	0.0442211	0.001106696
5	2.24618	0.0418876	0.000965436
6	2.31371	0.0407717	0.000917273
7	2.38086	0.0400819	0.000835317
8	2.44853	0.0379022	0.000733518
9	2.5153	0.0363079	0.000690937
10	2.58273	0.0353169	0.000637098
11	2.64949	0.0353796	0.000622351
12	2.71665	0.0340089	0.000608221
13	2.78362	0.0338626	0.000624697
14	2.85043	0.0333319	0.00062171
15	2.91767	0.031954	0.000632653
16	2.98459	0.0325116	0.000709274
17	3.05158	0.0322544	0.000751919
18	3.11814	0.0313508	0.000893723
19	3.18579	0.0316952	0.001031152
20	3.25283	0.0326999	0.001325898

Intergenic Regions

Bin	Coverage	DAF	stderr
1	1.97946	0.043142	0.001570726
2	2.04475	0.0501358	0.00178861
3	2.11312	0.04534	0.001221285
4	2.18104	0.044563	0.000980751
5	2.24701	0.0465953	0.000946051
6	2.31404	0.0458057	0.000920642
7	2.38165	0.0421392	0.000698095
8	2.44849	0.0454995	0.000674879
9	2.51526	0.0456842	0.00062616
10	2.58248	0.0436786	0.000569706
11	2.64953	0.0422026	0.000563664
12	2.71688	0.0402675	0.000531972
13	2.78365	0.040859	0.000523689
14	2.85066	0.0394748	0.000520314
15	2.91831	0.0393941	0.000531411
16	2.98463	0.0394173	0.000589611
17	3.05214	0.0375162	0.000697381
18	3.11896	0.0357543	0.000813551
19	3.18578	0.0365253	0.000932766
20	3.25257	0.0360121	0.001061836

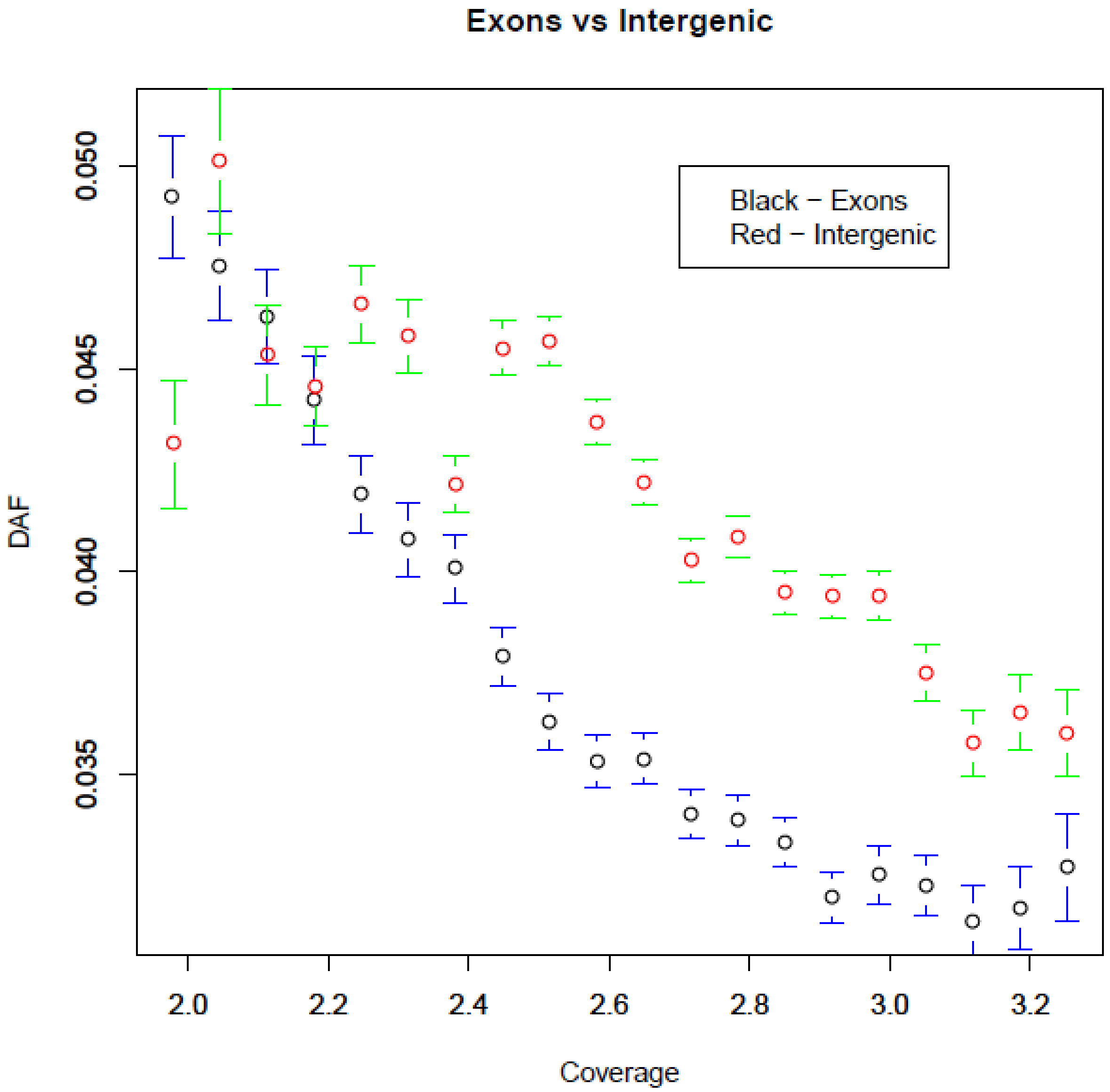

Figure 9: Binned regions for exonic regions (black) and intergenic regions (red) with standard error of the mean bars.

Figure 9 shows that the exonic regions generally have a lower DAF than that of randomized intergenic regions. However, the signal is more ambiguous when the coverage is low around 2. For randomized regions with very low coverage, the DAF is lower than that of the next bin.

miRNA vs. 3’UTR

miRNA

Bin	Coverage	DAF	stderr
1	2.03081	0.036165	0.002433
2	2.10278	0.034213	0.001938
3	2.17172	0.035	0.00162
4	2.24169	0.044015	0.001789
5	2.31064	0.033232	0.001271
6	2.38074	0.039829	0.001411
7	2.45179	0.045212	0.001408
8	2.52028	0.032867	0.001039
9	2.59001	0.036771	0.001155
10	2.66064	0.036656	0.001136
11	2.72988	0.033642	0.001035
12	2.7998	0.030969	0.001021
13	2.86958	0.032003	0.001038
14	2.93934	0.032827	0.001144
15	3.00857	0.033022	0.001222
16	3.07831	0.034995	0.001455
17	3.14787	0.025568	0.001251
18	3.21743	0.033884	0.001606
19	3.28785	0.024685	0.001766
20	3.35815	0.03137	0.002272

3’UTRs

Bin	Coverage	DAF	stderr
1	2.03189	0.051575	0.002283
2	2.10037	0.049371	0.001622
3	2.1729	0.055679	0.001816
4	2.24211	0.039538	0.001393
5	2.31166	0.041165	0.001417
6	2.38186	0.045509	0.001434
7	2.4501	0.046647	0.001283
8	2.52057	0.039703	0.001104
9	2.59107	0.047035	0.001095
10	2.66066	0.046401	0.001139
11	2.73003	0.043357	0.001112
12	2.80006	0.033534	0.000907
13	2.8702	0.038344	0.000985
14	2.93895	0.035648	0.001205
15	3.00842	0.040056	0.001199
16	3.07932	0.040329	0.001534
17	3.14699	0.030017	0.001445
18	3.21649	0.033999	0.0016
19	3.28766	0.026761	0.001542
20	3.35871	0.033632	0.002462

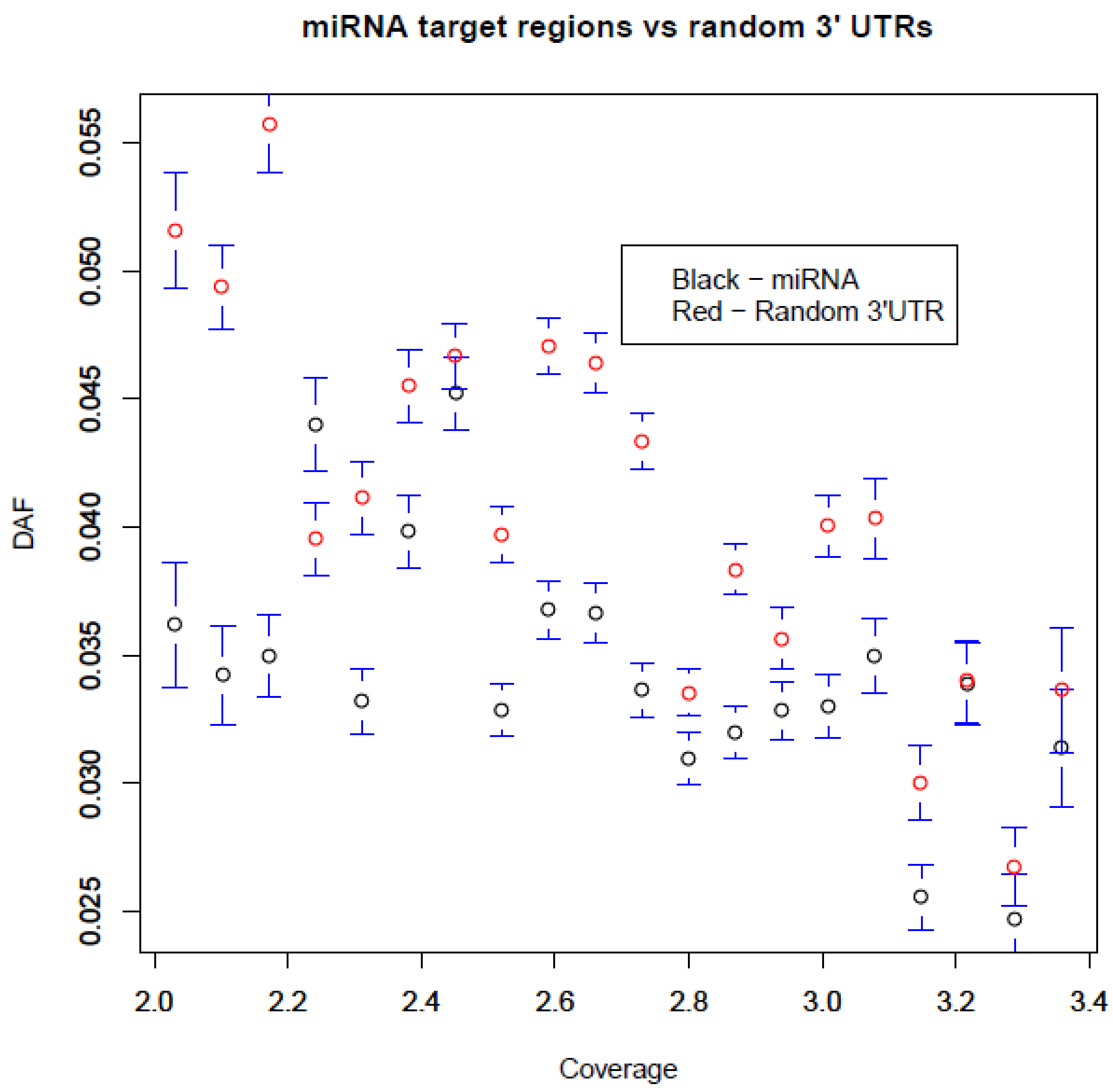

Figure 10: Binned regions for miRNA target regions (black)
and random 3’UTRs (red)
with standard error of the mean bars.

V. Discussion

The results strongly suggest that the DAF of exonic regions is lower than the DAF of general intergenic regions. The cause of this signal is most likely purifying selection, which agrees with the intuition that exons are functional.

There is a similar signal in the miRNA graph although the signal is not as clear is that in the previous test. This suggests that miRNA target sequences are relatively more constrained than random regions found within 3’UTRs but the signal is not as clear as in the previous test because 3’UTRs themselves may be relatively constrained when compared to regions such as intergenic regions.

However, the current pipeline does not correct for conservation, which is necessary for providing evidence of purifying selection in human-specific regions. Methods for modifying the pipeline are discussed in the next section. There is reason to believe that the signal for purifying selection would still be clear based on previous work done (*7*).

VI. Further Research

Because conserved regions generally have a low average DAF already, future work to identify human-specific selection would involve first separating the feature regions into conserved and unconserved. Regions would be divided into conserved regions and unconserved regions based on overlap with siphy elements (*17*), which are regions computed to have high conservation scores. A region would be considered conserved if 90% of the region overlapped with a siphy element and all other regions would be considered unconserved.

We also plan to extend this approach to other regulatory regions such as 5’UTR stem loops and Exonic Splicing Enhancers, as our pipeline is fully generalizable.

The 5’ Untranslated Region (5’ UTR) Stem Loop is a region on the 5’ end of DNA which sometimes forms a hairpin loop. There has recently been evidence that this region is essential for regulating translation of Growth Factor B1 (*18*).

Exonic Splicing Enhancers (ESEs) are sequences that somehow aid the splicing machinery in splicing and removing introns. Prior to translation, introns must be removed from the mRNA strand. The remaining regions known as exons are then transported to the ribosome to become translated into protein. Certain regions may help the splicing mechanism find places to delete (*19*).

Progress was also made on packaging the pipeline and making it generally available to the research community, so that other researchers may use it to quickly obtain a measurement of selection in putative functional regions. Currently the tool is functional, allowing users to run a command to compare selection and generate a plot of the results. However, efforts are being made to control for the other factors and make the tool run faster before being released.

Acknowledgments

I would like to thank all those who helped make this project possible. First, I would like to thank the PRIMES program for giving me the opportunity to perform this research. I would like to thank Dr. Manolis Kellis, who helped provide the idea for this project. I would also like to thank my mentor, Angela Yen, who guided me throughout the project. Finally, I would like to thank my parents who supported me throughout the project.

References

1. DeLisi, Charles (2008). "Meetings that changed the world: Santa Fe 1986: Human genome baby-steps". Nature 455 (7215): 876.

2. About the Human Genome Project. (n.d.).*Oak Ridge National Laboratory*. Retrieved May 4, 2013, from http://www.ornl.gov/sci/techresources/Human_Genome/project/about.shtml

3. BIOL414/614 at UMBC - RNA_Editing_Introduction. (n.d.). *UMBC: An Honors University In Maryland*. Retrieved May 5, 2013, from http://www.umbc.edu/bioclass/biol414/wiki/index.php?page=RNA_Editing_Introduction

4. The Lactose Operon. (n.d.). *Operons Uncovered: The dirty truth about the lactose and arabinose operons*. Retrieved May 4, 2013, from userpages.umbc.edu/~krebe1/biology302l/?p=LactoseOperon

5. Collins, F. S.; Brooks, L. D.; Chakravarti, A. (1998). "A DNA polymorphism discovery resource for research on human genetic variation". Genome research 8(12): 1229–1231.

6. Marı´n RM, Vanı´cˇek J (2012) Optimal Use of Conservation and Accessibility Filters in MicroRNA Target Prediction. PLoS ONE 7(2): e32208. doi:10.1371/journal.pone.0032208

7. Ward, L., & Kellis, M. (2012). Evidence of Abundant Purifying Selection in Humans for Recently Acquired Regulatory Functions. *Science*, *337*, 1675-1678.

8. An integrated map of genetic variation from 1,092 human genomes, McVean et Al, Nature 491, 56–65 (01 November 2012) doi:10.1038/nature11632.

9. *The Variant Call Format and VCFtools*, Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert Handsaker, Gerton Lunter, Gabor Marth, Stephen T. Sherry, Gilean McVean, Richard Durbin and 1000 Genomes Project Analysis Group, **Bioinformatics**, 2011

10. Harrow J, et al. (2012) GENCODE: The reference human genome annotation for The ENCODE Project" (**PubMed**)

11. Quinlan AR and Hall IM, 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 26, 6, pp. 841–842.

12. Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. Betel D, Koppal A, Agius P, Sander C, Leslie C., *Genome Biology* 2010 11:R90

13. The microRNA.org resource: targets and expression. Betel D, Wilson M, Gabow A, Marks DS, Sander C., *Nucleic Acids Res.* 2008 Jan; 36(Database Issue): D149-53.

14. Human MicroRNA targets. John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS., *PLoS Biol.* 2005 Jul;3(7):e264.

15. MicroRNA targets in Drosophila. Enright AJ, John B, Gaul U, Tuschl T, Sander C and Marks DS., *Genome Biology* (2003)**5**;R1

16. Grillo G, Turi A, Licciulli F, Mignone F, Liuni S, Banfi S, Gennarino VA, Horner DS, Pavesi G, Picardi E, et al. UTRdb and UTRsite (RELEASE 2010): a collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs. Nucleic Acids Res. 2009;38:D75–D80.

17. Garber, M. et al., *Identifying novel constrained elements by exploiting biased substitution patterns*. **Bioinformatics** 25, i54-62, doi:btp190 [pii] 10.1093/bioinformatics/btp190 (2009).

18. Jenkins RH, Bennagi R, Martin J, Phillips AO, Redman JE, et al. (2010) A Conserved Stem Loop Motif in the 59Untranslated Region Regulates Transforming Growth Factor-b1 Translation. PLoS ONE 5(8): e12283. doi:10.1371/journal.pone.0012283

19. Fairbrother WG, Holste D, Burge CB, Sharp PA (2004) Single nucleotide polymorphism–based validation of exonic splicing enhancers. PLoS Biol 2(9): e268.