Journal of
Biology, Biological Systems and Bioinformatics
Volume 4, Issue 1, 2013

Analysis of the Positively and Non-Positively Selected Non-Protein Coding Sequences of Human Chromosome 16

John Snedeker, undergraduate student, Wheaton College, IL, john.snedeker@my.wheaton.edu
Kyle Tretina, undergraduate student, Wheaton College, IL, kyletretina@gmail.com
Meredith Triplet, undergraduate student, Wheaton College, IL, meredith.triplet@my.wheaton.edu
David Lee, undergraduate student, Wheaton College, IL, david.m.lee@my.wheaton.edu
Anne Poirier, undergraduate student, Wheaton College, IL, annie.poirier@my.wheaton.edu
Pattle Pun, Professor, Wheaton College, IL, pattle.pun@wheaton.edu
John Hayward, Associate Professor, Wheaton College, IL, john.hayward@wheaton.edu
Ross Ka-Kit Leung, Post-doctoral fellow, The Chinese University of Hong Kong, China, ross@cuhk.edu.hk
Stephen Kwok-Wing Tsui, Professor, The Chinese University of Hong Kong, China, kwtsui@cuhk.edu.hk

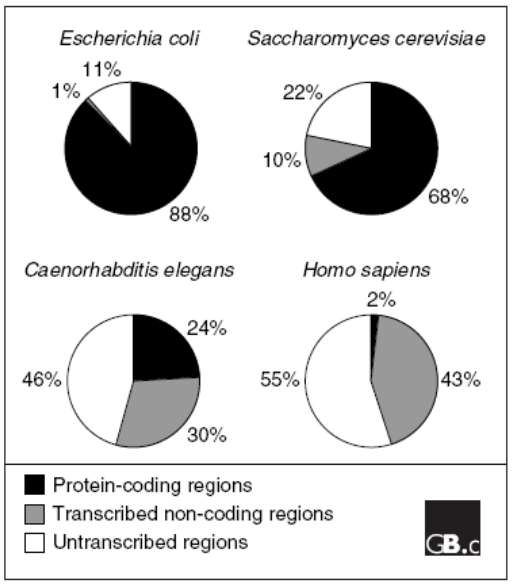
Abstract

The majority of the human genome consists of non-protein coding sequences of unknown function. A pipeline for predicting the functionality of these sequences utilizing selection algorithms from the HapMap project to identify SNPs, a mirror UCSC Genome Browser site to collect SNP flanking sequences, and finally the TRANSFAC database to discover homology to known regulatory sites is described herein. It was found that around three quarters of the non-coding SNP flanking sequences of human chromosome 16 (a) may play a significant role in transcription regulation, and (b) are on average 5kb closer to genes than non-coding SNPs as a whole.

Introduction

Francis Crick's "central dogma" of molecular biology can be simply interpreted as "DNA makes RNA, which makes proteins." The completion of the human genome project and the subsequent analysis of the functionality of human DNA have revealed that only about 2% of the genome participates in the process indicated by this central dogma, with around 98% of the human genome defined as non-protein coding. When this percentage is compared to other genomes, a puzzling discord arises between the amount of coding DNA and the apparent complexity of an organism. This is known as the G-value paradox [8] (Fig. 1). For example, humans have less than twice the number of genes (20,500) as a fruit fly (14,000**)** but are seemingly much higher in complexity [1, 17, 22]. Higher complexity seems to be correlated with the relative increase of the non-coding regions of an organism, but not the protein-coding region [21].

Figure 1. Ratios of the protein-coding, non-coding and untranscribed sequences in bacterial, yeast, nematode and mammalian genomes [20].



When faced with this paradox, scientists proposed mechanisms for increased protein diversity in mammals, such as the use of multiple transcription start sites, alternative pre-mRNA splicing and polyadenylation, pre-mRNA editing, and post-translational modifications, with alternative splicing considered as the most important source of protein diversity 15. However, this idea warrants serious reconsideration in light of the discovery that mammals do not have considerably higher levels of alternative splicing than organisms in other phyla, such as insects, which are thought to be much less complex [3].

Current theories that seek to explain the complexity of organisms must rely not solely on the size and diversity of the transcriptome, but also the complex regulation of gene expression [20]. Gene regulation in eukaryotes involves a complicated system of transcription factors that activate or repress genes via DNA binding sites [14]. Within an organism, each cell type in a certain tissue at each developmental stage has its own custom gene expression profile due a unique combination of transcription factors. The specialized control of transcription factors over gene expression seems to offer a possibility for further explanation of the G-value paradox. If the non-coding region is rich with transcriptional control sites and these transcriptional controls are critical in the complex assortment and processing of gene products in humans, then the importance of a non-coding region for complexity as a pattern throughout nature would be an intriguing hypothesis.

Genetic variation has been studied in the context of diseases that afflict humans [10, 18]. The search for disease genes was built off the first human genome reference sequence and the subsequent HapMap [23]. HapMap describes blocks of DNA tagged by common single nucleotide base variants (single nucleotide polymorphisms, or SNPs) by determining whether a particular SNP is positively selected or not. Positive natural selection describes the phenomenon that the prevalence of traits advantageous to reproduction tends to increase in a given population. Positive selection is detected using five “signatures” or patterns of genetic variation in DNA sequences: high proportion of function-altering mutations, reduction in genetic diversity, high frequency derived alleles, differences between populations, and long haplotypes [19]. Only SNPs that were recently under selective pressure and were found in all four HapMap ethnic groups were considered here in order to conservatively denote the most likely sites of transcription factor binding (see methods). These recently selective pressures could include any widespread selective effects around the world, from rising average global temperatures to technological changes in modern society. A study of non-genic sequences of chromosome 21 showed that highly conserved SNPs are present within non-coding regions on the chromosome [6]. The fact that some non-protein coding SNPs are highly conserved indicates that they are either functionally important, possibly possessing structural roles or roles involved with gene expression, or that they are simply genetic markers of a more extended haplotype that are merely co-inherited with the causal variant. This ambiguity in the interpretation of genetic variation is a common issue in linkage studies [11]. For example, a non-coding SNP may be located at a site which allows for both transcription factor binding and the transcription of non-coding RNA, making it difficult to determine the reason for selection at any given site.

While prokaryotic gene expression is generally divided into discrete operons with transcription factors relatively nearby and upstream of the site of transcription initiation, in eukaryotes this is not the case. For example, eukaryotic cis-acting transcription factors have been known to be more than 50 kilobases away from the genes that they regulate [7], either upstream or downstream, acting on those genes via a combination of binding to their appropriate sites and DNA looping. Since sequence proximity correlates to mitotic segregation percentages (as it is the basis of linkage studies), being close to genes may actually indicate that these SNPs are linkage markers rather than transcription factor binding sites. On the other hand, proximity to genes may have higher correlation with certain gene regulation functions. Since TRANSFAC includes both cis and trans-acting elements, further research would need to be performed in order to determine the relevance of gene proximity to transcription factor binding sites in eukaryotes.

The role of the vast amount of non-coding sequences in the human genome may be further elucidated through the analyses of SNPs, transcription factor related areas, and the effect of selection on the sequences. A technique for collecting transcription regulation information of the non-coding regions of chromosome 16, as well as a process for characterizing some regulatory differences between recently positively selected (Pos) SNP flanking sequences and recently non-positively selected (NonPos) sequences is described herein. SNPs from the non-protein coding regions identified from UCSC Genome Browser were screened using Haplotter in the Phase II HapMap Database to determine selection. SNP flanking sequences from dbSNP build 135 with perfect re-alignment in the Genome Browser were processed through the PATCH program of the TRANSFAC database to search for potential roles in transcription.

Materials and Methods

HapMap Data Collection

The International HapMap Consortium has characterized 3.1 million human SNPs, genotyped in 270 individuals from four geographically diverse populations. In this study, their Phase II data [9] are utilized. Their database has been made

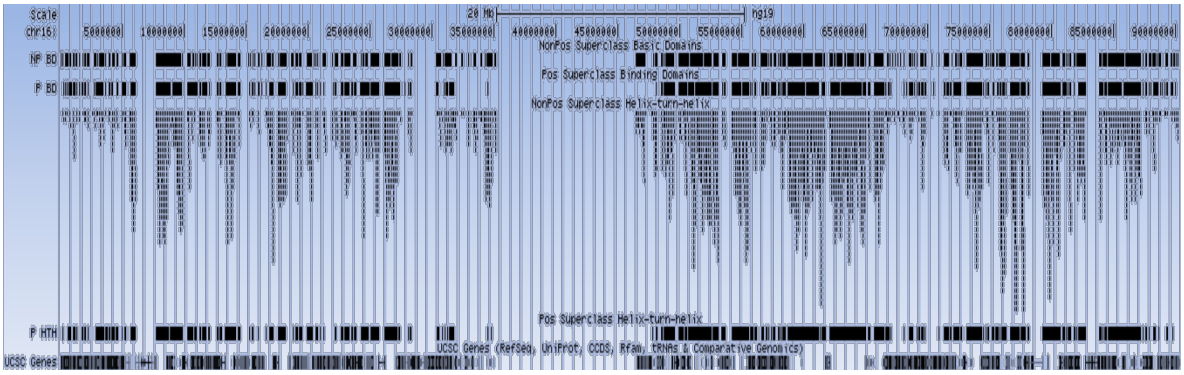
available via a web application that displays the results of a scan for positive selection in the human genome [23]. The HapMap includes only the most common markers, which are found in at least 5% of the population. Using this application, a list of the most non-positively selected non-protein coding SNPs was generated (7233 total), and the most positively selected non-protein coding SNPs (24630 total) on chromosome 16 as recorded by RefSNP accession IDs (rsnumbers) was analyzed in all four HapMap ethnic groups (CEU/YRI/CHB/JPT). The HapMap database, while available via a web application, is also available for download. While the web application was used to do spot checks, the data were downloaded and used directly for this work.

UCSC Genome Browser Mirror

The University of California at Santa Cruz (UCSC) has developed a Genome Browser website as a graphical interface for a large database of publicly available sequence and annotation data along with tools for comparing and aligning genome sequences between species and among different datasets [4]. The browser sequences and annotation data were downloaded from the UCSC website, as well as the Genome Browser source code. The source code was used to incrementally create a mirror site as per the Genome Browser instructions. For this project, the track SNP135 was used, which contains data about SNPs and small insertions and deletions from a provisional release of dbSNP build 135, available from the National Center for Biotechnology Information (NCBI). dbSNP provided a provisional release of genomic coordinates mapped from NCBI build 36 to the latest human assembly, GRCH37 (also hg19). Using this mirror site of the UCSC genome browser, each flanking sequence for the non-protein coding SNPs in chromosome 16 was mined to include only those regions around the SNPs that were found in each of the four HapMap populations (CEU/YRI/CHB/JPT). Chromosome 16 was chosen as a randomized sample as it appeared first on our initial search on the UCSC genome browser. Non-coding DNA sequences flanking the SNPs were gathered and aligned from the UCSC database by an automated process. The sequences were then imported, via automation, into the TRANSFAC web server that examines the DNA for identifiers and the associated factors that correlate with the DNA codes. The source code was modified to enable automatic search for flanking sequences. Before processing with TRANSFAC, those SNPs which were in genes regions were first excluded and those SNPs with flanking sequences that were in gene regions were also excluded.

An example of the UCSC graphical representation of gene or superclass of various data can be seen along the chromosome in Figure 2.

Figure 2. Superclass SNP locations plotted in the UCSC Genome Browser



TRANSFAC

TRANSFAC is a relational database available via the web as six flat files including various data concerning transcription factors, DNA-binding sites, and target genes (http://www.biobase-international.com/product/transcription-factor-binding-sites). The program PATCH uses the single site sequences which are stored in the SITE table and transcription factors are assigned a certain superclass and class, as well as other categorical determinations according to the DNA-binding domain [16]. Through collaboration with the Chinese University of Hong Kong, several fields from the tables of the 2009 version of TRANSFAC were collected and organized into two spreadsheets. Public versions of this database were available, but they did not contain the updated and extensive data that was desired in this study. The sequences mined from the UCSC Genome Browser mirror were analyzed by the PATCH program using automation, and the resultant data, which mostly concerned certain identifier sequences and binding factors, was returned for further analysis. Data from the PATCH program was compiled into two MYSQL tables, one each for the positively and non-positively selected SNP data. These tables contain data concerning identifiers found in the

flanking sequences of the SNPs. This includes the classification of the identifier, binding factors, functional properties, location, and expression information.

Refining of TRANSFAC data

TRANSFAC sequences which were present in the flanking sequences of both a non-positively selected SNP and a positively selected SNP were eliminated along with the two oppositely selected SNPs in order to ensure unambiguous selection association for analysis. A total of 6.4% of the SNPs found in TRANSFAC were eliminated in this process. The remaining sequences not shared by the NonPos and Pos cohorts are named "mono-selected."

Data Analysis

Proximity test script.

The distances of positively selected and non-positively selected SNPs from their nearest known gene were determined, collected, and analyzed for statistically significant differences between the two cohorts. The SNPs with flanking sequences overlapping with known genes were excluded. The SNP location on chromosome 16 was compared to genes from the UCSC Genome Browser list of "known genes" and the smallest absolute distance regardless of upstream or downstream orientation was collected. Statistical significance was determined by Student's t-test of the natural logs of the determined distances. The natural log was used because it showed a better distribution of data.

Categorization of TRANSFAC data.

The transcription factors predicted to bind to the various collected non-coding sequences were categorized in accordance with "superclass" distinctions provided by the TRANSFAC database. Differences in superclass prevalence between selection cohorts as well as proximity differences between the superclasses were analyzed. Statistically significant distribution differences between the superclasses for the two selection cohorts were determined using chi-squared tests. Student's t-tests were again used to determine statistical significance for the difference in proximity of superclasses to genes.

TRANSFAC Superclass Descriptions.

There are five broad superclasses given by TRANSFAC to include the different types of known transcription factors in the human proteome. These superclasses differentiate the transcription factors according to their DNA binding domain, the part of the protein which binds to specific DNA sequences flanking the SNPs collected. The names of these superclasses are Basic Domain, beta-Scaffold Factors with Minor Groove Contacts, Helix-Turn-Helix, Zinc-coordinating DNA binding domains, and Other [16] (Fig. 3).

Basic Domain (Basic). DNA binding domains rich in basic amino acids, which bind to DNA as a dimer [12, 24].

Beta-Scaffold Factors with Minor Groove Contacts (Beta). DNA binding domains which are globular and not elongated, possibly providing extensive contact with the binding sequence. Domains bind to the minor groove of DNA unlike other transcription factors which bind to the major groove [12, 24].

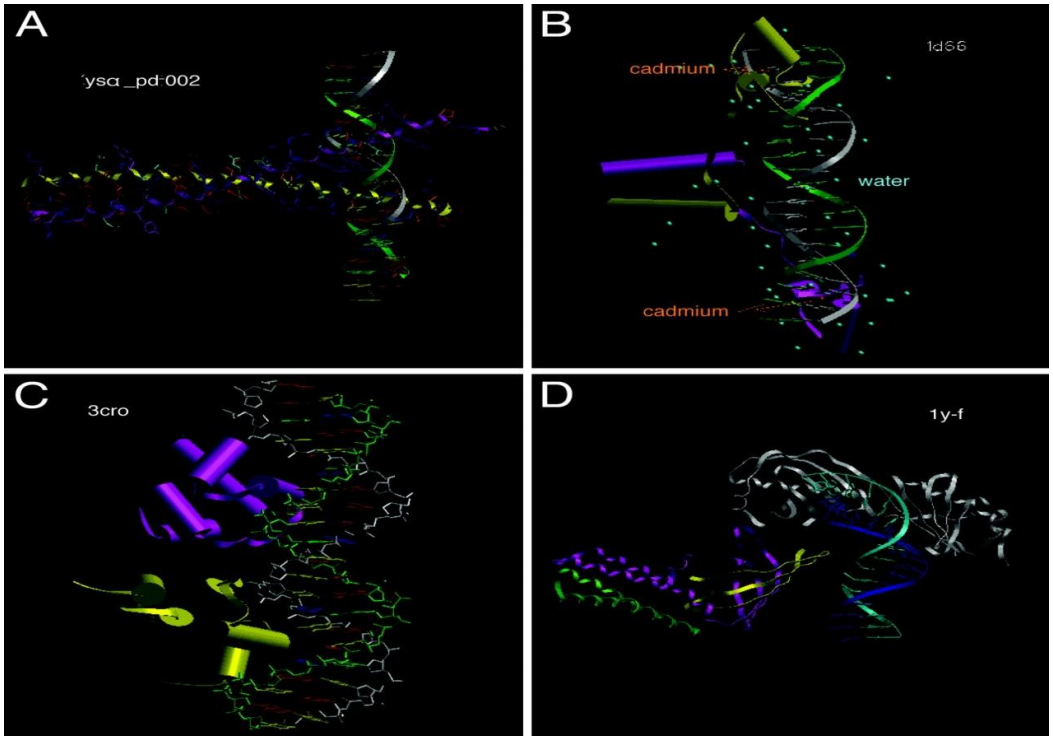
Helix-Turn-Helix (HTH). DNA binding domains composed of two short α-helices with a β turn in between. One α-helical segment is the recognition helix as it has the amino acids which interact with a specific binding sequence. This helix sits in a major groove with the other helix 34 Å away in another major groove [12, 24].

Zinc-coordinating (Zinc). DNA binding domains comprised of varying amounts of zinc finger domains. Zinc fingers are elongated loops which bind to the major groove of DNA wrapping partially around the DNA double helix and are held together by one Zn^{2+} ion at the base. The zinc ion is coordinated to either four cysteines, or two cysteines and two histidines [12, 24].

Other. This is a superclass for DNA binding domains not found in one of the four other superclasses. Our TRANSFAC results only produced one transcription factor characterized as "Other" by TRANSFAC, transcription factor HMGI(Y). "Other" will henceforth be called HMGI(Y). HMGI(Y) transcription factors have DNA binding domains with "A-T hooks" [2]. The binding sequences for all of our collected HMGI(Y) transcription factors were poly-adenine.

Unique Sequence Scan. Scans for sequences unique to either non-positive or positive selection were performed.

Figure 3. Examples of 4 superclasses: A) Leucine Zipper, a common Basic Domain transcription factor; B) Zinc fingers of the Zinc-coordinating superclass; C) Helix-Turn-Helix; D) beta-Scaffold Factors with Minor Groove Contacts [24].



Results

Data collection

Table 1. A summary of SNPs and transcription factor binding site sequences collected

	NonPos	Pos
Total SNPs	**7233**	**24630**
SNPs with TRANSFAC Data	**5373**	**18568**
% SNPs with TRANSFAC Data	**74.3%**	**75.4%**
SNPs in TF & Mono-Selected	**4606**	**17799**
% SNPs in TF & Mono-Selected	**63.7%**	**72.3%**
Unique Sequences (Shared - 760)	**32**	**226**
% Unique Sequence of Total Sequence (Total Sequence - 1018)	**3.1%**	**22.2%**

As seen in Table 1, about three quarters of all non-coding SNPs collected were found to have a potential role in transcription regulation. Moreover, 6.4% of SNPs found in TRANSFAC were located within a sequence which also contained an SNP of opposite selection. These SNPs and their corresponding sequences were eliminated during the refining of TRANSFAC data, before any steps of data analysis were conducted. The impact upon the percentage of mono-selected SNPs in TRANSFAC was greater for the non-positive cohort because both cohorts had roughly the same numbers of SNPs eliminated (767 NonPos and 769 Pos) but the non-positive cohort had less SNPs originally collected.

The Unique Sequence Scan found several transcription factor binding sequences which were either solely positively selected or non-positively selected. Over 7 times more sequences were found to be uniquely positively selected than sequences which were uniquely non-positively selected. The size of the cohorts does not fully explain this difference as the positive cohort is just less than 4 times larger than the non-positive, while containing 7 times more sequences.

Proximity Tests

Figure 4. Proximity Test 1: Average distances of SNPs from nearest known gene comparing positively and non-positively selected sequences in the 3 stages of collection: all SNPs found in non-coding region, non-coding SNPs with TRANSFAC sequences, and SNPs with TRANSFAC sequences found exclusively in positive or non-positive flanking regions or “mono-selected.”

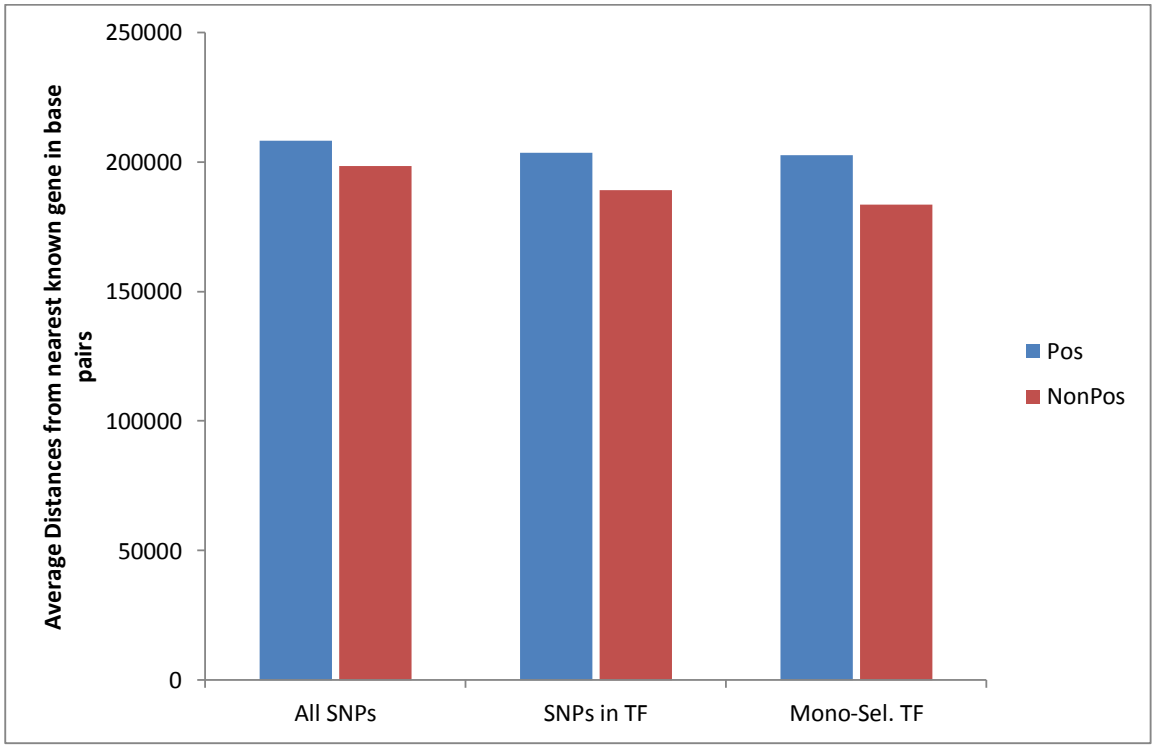
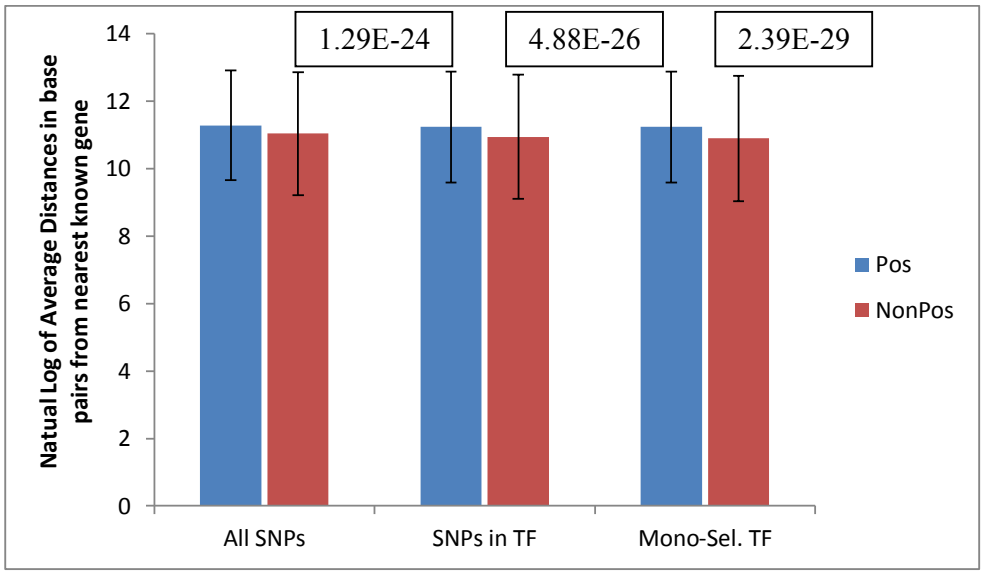


Figure 5. Proximity Test 1 statistical analysis: The natural logs of SNPs average distances with standard deviation and *P*-values above bars indicating significant differences between NonPos and Pos groups at each stage.



Proximity Test 1 (Figs. 4 and 5) found that there was a statistically significant difference in proximity to genes between positively selected SNPs and non-positively selected SNPs at every stage of our data collection. Figure 4 shows the actual distances with non-positively selected SNPs on average about 10,000 (All SNPs) to 20,000 (Mono-Selected in TRANSFAC) base pairs closer to genes than positively selected SNPs.

Figure 6. Proximity Test 2: Average distances of SNPs from nearest known gene comparing 2 stages of collection, All SNPs in non-coding regions and non-coding SNPs with TRANSFAC sequences, in positive, non-positive, and combined cohorts.

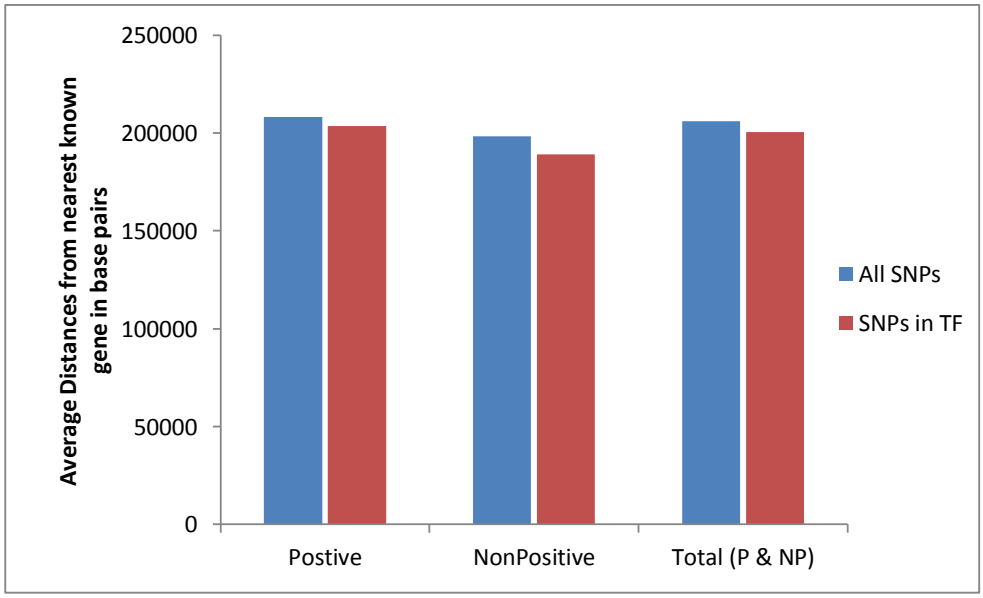
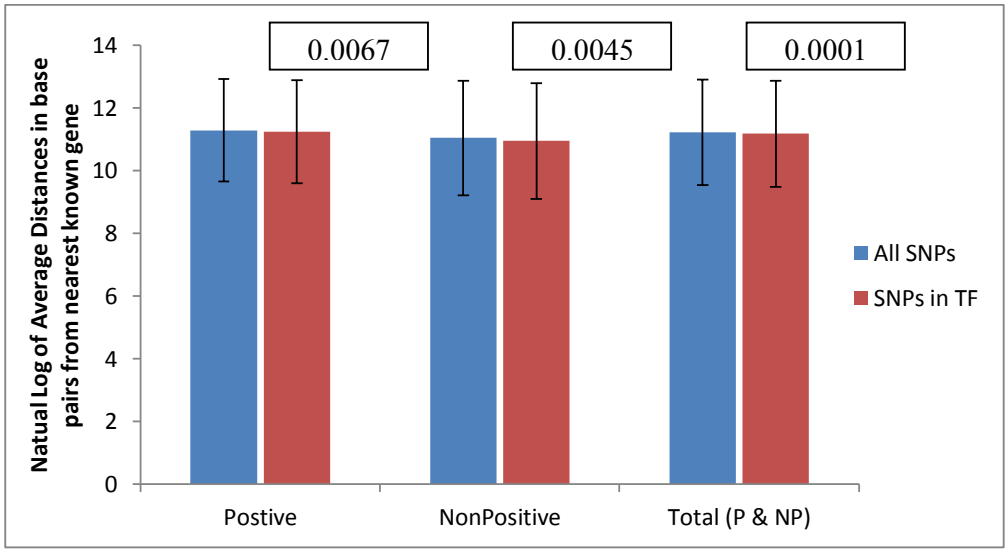


Figure 7. Proximity Test 2 statistical analysis: The natural logs of SNPs average distances with standard deviation and *P*-values above bars indicating significant differences between collection stage 'All SNPs' and collection stage 'SNPs with TRANSFAC sequences' in each selection cohort.



Proximity Test 2 (Figs. 6 and 7) found that there was a statistically significant difference in proximity to genes between all the SNPs found in the non-coding regions and non-coding SNPs with TRANSFAC sequences. This difference was found in the positive and non-positive cohorts as well as the two cohorts combined. Figure 6 shows that regardless of selection non-coding SNPs with sequences which have potential roles in transcription regulation are on average about 5,000 base pairs closer to genes than non-coding SNPs as a whole.

Categorization of TRANSFAC data

Table 2. Percentage of superclasses for all Pos and NonPos mono-selected TRANSFAC sequence data with chi-squared values contributed by each superclass to overall chi-squared test.

Superclass (Chi-Squared Value)	%Pos	%NonPos
Basic Domains (4.28E-5)	**13.20%**	**13.21%**
beta-Scaffold Factors with Minor Groove Contacts (18.15)	**23.36%**	**21.39%**
Helix-turn-helix (4.83)	**34.05%**	**32.82%**
HMGIY (69.67)	**2.34%**	**3.68%**
Zinc-coordinating DNA-binding domains (13.98)	**27.04%**	**28.94%**

A chi-squared test was performed to test if differences in relative amounts of different superclass sequences are related to selection. A chi-squared value of 106.6 of all of the superclasses combined (which is higher than the critical value) showed statistically significant differences between the relative amounts of different positively selected superclass sequences versus non-positively selected superclass sequences. To elucidate differences caused by selection between individual superclasses, the individual chi-squared values of each superclass were examined (Table 2). Of the five different superclasse**s,** beta-scaffold factors with minor groove contacts, HMGI(Y), and zinc-coordinating DNA-binding domains each have large chi-squared values, higher than the critical values for significance. The statistical differences between these three superclasses therefore explain most of the differences between the positively and non-positively selected groups overall. The chi-square values of basic domains seem to

indicate that their relative abundance has almost no relation to selection, and the effect of selection on helix-turn-helix transcription factors seems minimally present if at all, as they are smaller than the critical values.

Proximity of Superclasses to Genes

Figure 8. Positively selected, non-positively selected, and combined (Pos and NonPos combined) transcription factor superclass average distance from genes in number of base pairs.

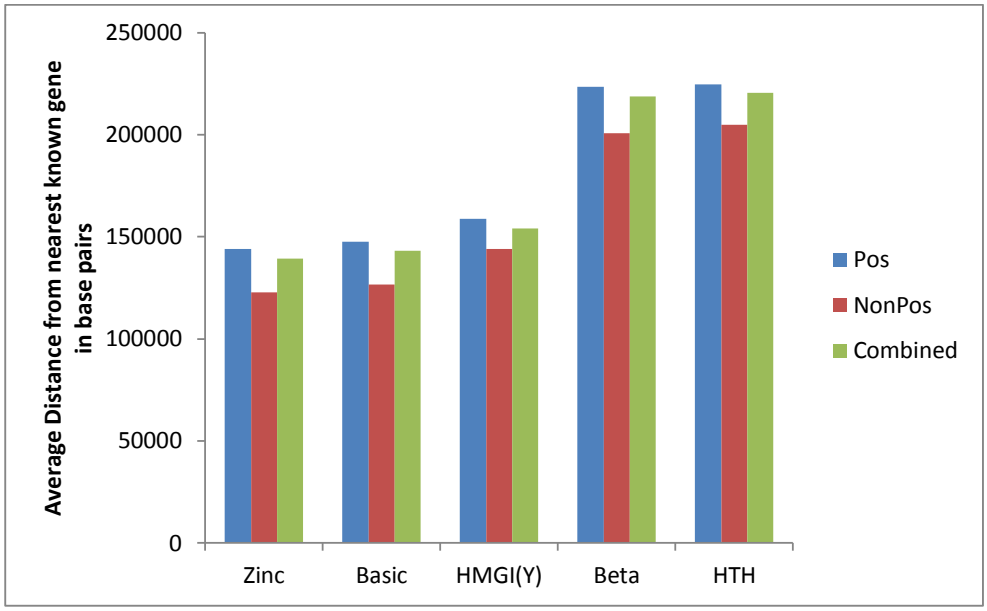


Figure 9. Positively selected, non-positively selected, and combined (Pos and NonPos combined) transcription factor superclass natural log of average distance from genes in number of base pairs for statistical analysis with standard deviation.

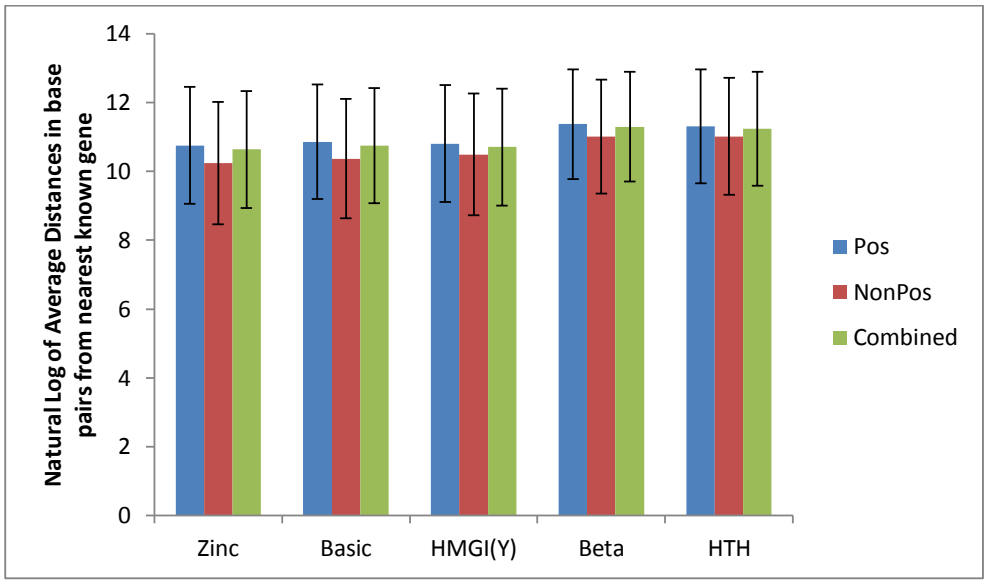


Table 3. Average distances and natural logs of average distances for non-positively selected transcription factor superclasses away from genes in number of base pairs (bp). *P*-values of statistical analysis comparing distances between each superclass. Bold numbers represent statistical significance between two superclasses' proximity to genes.

NonPos	Avg Dist (bp)	Avg Dist (ln)	Zinc (p-value)	Basic (p-value)	HMGIY (p-value)	Beta (p-value)	HTH (p-value)
Zinc	122680.72	10.24	---	**0.013**	**0.005**	**4.65E-62**	**5.29E-78**
Basic	126710.00	10.37	**0.013**	---	0.20	**6.78E-30**	**6.23E-35**
HMGIY	143983.88	10.49	**0.005**	0.20	---	**3.59E-8**	**8.94E-9**
Beta	200736.16	11.00	**4.65E-62**	**6.78E-30**	**3.59E-8**	---	0.80
HTH	204733.14	11.01	**5.29E-78**	**6.23E-35**	**8.94E-9**	0.80	---

Table 4. Average distances and natural logs of average distances for positively selected transcription factor superclasses away from genes in number of base pairs. *P*-values of statistical analysis comparing distances between each superclass. Bold numbers represent statistical significance between two superclasses' proximity to genes.

Pos	Avg Dist (bp)	Avg Dist (ln)	Zinc (p-value)	Basic (p-value)	HMGIY (p-value)	Beta (p-value)	HTH (p-value)
Zinc	144147.4	10.75	---	**5.39E-5**	0.36	**2.47E-185**	**8.45E-169**
Basic	147624.1	10.86	**5.39E-5**	---	0.31	**3.22E-88**	**9.05E-73**
HMGIY	158657.0	10.80	0.36	0.31	---	**2.31E-26**	**3.32E-21**
Beta	223468.0	11.37	**2.47E-185**	**3.22E-88**	**2.31E-26**	---	**0.0004**
HTH	224808.8	11.30	**8.45E-169**	**9.05E-73**	**3.32E-21**	**0.0004**	---

Table 5. Average distances and natural logs of average distances for combined positively and non-positively selected transcription factor superclasses away from genes in number of base pairs. *P*-values of statistical analysis comparing distances between each superclass. Bold numbers represent statistical significance between two superclasses' proximity to genes.

Both P & NP	Avg Dist (bp)	Avg Dist (ln)	Zinc (p-value)	Basic (p-value)	HMGIY (p-value)	Beta (p-value)	HTH (p-value)
Zinc	139190.48	10.63	---	**6.97E-7**	0.11	**1.62E-251**	**3.27E-247**
Basic	143040.90	10.75	**6.97E-07**	---	0.37	**1.6E-117**	**3.22E-105**
HMGIY	154175.02	10.71	0.11	0.37	---	**2.63E-36**	**6.30E-31**
Beta	218821.50	11.30	**1.62E-251**	**1.6E-117**	**2.63E-36**	---	**0.002**
HTH	220535.38	11.24	**3.27E-247**	**3.22E-105**	**6.30E-31**	**0.002**	---

The superclass proximity tests (Figs. 8 and 9) found statistically significant differences in the proximity to genes between several superclasses in positive (Table 4), non-positive (Table 3), and combined selection cohorts (Table 5). One of the major differences found in all selection cohorts was a large distance between average Zinc, Basic, and HMGI(Y) proximities compared to Beta and HTH proximities. Zinc, Basic, and HMGI(Y) binding sequences were on average about 60,000 to 80,000 base pairs closer to genes than the average Beta or HTH binding sequences. Zinc was found to be significantly closer to genes than Basic in all cohorts by about 4,000 base pairs, while the difference between Beta and HTH cannot be considered statistically significant across the selection cohorts. HMGI(Y) also cannot be considered significantly different than the two closest superclasses, Zinc and Basic.

Discussion

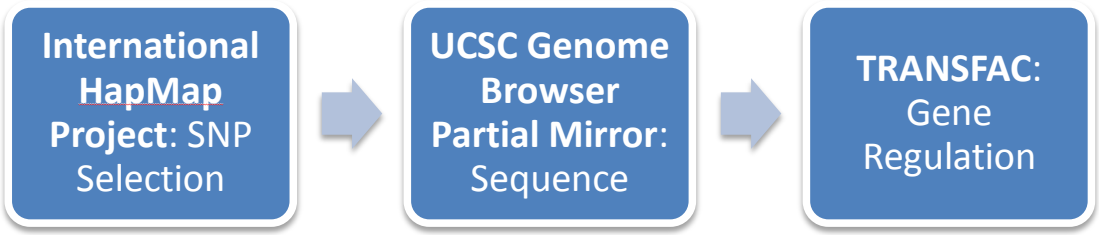
The implications of gene proximity in relation to selection are difficult to ascertain without further work, but the patterns elucidated in this paper provide evidential support for the theory that transcription factor binding sites exist in significant numbers in non-coding DNA as another component of human gene regulation, given that over 74% of SNPs in the non-coding regions of human chromosome 16 have flanking sequences that are homologous to transcription factor binding sites (Table 1). The regulatory role of non-coding DNA, however, seems to be a significantly dynamic part of gene regulation in humans. Since non-positively selected SNPs were on average about 10,000-20,000 base pairs closer to genes than positively selected SNPs (Figs. 4, 5; proximity test 1), and if this reveals a relationship to the functions that these SNPs play in gene regulation, then it indicates a declining role for that function. However, since SNPs linked to transcription regulation in the TRANSFAC database were on average 5,000 base pairs closer to genes than non-coding SNPs as a whole (Figs. 6 and 7), there may be conflicting selective forces at work.

The potential effects on the selection of transcription factors that bind to the sites predicted in this paper, however, may begin to clarify the nature of these forces. Interestingly, the chi-square test of the prevalence of superclass groups suggested that distribution between superclasses is affected by selection (Table 2). While these differences may be graphically minute, there is no standard for a biologically relevant amount of selection, and the selective differences in superclasses found here are statistically significant. Analysis of individual chi-squared values indicates that the HMGI(Y) transcription factor group is by far the most enriched by selection, with Zinc-coordinating DNA-binding domains and beta-Scaffold Factors with Minor Groove Contacts the next most affected groups. Basic Domain transcription factors and Helix-turn-Helix transcription factors show very minimal effects of selection. Whether these distributions reveal an actual trend in the role of different transcription factors in chromosome 16 alone could be investigated by using similar techniques for other chromosomes or even the genome as a whole to see if this pattern continues.

The differences in the proximity of superclass groups to genes were significant (Figs. 8, 9, Tables 3-5). Basic Domains, Zinc-Coordinating factors, and HMGI(Y) were on average about 60,000-80,000 base pairs closer to genes than Beta-Scaffold factors and Helix-turn-helix transcription factors. This phenomenon could be rationalized by the differences in the nature of the transcription factors involved in DNA binding and their properties for gene regulation. For example, Zinc-Coordinating factors may require closer proximity to genes than Helix-turn-helix factors because of their differential ability to navigate steric hindrance from local chromatin structures [12, 24]. Performing this analysis on other chromosomes in the human genome could test whether the average location of certain superclass binding sites is due to something inherent in the function of the transcription factors or due to DNA-chromatin structures specific to chromosome 16.

Other procedures have been developed for investigating possible functionality of non-coding SNPs, including using sequence conservation in order to determine the relevance of certain potential transcription factor binding sites. Interestingly, one study found that conserved segments in non-coding regions have an over-representation of transcription factor binding sites [13]. However, a problem similar to what they encountered in their data interpretation arose: in the literature, transcription factor binding sites may be conserved and play a role in regulation without actually binding transcription factors.

Figure 10. A novel pipeline for predicting functionality in non-coding DNA.



In this study, a novel and proof of concept pipeline was introduced and tested (Figure 10) for predicting functional transcription factor binding sites in non-coding DNA. The next step would be to verify that at least some of these sites are in fact functional via biochemical assays, e.g. chromatin immunoprecipitation assays [5], for a subset of those proteins which are known to bind these sites (and for which antibodies exist), followed by qPCR for our corresponding non-coding sequences. Two problems with this might be (a) determining what kind of samples to use for DNA extraction, since cis acting regulatory elements in eukaryotes often only bind at certain sites in one or a few different cell types within an organism [7], and (b) determining a subset of transcription factors to test. While overarching patterns across the entire structure of a human chromosome are provided herein, specific interests, particular applications and further studies will determine their significance.

The non-coding sequences comprise the majority of the human genome. These sequences have often been overlooked in comparison to their coding counterparts. Our analysis has shown over 74% of the non-positively or positively selected non-protein coding SNP flanking sequences in chromosome 16 may be functional as predicted from the TRANSFAC database. Regardless of selection status, these predicted sites of functionality have also been shown to be closer to genes than SNPs as a whole.

The discovery of the notable characteristics of the non-coding region presented here offers a possible route towards a better understanding of gene regulation and, along with the implications for biology's "central dogma" suggested herein, encourage further investigation into the non-coding regions.

Acknowledgement:

We thank the Wheaton College Aldeen grant and the Wheaton College Student Summer Researcher in Residence program for supporting Kyle Tretina, David Lee and John Snedeker and the Hong Kong Bioinformatics Center for sharing their computing resources. We thank Dr. John Vessey of the Psychology Department for his critical review on statistics. We also thank undergraduate Melissa Bragg and the students in the Bioinformatics class of the Fall semester of 2009 for collecting some of the initial data of the project. Preliminary data of this project were presented at Experimental Biology 2012. Kyle Tretina and John Snedeker contributed equal amounts of work and can both be considered first authors of this paper.

References

1. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, George RA, Lewis SE, Richards S, Ashburner M,Henderson SN, Sutton GG, Wortman JR, Yandell MD, Zhang Q, Chen LX, Brandon RC, Rogers YH, Blazej RG, Champe M, Pfeiffer BD, Wan KH, Doyle C, Baxter EG, Helt G, Nelson CR, Gabor GL, Abril JF, Agbayani A, An HJ, Andrews-Pfannkoch C, Baldwin D, Ballew RM, Basu A, Baxendale J, Bayraktaroglu L, Beasley EM, Beeson KY, Benos PV,Berman BP, Bhandari D, Bolshakov S, Borkova D, Botchan MR, Bouck J, Brokstein P, Brottier P, Burtis KC, Busam DA, Butler H, Cadieu E, Center A, Chandra I, Cherry JM, Cawley S, Dahlke C, Davenport LB, Davies P, de Pablos B, Delcher A, Deng Z, Mays AD, Dew I, Dietz SM, Dodson K, Doup LE, Downes M, Dugan-Rocha S, Dunkov BC, Dunn P,Durbin KJ, Evangelista CC, Ferraz C, Ferriera S, Fleischmann W, Fosler C, Gabrielian AE, Garg NS, Gelbart WM, Glasser K, Glodek A, Gong F, Gorrell JH, Gu Z, Guan P, Harris M, Harris NL, Harvey D, Heiman TJ, Hernandez JR, Houck J, Hostin D, Houston KA, Howland TJ, Wei MH, Ibegwam C, Jalali M, Kalush F, Karpen GH, Ke Z, Kennison JA,Ketchum KA, Kimmel BE, Kodira CD, Kraft C, Kravitz S, Kulp D, Lai Z, Lasko P, Lei Y, Levitsky AA, Li J, Li Z, Liang Y, Lin X, Liu X, Mattei B, McIntosh TC, McLeod MP, McPherson D, Merkulov G, Milshina NV, Mobarry C, Morris J, Moshrefi A, Mount SM, Moy M, Murphy B, Murphy L, Muzny DM, Nelson DL, Nelson DR, Nelson KA, Nixon K, Nusskern DR, Pacleb JM, Palazzolo M, Pittman GS, Pan S, Pollard J, Puri V, Reese MG, Reinert K, Remington K, Saunders RD, Scheeler F, Shen H, Shue BC, Sidén-Kiamos I, Simpson M, Skupski MP, Smith T, Spier E, Spradling AC, Stapleton M, Strong R, Sun E, Svirskas R, Tector C, Turner R, Venter E, Wang AH, Wang X, Wang ZY, Wassarman DA, Weinstock GM, Weissenbach J, Williams SM,

WoodageT, Worley KC, Wu D, Yang S, Yao QA, Ye J, Yeh RF, Zaveri JS, Zhan M, Zhang G, Zhao Q, Zheng L, Zheng XH, Zhong FN, Zhong W, Zhou X, Zhu S, Zhu X, Smith HO, Gibbs RA, Myers EW, Rubin GM, Venter JC. 2000. *The Genome Sequence of Drosophila Melanogaster. Nature* 287:2185-2195.

2. Banks G, Moore B, Reeves R. 1999. *The HMG-I(Y) A-T-hook peptide motif confers DNA binding specificity to a structured chimeric protein. The Journal of Biological Chemistry.* 274(23):16536-44

3. Brett D, Pospisil H, Valcarcel J, Reich J, Bork P. 2002. *Alternative splicing and genome complexity. Nature Genet.* 30:29-30.

4. Brooke R, Karolchik D, Kuhn RM, Hinrichs AS, Zweig AS, Fujita PA, Diekhans M, Smith KE, Rosenbloom KR, Raney BJ, Pohl A, Pheasant M, Meyer LR, Learned K, Hsu F, Hillman-Jackson J, Harte RA, Giardine B, Dreszer TR, Clawson H, Clawson GP, Haussler D, Kent WJ. 2009. *The UCSC Genome Browser Database: Update 2010. Nucleic Acids Research* 38:613-619.

5. Buck MJ, Live JD. 2004. *ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. Genomics* (83): 349-360.

6. Dermitzakis ET, Reymond A, Lyle R, Scamuffa N, Ucla C, Deutsch S, Stevenson BJ, Flegel V, Bucher P, Jongeneel CV, Antonarakis SE. 2002. *Numerous Potentially Functional but Non-genic Conserved Sequences on Human Chromosome 21. Nature* 420:578-82.

7. Griffiths AJF, Wessler S, Lewontin R, and Caroll S. 2012. *Introduction to Genetic Analysis.* 10th edition. New York: W. H. Freeman.

8. Hahn MW, Wray GA. 2002. *The G-value Paradox. Evolution & Development* 4(2): 73-75.

9. *International HapMap Consortium. "A second generation human haplotype map of over 3.1 million SNPs." Nature* 449.7164 (2007): 851-61. Print.

10. Kaisar J. 2008. *DNA Sequencing: A Plan to Capture Human Diversity in 1000 Genomes. Science* 319(5868):1336.

11. Knight JC. 2003. *Functional Implications of Genetic Variation in Non-coding DNA for Disease Susceptibility and Gene Regulation. Clinical Science* 104:493-501.

12. Latchman D. 2010. *Gene Control.* New York: Garland Science. p. 137-142.

13. Levy S, Hannenhalli S, Workman C. 2001. *Enrichment of regulatory signals in conserved non-coding genomic sequence. Bioinformatics* 17(10): 871-877.

14. Maniatis T, Tasic B. 2002. *Alternative pre-mRNA splicing and proteome expansion in metazoans. Nature* 41(8):236-243.

15. Mattick JS. 2001. *Non-coding RNAs: the Architects of Eukaryotic Complexity. European Molecular Biology Organization* (EMBO) 2(11): 986-91.

16. Matys V, Fricke E, Geffers R, Gobling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, Kloos DU, Land S, Lewicki-Potapov B, Michael H, Munch R, Reuter I, Rotert S, Saxel H, Scheer M, Thiele S, Wingender E. 2003. *TRANSFAC: Transcriptional Regulation, from Patterns to Profiles. Nucleic Acids Research* 31(1): 374-78.

17. Mullikin JC, Hunt SE, Cole CG, Mortimore BJ, Rice CM, Burton J, Matthews LH, Pavitt R, Plumb RW, Sims SK, Ainscough RM, Attwood J, Bailey JM, Barlow K, Bruskiewich RM, Butcher PN, Carter NP, Chen Y, Clee CM, Cogill PC, Davies J, Davies RM, Dawson E, Francis MD, Joy AA, Lamble RG, Langford CF, Macarthy J, Mall V, Moreland A, Overtonlartey EK, Ross MT, Smith LC, Steward CA, Sulston JE, Tinsley EJ, Turney KJ, Willey DL, Wilson GD, McMurray AA, Dunham I, Rogers J, Bentley DR. 2000. *An SNP Map of Human Chromosome 22. Nature* 407(6803):516-20.

18. Pennisi E. 2007. *Breakthrough of the Year: Human Genetic Variation. Science* 318:1842-843.

19. Sabeti, Pardis C., Patrick Varilly, Ben Fry, Jason Lohmueller, Elizabeth Hostetter, Chris Cotsapas, Xiaohui Xie, Elizabeth H. Byrne, Steven A. McCarroll, Rachelle Gaudet, Stephen F. Schaffner, and Eric S. Lander. *Genome-wide detection and characterization of positive selection in human populations. Nature* 449.7164 (2007): 913-18.

20. Shabalina SA, Spiridonov NA. 2004. *The Mammalian Transcriptome and the Function of Non-coding DNA Sequences. Genome Biol*. 5(4): 105.

21. Taft RJ, Pheasant M, Mattick JS. 2007. *The Relationship between Non-protein-coding DNA and Eukaryotic Complexity. BioEssays* 29: 288-89.

22. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Miklos GLG, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Francesco VD, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang ZY, Wang A, Wang X, Wang J, Wei MH, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu SC, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigó R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays AD, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek K, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X. 2001. *The Sequence of the Human Genome*. *Science* 291:1304-1351.

23. Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. *A map of recent positive selection in the human genome.* PLoS Biology 4(3):72.

24. Xing E, Karp R. 2004. *MotifPrototyper: A Bayesian profile model for motif families*. *PNAS* 101(29):10523-8.