Genome-Wide Analysis of Single Nucleotide Polymorphisms in and near Genes and Evolutionary Conserved DNA

Claudia Fried[†], Peter F. Stadler[†,‡], Peter Ahnert[♣]

[†]Bioinformatics Group, Department of Computer Science, University of Leipzig, Härtelstr. 16-18, D-04107 Leipzig, Germany.

[‡]Institute for Theoretical Chemistry and Structural Biology, University of Vienna, Währingerstraße 17, A-1090 Wien, Austria

[♣]Institute for Clinical Immunology and Transfusion Medicine, Center for Biotechnology and Biomedicine, University of Leipzig, Johannisallee 30, 04103 Leipzig, Germany

[*]Address for correspondence:

Dr. Peter Ahnert, Center for Biotechnology and Biomedicine, Institute for Clinical Immunology and Transfusion Medicine, University of Leipzig
Johannisallee 30, 04103 Leipzig, Germany
Tel.: +49-(0)341-9725484 Fax.: +49-(0)341-9725819
ahnert@uni-leipzig.de

Abstract.
Variation databases promise to enable the assessment of recent selection pressure on genomic sequence elements. Evidence of recent selection would suggest recent functional relevance of elements potentially important for understanding the organization of the human genome and resulting complex phenotypes and diseases. In contrast, comparative sequence analysis can be employed to focus on those regions that have have been under long-term stabilizing selection. Previously observed biases in variation databases appear to have been reduced to the point that their data can now be used to for investigate the relationships of long-term sequence preservation and recent selective pressures.
In a genome-wide study, we identified phylogenetic footprints (PFs) in the vicinity of human genes. In agreement with the distribution of known regulatory sites, the density of these PFs was highest within two thousand base pairs upstream and downstream of genes. Stabilizing selection acting on these PFs was most strongly indicated by significantly reduced single nucleotide polymorphism (SNP) density. Weak correlation between SNP densities of PFs and coding sequences suggests that gene regulation and function often evolve independently. Decreasing diversity in human genes with increasing time of conservation suggests that most old genes have not ceased to be functionally important today. On average, intergenic sequences are under the least selection pressure in the vicinity of genes and may serve as a preferred model for estimating neutral evolution.
Intriguingly, we observed increased SNP densities in coding sequences and introns as compared to non-coding sequences of genes conserved only among primates. These genes appear to be mainly involved in the regulation of gene expression and raise the question about mechanisms for adaptive evolution and their role in primate development.

NCBI MESH-Terms: Sequence Homology, Nucleic Acid, Evolution, Phylogeny, Single Nucleotide Polymorphism, Genome, Transcription Factors, Gene Expression Regulation
Other Keywords: Divergence, Diversity, Phylogenetic Footprint, Transcription Factor Binding Site

1. Introduction

Recently, there have been various new insights into the structure and function of the human genome. Prominently, regulatory untranslated RNAs have been in focus (Mattick, 2003). However, in connection with understanding metabolic and signaling pathways, there also has been considerable interest in regulatory regions of the genome, such as transcription factor binding sites, splicing regulators, and other motifs that exert their function at mRNA level. Databases of proven and putative transcription factor binding sites have been established (Heinemeyer *et al.*, 1998; Sandelin *et al.*, 2004) and various methods have been developed to predict new potential regulatory sites *in silico*, see e.g. (Blanchette and Tompa, 2002; Xie *et al.*, 2005). Whether or not predicted regulatory sites are of biological relevance usually remains elusive.

Binding sites for transcription factors are usually short and variable nucleotide sequences mostly upstream of genes. As a result of the potential degeneracy of the binding sequence these sites are hard to identify unambiguously, in particular if the transcription factors involved are not known *a priori* (Tautz, 2000; Ludwig *et al.*, 2000). For many purposes, it is of interest to establish whether or not predicted regulatory sites are biologically relevant. For instance the understanding of regulatory mechanisms and the interpretation or even prediction of gene expression patterns depends on this. Another area where knowledge about functional properties of genomic sequences is desirable is the search for genomic variants with association to diseases. While there is an advent of genome wide single nucleotide polymorphism (SNP) genotyping technologies and tagging strategies, these are still limited. Therefore, and in order to increase power of studies and to be able to detect actual causative SNPs, it is still desirable to choose SNPs with a high probability of functional relevance.

The functional relevance of predicted regulatory genomic sequences can in principle be established experimentally. If a specific sequence is under detailed investigation, this clearly would be the approach of choice. If, however, a large number of sequences is under investigation or a genome wide analysis is desired, experimental techniques so far are prohibitively resource intensive. A more accessible approach is the assessment of recent evidence of selection pressure on these sequences.

Here we present a genome wide study on the presence of recent selection pressure as reflected by the density and properties of SNPs in genes and their vicinity. We especially focus on phylogenetic footprints near genes conserved over different phylogenetic distances.

It has been noted for a long time that non-coding sequences can contain islands of strongly conserved segments. These so called phylogenetic footprints (Tagle *et al.*, 1988) are supposed to be regulatory elements that are conserved over a long period of time due to evolutionary pressure to keep functional regions constant. In a number of cases it has been shown that these phylogenetic footprints are indeed indicative of functional cis-regulatory elements (Duret and Bucher, 1997; Fickett and Wasserman, 2000). Phylogenetic Footprinting is a technique used to identify at least a subset of regulatory sites by comparing orthologous regions from multiple species thereby utilizing the feature of conservation.

The most common variations in the human genome SNPs. These variations are associated with divergence between species and diversity within species populations, susceptibility to diseases, and individual drug response. SNPs are the simplest form of mutations, which can cause changes in the protein structure and alter the function of the encoded protein, (Wang *et al.*, 2005). SNPs lying in a coding region can be distinguished into non-synonymous SNPs that can alter the encoded amino acid (nonsense and mis-sense mutations) and are likely to be deleterious and therefore may be eliminated by natural selection. The other form of SNPs occurring in coding regions are synonymous SNPs (silent mutations), that do not alter the amino acid sequence and therefore are subject to random drift. (Kimura, 1983). Therefore, SNPs are reduced in number and minor allele frequency in the coding region at non-synonymous sites compared to synonymous sites (Miller *et al.*, 2001; Sunyaev *et al.*, 2000; Zhao *et al.*, 2003; Freudenberg-Hua *et al.*, 2003; Hughes *et al.*, 2005), which reflects recent selection, which has been active in the past one to two million years.

Only 1-2 percent of the human genomic sequence encodes proteins ((The International Human Genome Sequencing Consortium, 2001; The Human Genome Sequencing Consortium, 2004)). The vast majority of SNPs are located in the non-coding regions and are believed not to be under natural selection (Barkur, 2002). However at least 5% of the human genome are under stabilizing selection (Waterston *et al.*, 2002). One has to distinguish between presumed non-functional regions and functional regions (such as non-coding RNAs and transcription factor binding sites). DNA sequences that are regulatorily active presumably cover a substantial part of these intergenic regions. Furthermore, it seems obvious, that not all functional changes are based on variations of the protein coding sequence. It is known that alteration, deletion, or destruction of regulatory sites can disrupt the use of its target genes or lead to an altered (higher or lower) gene expression (Knight *et al.*, 2003). Since the nucleotide composition of regulatory elements is important for binding of transcription factors, it is likely that polymorphisms located in regulatory sites of a gene can act directly on the expression level of the regulated gene. While the sequence of transcription factor binding sites is somewhat degenerate (Dermitzakis and Clark, 2002), SNPs may still affect the binding affinity for a particular transcription factor, lead to recognition by a completely different transcription factor, or destroy the binding site altogether.

This would implicate that SNPs in regulatory regions can be function-modifying and selection pressure should act on these regions. Consistent with this hypothesis, Majewski and Ott (2002) showed that, within coding regions, the ratio of non-synonymous to synonymous changes decreases at sites with higher density of regulatory elements (e.g. splice site regulators). In our study we focus on the analysis of putative regulatory polymorphisms, coding polymorphisms and polymorphisms in sequences located in the vicinity of genes but which are not presumed to have a regulatory function.

For the identification of potentially regulatory sites, we here use a relatively conservative definition based on phylogenetic footprinting, i.e., evolutionary conservation of noncoding DNA sequence elements. Phylogenetic footprints as detected by `tracker` (Prohaska *et al.*, 2004) are presumed to exist due to the influence of selection. In

this study, we consider evolutionarily conserved non-protein-coding DNA that is conserved at least across the major mammalian clades an sometimes even originate before the last common ancestor of tetrapods and actinopterygian fishes at least 450 million years ago (Kumar and Hedges, 1998).

Here we are asking whether these regions, which have been under long term selection, have still been of functional relevance in the past one to two million years, a time frame relevant to the recent development of humans. If these footprints are still functionally relevant, one would expect that mutations should be selected against in coding and regulatory sequences and that the rate of occurrence of SNPs is lowest in these DNA regions, while SNPs are more common in non-coding, non-functional DNA since these regions are not subject to selective pressure.

Many studies have been concerned with nucleotide diversity in coding regions (Fay *et al.*, 2001; Majewski and Ott, 2002; Miller *et al.*, 2001; Sunyaev *et al.*, 2000). This may be due to the fact that the influence of non-synonymous SNPs in relationship to synonymous SNPs can be easily understood. Consequences of the polymorphisms for protein structure and function can be reasonably well predicted as the structure and functions of genes are in many cases well known. However, regulatory DNA elements and other regions not coding for proteins (in particular non-coding RNAs) may be equally important for understanding the functioning of the human genome. Therefore, we started looking at such non-coding regions with potential functions, in particular phylogenetic footprints determined by `tracker`.

2. Materials and Methods

2.1. Sequence Data and SNP Data. We use 7111 human *(Homo sapiens)* genes with homologous genes ($\geq$ 40% identity to the human coding regions according to ENSEMBL annotation) in pufferfish *(Fugu rubripes)* and zebrafish *(Danio rerio)* and mouse *(Mus musculus)* throughout this study. In all cases the DNA sequence extending up to 9999nt upstream and downstream of the gene was retrieved (to the extent available) from the EBI databases using `ENSMART`[1]. Database versions: Human NCBI 35, dbSNP 123, HGVbase 15 TSC (Ensembl v30), Fugu FUGU 2 (Ensembl v30), mouse NCBI 33 (Ensembl v30), zebrafish ZFISH 4 (Ensembl v30).

For these genes, SNP information was retrieved using `ENSMART` from `dbSNP`[2]. There are several types of sequence variants contained in dbSNP. Variation in the repeat number of a motif (microsatellite markers), indels (small deletions and insertions), polymorphic insertion elements such as retrotransposons, invariant regions of sequence and single nucleotide polymorphisms (Sherry *et al.*, 2001).The quality of SNP data depends on the SNP detection method and on independent verification. We distinguish validated from non-validated SNPs in our analysis according to the annotation in `ENSMART`.

The sequences studied here cover 15.6 % of the Human genome and contain about 14.7 % of the known SNPs. The sequences were selected to be centered around an

[1]URL: `http://www.ensembl.org/EnsMart/`

[2]URL: `http://www.ncbi.nlm.nih.gov/SNP/index.html`

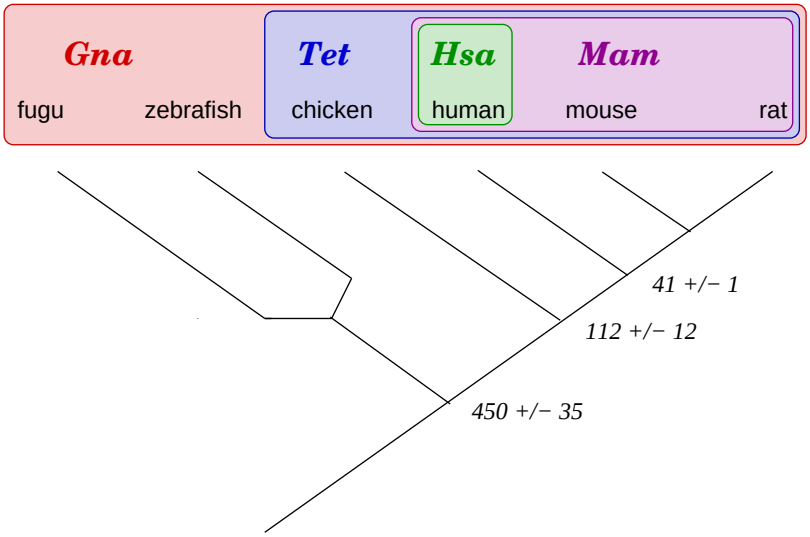

Figure 1. We used four sets of human genes that are differentially conserved. In the tree above these different sets of genes are marked by colored boxes. For example, human genes with homologs in mouse and rat but not in chicken, fugu, or zebrafish are labeled with a green box. Divergence times are taken from the work of Kumar and Hedges (1998)

Table 1. Summary statistics of analyzed data.

	SNPs			investigated
Region	all	validated	non-val.	sequence
CDS	26545	12112	14433	12235070
IN	1046727	585744	460983	389461359
UTR	239157	130610	108547	55316108
PF	21312	11289	10022	7277934
NC	144889	75033	69856	45519710

annotated gene and contained the immediate flanking regions. Therefore our analysis does not represent genome sequences far from known genes. The basic descriptive statistics of sequences and SNPs used here are summarized in table 1. We distinguish coding sequences (CDS), introns (IN), total UTR (defined as all parts of an annotated gene that are not explicitly designated as CDS or intron), phylogenetic footprints (PF), and non-coding non-regulatory sequences (NC).

In addition, we used SNP information for the following specific sets of genes, see fig 1: ***Hum*** 3461 (3457 genes for which SNPs where available) human genes that have no known homologs in either mouse, rat, chicken, fugu, or zebrafish. ***Mam*** 1323 (1323) human genes with homologs in mouse and rat but not in chicken, fugu, or zebrafish; ***Tet*** 797 (797) human genes with homologs in mouse, rat, and chicken, but not in the teleosts; ***Gna*** 6585 (3007) human genes with homologs in mouse, rat, chicken, zebrafish, and fugu.

2.2. Phylogenetic Footprinting. Conserved non-coding sequences (phylogenetic footprints) were detected using the `tracker` software (Prohaska *et al.*, 2004). This

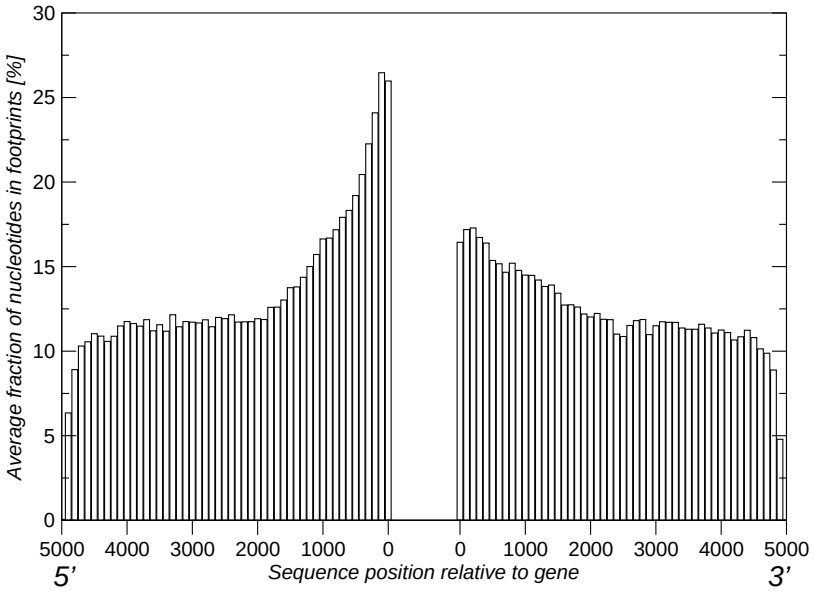

Figure 2. Distribution of sequence conservation in the vicinity of annotated genes. The footprint density is computed as the average number of nucleotides located in phylogenetic footprints in a each 100 windows. Only a few sequences were available further than 5000 nucleotides away from genes. Therefore a calculation of the mean number of nucleotides in phylogenetic footprints was not meaningful and intervals further away than 5000 bp were ignored. The gene itself gap from 0 to 0 which does not reflect the actual length of the gene.

program is based on `BLAST` (Altschul *et al.*, 1990) for the initial search of all pairs of input sequences. Comparisons are restricted to homologous intergenic regions. The resulting list of pairwise sequence alignments is then assembled into groups of partially overlapping regions that are subsequently passed through several filtering steps. Typically `tracker` detects clusters of such footprints which are termed *cliques*. These cliques are therefore composed of true binding sites interspersed by short stretches of DNA that are less conserved and presumably non-functional.. For the purpose of the present study we used `tracker` to detect phylogenetic footprints in human genes with homologs in mouse, zebrafish and pufferfish.

3. Results

3.1. Phylogenetic footprints are clustered near genes, especially upstream. Phylogenetic footprints represent one way to identify potential regulatory non-coding sequences. In Fig. 2 we show the distribution of phylogenetic footprints in the vicinity of the genes analyzed in this study. We observe a rather broad distribution with significantly enhanced footprint frequency within an interval of at least 1000nt away from the coding sequence and pronounced bias towards 5' flanking regions.

3.2. SNPs are Underrepresented in Coding Regions, Introns, and Phylogenetic Footprints. We used Fisher's exact test (Fisher, 1935, 1962) to assess the statistical significance of the differences in SNP frequency in exons, introns, UTR, and phylogenetic footprints compared to the remaining non-coding non-conserved DNA.

Table 2. Distribution of SNPs in the set of 7111 genes for which phylogenetic footprints where detected. The density ρ of SNPs is listed in units of SNPs per 10kb for genes (ρ_{CDS}), introns (ρ_{IN}), phylogenetic footprints (ρ_{PF}), and non-coding non-conserved DNA (ρ_{SNC}). Significant over- or under-representation of SNPs relative to non-conserved non-coding regions is indicated by ($\uparrow$) or ($\downarrow$), respectively, according to Fisher's exact test.

7111 genes	SNPs		
	all	validated	non-val.
ρ_{CDS}	21.7	9.90	11.80
FT(SNP $\uparrow$)	1	1	1
FT(SNP $\downarrow$)	0 $\downarrow$	0 $\downarrow$	0 $\downarrow$
ρ_{SIN}	26.88	15.04	11.84
FT(SNP $\uparrow$)	1	1	1
FT(SNP $\downarrow$)	0 $\downarrow$	0 $\downarrow$	0 $\downarrow$
ρ_{SPF}	29.28	15.51	13.77
FT(SNP $\uparrow$)	1	1	1
FT(SNP $\downarrow$)	0 $\downarrow$	0 $\downarrow$	0 $\downarrow$
ρ_{SNC}	31.83	16.48	15.35

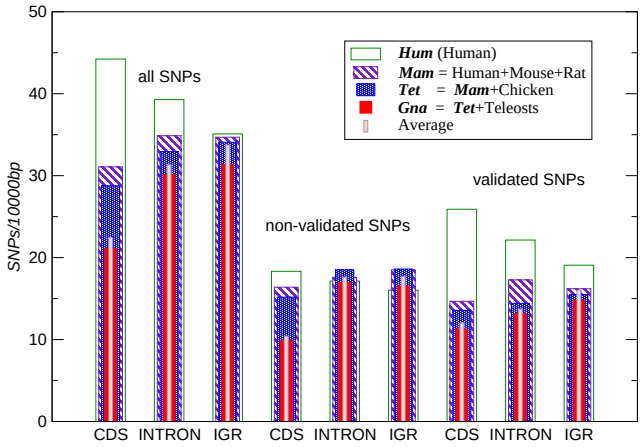

Figure 3. Comparison of SNP densities within coding regions, gene region but non-coding (denoted as intron in the figure. This region include UTRs, and introns according to `ENSMART` annotation) and in 10kb flanking regions both up- and down-stream. Human genes that have no known homologs in non-hominid species have an increased density of SNPs compared to evolutionarily older genes. This difference is detectable for both the genes and for their flanking regions.

In tab 2 the densities of SNPs and the results of the statistical analysis are summarized. Data for UTRs remained inconclusive, however, due to poor annotation data of UTR locations (not shown).

The distribution of SNPs in `tracker`-predicted phylogenetic footprints and exonic region does significantly differ from distribution in non-coding non-conserved DNA on average. SNPs are significantly underrepresented, i.e., selected against, in both coding regions and in conserved non-coding regions in the vicinity of genes.

3.3. SNPs are Underrepresented in Old Genes. Fig. 3 shows the density of SNPs in genes that are different evolutionary age or alternatively, that are under unequal selection pressure. For this purpose we distinguish between genes showing low inter-species divergence and conservation over a long evolutionary time (old genes) and human genes where no homologs can be found in non-hominid organisms (new genes). We have distinguish different groups of old genes: ***Gna*** has homologs in mouse, rat, chicken, fugu and zebrafish), ***Tet*** consists of genes conserved in mammals and aves (homologs in mouse, rat and chicken), ***Mam*** refers to mammal-specific genes (homologs in mouse and rat). New genes, designated ***Hum***, by definition have no known homologs in any of these species. In the CDS of newer genes we clearly see a higher SNP density than in evolutionarily older genes. The same trend is also observed for introns and intergenic regions, albeit it is much less pronounced in these regions.

Using the four gene sets with different evolutionary conservation, we observe that a reduced amount of interspecies divergence is correlated with a reduced level of diversity for CDS and introns. The effect is most pronounced in the CDS. In contrast it is barely detectable in the non-coding flanking sequences.

It is noteworthy that for new genes (***Hum***), the density of SNPs in INTRON and especially in CDS is higher than in IGR. Since this may reflect a data bias, we analyzed the annotation of the Hum genes. They are mainly involved in the regulation of protein expression, as potential transcription factors, zinc finger proteins, and protein components of ribosomal subunits.

As one would expect, genes without known homologs are most frequently annotated as "Putative/PREDICTED", while the fraction of "Putative/PREDICTED" genes is lowest in the group that has been conserved throughout vertebrate evolution. Since genes annotated "Putative/PREDICTED" might in fact be just genomic background, we repeated the entire analysis without this class of genes. In this more conservative dataset, the same trends were observed: there is an elevated SNP density in both CDS and introns of genes without known homologs.

The higher CpG content of the synonymous sites in coding regions of new genes which were recently formed as pseudogenes from conserved genes is known to be responsible for a higher mutation rate (and therefor SNP density) in CDS compared to intron and IGR (Subramanian and Kumar, 2003). In our analyses, the CpG content in exons is highest in the most conserved group (5.5%) and lower in the less conserved groups (4,9%, 5.06%, 4.74%), without an exception in ***Hum***. CpG content at the different codon positions is: ***Gna*** (1,2: 33,16% 2,3: 32.86% 3,1: 33.97%); ***Hum*** (1,2: 32.92% 2,3: 31.83% 3,1: 35.26%).

3.4. SNPs with low heterozygosity are enriched in CDS. Heterozygosity of SNPs also potentially conveys information of selective pressures. Stabilizing selection is expected to reduce minor allele frequencies (MAF) and, hence, heterozygosity.

We compared the percentage of SNPs with different levels of MAF among the different categories of genomic locations, see tab 3. To this end, we divided SNPs into frequency classes in the same manner as described by Cargill *et al.* (1999); Fay *et al.* (2001),

Table 3. Percentage of SNPs in three minor allele frequency classes (MAF <5%, 5-15% and 15-50%, respectively), for the different functional categories. Only SNPs for which heterozygosity data was available in dbSNP 123 are taken into account.

region	< 5%	5-15%	15-50%
CDS	26.34	18.16	55.50
Intron	14.03	16.52	69.45
PFs	15.64	17.00	67.36
NCR	14.13	15.84	70.03

Table 4. Relative frequency of SNP variation types (in %). STR: short tandem repeat; DIP: deletion/insertion polymorphism; Un: undefined SNP type A T G C seen in this position (maybe data from plus and minus strand analyzed); Transversion; Transition; T&T: Transversion and Transition.

Variation	CDS	Intron	PF	NCR	UTR
	validated SNPs				
Transversions	22.18	30.77	31.75	32.09	30.89
Transitions	77.14	68.42	67.24	66.89	68.30
T&T	0.19	0.12	0.20	0.17	0.11
DIP	0.10	0.60	0.66	0.76	0.61
STR	0.00	0.01	0.01	0.01	0.01
Un	0.39	0.09	0.14	0.09	0.08

termed "low SNPs", "moderate SNPs", and "common SNPs" (MAF <5%, 5-15%, and 15-50%, respectively). The frequency of low SNPs was higher in CDS than in the other sequences types. Conversely, the fraction of common SNPs was lower in coding regions than elsewhere. In introns and phylogenetic footprints we observe the same trend, although the effect is much smaller.

3.5. Transition mutations are enriched in coding regions. Different types of mutations have different potential to disrupt functional sequence elements. As one would expect, deletion/insertion polymorphisms are very rare in coding sequences. In our analysis, their frequency was reduced by a factor of 6 in this category. Freudenberg-Hua *et al.* (2003) reported that "radical" changes of a coding sequence are more likely to result from transversion while transitions account for conservative and synonymous changes. Even at the level of nucleic acids one might argue that transition are smaller changes than transversions because the substitutes belong to the same chemical group (purine to purine and pyrimidine to pyrimidine), whereas transversion lead from one type of base to the other (Guo and Jamison, 2005). One would therefore expect a reduced ratio of transversions to transitions with increased levels of stabilizing selection. This was indeed observed for coding regions, Tab. 4.

By the same argument, transversion should be biased towards the low-frequency category of SNPs. Consistent with this expectation, we found that 28.06% of transversions located in exons are low SNPs, while only 25.7% of transitions fall into the same category. This difference is nearly significant with a p-value of 0.054 according to Fisher's exact test.

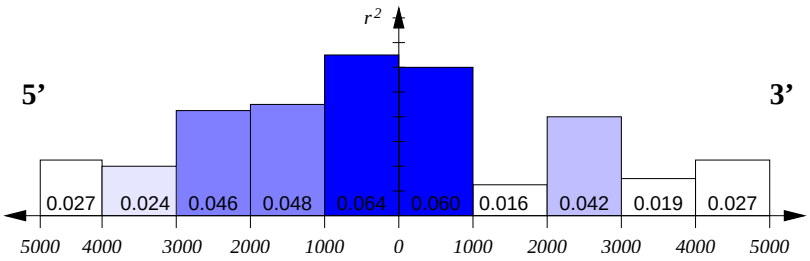

Figure 4. Correlation (Kendall's τ) between SNP density in CDS and surrounding phylogenetic footprints in 1kb-windows. Shading indicates significance, for dark to white: $p < 0.0001$, $p < 0.001$, 0.01, 0.05, and not significant.

For Intron, PF, and UTR, the rate of transversions is also lower than for NCR. However, these differences do not reach significance. Deletion/Insertion polymorphisms are also most abundant in NCR, but again this is not significant.

3.6. Weak Correlation between SNP densities of CDS and Footprints. (Wagner *et al.*, 2005) report a correlation between evolution rates of coding sequences and conserved non-coding sequences of posterior *HoxA* genes following the teleost specific genome duplication. In principle, the correlation of SNP density with selection pressure should allow us to detect a correlation between selection pressures on functional non-coding regions and coding sequence provided such a correlation is a strong and generic feature in molecular evolution. In order to test this hypothesis, we correlated SNP densities in phylogenetic footprints located in windows of 1000nt up- and downstream of the translation start sites of genes with the SNP densities in the coding sequence (see Fig. 4 and supplemental data). Indeed we see an correlation between SNP densities in phylogenetic footprints and SNP densities in the coding sequence although it is a very weak one. In the close vicinity of genes we find a higher correlation of the SNP densities than in regions further away. Using the entire flanking region, we obtain the following correlation coefficients: Pearson's $r^2 = 0.053$ (significant at $p = 4 \times 10^{-6}$) and Kendall's $\tau = 0.083$ (significant at $= 2 \times 10^{-16}$).

4. Discussion

Ongoing efforts on high-throughput genotyping methods have enabled and spawned many studies attempting to identify the role of genetic variation in the etiology and pathogenesis of many diseases. One approach to identify genes associated with a disease is to select candidate genes according to their known or assumed role in biological processes. The association of variants of these genes with the disease are then tested by genotyping a number of SNPs distributed along their genomic DNA. These SNPs either modify the function of the gene directly or are in linkage with function-modifying polymorphisms. Clearly, SNPs located in the coding sequence or a functional element (determined for instance as a phylogenetic footprint) of the gene are most likely to be function-modifying and are therefore the best candidates for further experimental study, in combination with tagging SNPs from HapMap or other efforts for capturing general variation.

An important issue is the question whether or not predicted functional non-coding regions are indeed functionally relevant. Short of direct experimental verification, evidence of selection is a strong indicator of functional relevance. Phylogenetic footprints by their nature are evidence of stabilizing selection over long phylogenetic distances, suggesting functional relevance. In the context of research into human diseases, the question arises, whether or not these phylogenetic footprints have been under stabilizing selection more recently, during a time scale relevant to human evolution. The distribution of SNPs in exons, regulatory elements, and non-functional background potentially provides direct information on selection pressures acting on various components of a gene within the past two million years or so.

We determined phylogenetic footprints in the vicinity of human genes conserved in pufferfish, zebrafish, and mouse. Our analysis of SNP distribution and features for these genes and related sequences showed mainly that the density of SNPs is lower in coding region, introns, and phylogenetic footprints compared to non-coding non-conserved DNA. This suggests that, on average, the determined phylogenetic footprints have been under recent stabilizing selection. Due to our overall analysis of sequences within or near known genes, our analysis do not represent genome sequences far from known genes and also do not allow the assessment of the specific situation in particular loci.

4.1. Evidence of functional relevance — phylogenetic footprints are clustered near genes, especially upstream. A first important aspect in assessing the functional relevance of phylogenetic footprints as detected by `tracker` (Prohaska *et al.*, 2004) is whether or not their overall distribution reflects that of known regulatory sites. There has been various evidence that regulatory sites in the genome cluster in close vicinity to genes, mostly upstream. Some elements, like promoters, are often very close to the transcription start site while other elements, like transcription factor binding sites, enhancers, and silencers, may be close or further away. The phylogenetic footprints detected by tracker are presumed to occur due to the influence of long-term stabilizing selection based on actual regulatory function. The more frequent occurrence of phylogenetic footprints close to genes (Fig. 2) therefore supports the notion that at least part of the detected footprints are functionally relevant as transcription regulators. The difference of footprint densities between upstream and downstream flanking regions of genes reflects known differences in the arrangement of regulatory sites near genes. The relatively high background level of phylogenetic footprints more than a thousand bases upstream and downstream of genes resembles that of transcription factor binding sites predicted by recognition matrices. The functional relevance of both remains largely elusive.

4.2. Evidence of functional relevance — SNPs are Underrepresented in Coding Regions, Introns, and Phylogenetic Foot-prints. In the analysis of 7111 genes with homologs in mouse, pufferfish, and zebrafish and for which SNPs were available SNP density was significantly decreased in coding sequences, introns and phylogenetic footprints in comparison to extragenic sequences in the vicinity of genes. The strong decrease of SNP density in coding sequences is clear evidence of more recent selection pressure. This is not surprising, assuming that most genes which have

been functionally important since our ancestors parted way with those of pufferfish and zebrafish are still likely to be functionally important. The average SNP density in introns and phylogenetic footprints is much closer to that of extragenic regions. Still, it is significantly reduced, indicating the influence of recent selection pressure. The smaller effect of stabilizing selection on introns and phylogenetic footprints as opposed to coding sequences is not unexpected. In introns, regulatory sites, for instance for splicing, only make up part of the sequence. The regulatory sites themselves show a degree of degeneracy in regard to their sequence requirements. Therefore, the same strength of functional preservation will leave less of a detectable footprint in form of reduced SNP density. Along the same line of arguments, phylogenetic footprints are mostly clusters of binding sites with degenerate recognition sequences and consist of highly conserved parts as well as connecting sequences which are not as highly conserved resulting in possible dilution of visible effects of selection. We therefore conclude that the observed evidence of selection corroborates the functional relevance of phylogenetic footprints.

4.3. Phylogenetic distance of conservation — a measure of functional relevance for basic processes of live? (SNPs are Underrepresented in Old Genes). Figure 3 shows that SNP densities in the sequence categories CDS, INTRON, and IGR depend on the phylogenetic age of genes. When either all SNPs represented in the database are used or only he SNPs annotated as validated, the SNP densities for the oldest genes ***Gna*** are smaller than in the respective sequence categories of any other group. This effect continues for ***Tet*** and ***Mam*** and is most pronounced in CDS but can also be observed in INTRON and even in IGR. Since the phylogenetic age of the genes analyzed is much larger than the life time of SNPs, this suggests that very old genes which have been under stabilizing selection for a very long time may still be functionally more essential than younger genes and are under stronger selection pressure. On average, even their regulatory elements appear to be under stronger recent selection, as indicated by the reduced SNP density in INTRON and IGR.

4.4. The surprising phenomenon of SNP density being highest in CDS of new genes. A surprising observation is evident in Figure 3. While the increase of SNP densities from phylogenetically older genes to newer genes is at least plausible, the observation of higher SNP density in CDS of Hum genes as opposed to IGR of Hum genes is quite surprising. One would normally expect that in the absence of selection the SNP density in CDS should maximally reach that of IGR. This of course raises questions about the nature of this phenomenon. Is it due to some bias in the available data or is it a biologically relevant observation? Is the SNP density in all analyzed IGR lower due to some selection effect which has been released for CDS in ***Hum*** or is there some mechanism actively increasing the variation rate in the CDS of genes specific to humans? Analysis of the gene ontology annotations of the ***Hum*** genes and manual inspection showed that they are mainly involved in the regulation of protein expression. Since these genes are primate specific, they would be relatively new and an increased SNP density is at least plausible.

One possible explanation of the phenomenon could be an over-representation of putative genes or pseudogenes in ***Hum***. These may actually represent genomic background far from genes and may be under even less selection than IGR near actual genes. However, this would not explain the differences between the sequence classes within ***Hum***. Excluding putative genes from analysis does not change the observed phenomenon, which therefore could indeed be of biological significance.

If a substantial number of genes in ***Hum*** were to have recently formed from pseudogenes which in turn rose from genes with a high CpG content preserved by strong selection, than this could explain the high SNP density in CDS of ***Hum*** genes (Subramanian and Kumar, 2003). When selection pressure is released from sequence elements, the higher mutation rate of CpG sites becomes visible as increased SNP density as long as the selection pressure has not been released too long ago so that SNPs would be fixed or eliminated by drift. In our analysis, the CpG content in exons is highest in the most conserved group and lower in the less conserved groups without an exception in ***Hum***. This would suggest that the mechanism described above has no been active in the past two million years. It could be argued, that this process has just finished for these genes and therefore the CpG content is not especially high in CDS of ***Hum*** genes. Another aspect, however, makes the CpG hypothesis to appear less likely: It is difficult to see how enough genes to influence the statistics would have risen from well-conserved genes without showing homology to other species.

On he more speculative side, active mechanisms could be at work which increase the variation density in CDS (and to a lesser degree in INTRON) of genes in the Hum category. Are there so far unknown regions of hypermutation, similar to that of immune globulin genes, the products of which are then sent throughout the genome by a mechanism similar to retrotransposons? Similarly speculative is the idea that a large proportion of ***Hum*** genes are located in regions which are difficult to sequence and therefore contain a higher level of sequencing mistakes. But why would sequencing effects be specific for exons?

4.5. Other indicators of selection pressure — heterozygosity, transition mutation, and deletion/insertion frequencies. Natural selection may not only influence the occurrence of SNPs but also their heterozygosity. Stabilizing selection is expected to reduce minor allele frequencies (MAF) and, hence, heterozygosity. This effect is clearly observed for SNPs in CDS (Table 3) where stabilizing selection is expected to be largest overall. The data for phylogenetic footprints are more tending towards the effect seen in CDS than those for introns and non-coding regions, but a significant effect can not be observed. A separate analysis for ***Hum*** genes was not performed since we cannot distinguish between functional and non-functional non-coding DNA in the vicinity of these genes.

Transition mutations have an especially high propensity to be function modifying (Freudenberg-Hua *et al.*, 2003; Guo and Jamison, 2005). Therefore, one would expect a reduced ratio of transversions to transitions with increased levels of stabilizing selection. This was indeed observed for coding regions (Table 4) and transversions were biased towards the low-frequency category of SNPs.

Along the same lines one would expect deletion/insertion polymorphisms to be very rare in coding sequences due to the stringency of codon structure. In our analysis, their frequency was reduced by a factor of six in this category as compared to the other categories. A similar effect on introns or phylogenetic footprints would not be expected and was not observed (Table 4).

Overall, measures of stabilizing selection other than SNP density show expected signatures in coding sequences but not in the other sequence categories analyzed in this genome wide, gene centered analysis.

4.6. Weak Correlation between SNP densities of CDS and Footprints. An interesting question is the relationship between functional and regulatory evolution of genes. Wagner *et al.* (2005) showed a correlation between evolution rates of coding sequences and conserved non-coding sequences of posterior HoxA genes following the teleost specific genome duplication. On the other hand, it was suggested recently that differences in the functioning of the human brain as opposed to those of other primates is mainly due to gene expression regulation and not to the occurrence of new genes (Khaitovich *et al.*, 2005). Therefore, the question arises whether co-selection of genes and regulator sites or separate evolution of these sequence elements are predominant in evolution.

In our analysis we see a correlation between SNP densities in phylogenetic footprints and SNP densities in the coding sequence, although a very weak one. This effect is strongest in close vicinity of genes, which is in agreement with the higher propensity of phylogenetic footprint in the vicinity of genes to be functionally relevant.

4.7. Database Biases. An important consideration for the interpretation of large scale statistical analysis of SNP data based on databases is a possible bias of the available SNP data towards well-studied regions. In a recent paper, (Guo and Jamison, 2005) observe, based on NCBI build 33, an increased frequency of SNPs in the close proximity of transcriptional start sites and that SNPs are overrepresented in predicted transcription factor binding sites in comparison to background levels. Similarly, we also found a strong bias towards SNPs in coding regions in earlier versions of ENSEMBL/NCBI. This effect is also described in (Zhao *et al.*, 2003) for the RefSNP database.

In the versions of the data sources used here, (NCBI 35/ENSEMBL 30), these biases are at least greatly reduced. Our findings are indeed consistent with (Zhao *et al.*, 2003), who found that SNP densities are smaller in exonic regions based on Celera's CgsSNP.

4.8. Conclusions for measuring neutral evolution. It is common practice to measure neutral evolution rates from 4-fold degenerate sites in protein coding sequences; The publications reporting the genomes of mouse (Waterston *et al.*, 2002) and the teleost fish *Tetraodon nigroviridis* (Jaillon *et al.*, 2004) may serve as recent prominent examples. In a human/chimpanzee comparison it was demonstrated, however, that about a third of the 4-fold degenerate non-CpG sites evolve under constraints (Hellmann, 2005). This suggests that introns would provide a better model

for neutral evolution despite the fact that some introns contain ncRNAs and other evolutionary conserved RNA secondary structures, see e.g. (Washietl *et al.*, 2005).

In our analyses, we indeed observe a higher SNP density in Introns than in coding sequences (Figure 3). However, the SNP densities observed in introns of genes with phylogenetic homologs beyond mammals (***Tet***, ***Gna***) are still lower than those for the respective intergenic regions. This reduced SNP density in old genes indicates stabilizing selection on intronic DNA. Our data suggest that the least selective pressure is exerted on noncoding intergenic sequences in the vicinity of genes even when regulatory sites are not explicitly excluded from the data. This suggests that, on average, non-coding regions in the vicinity of genes may provide a better model for neutral evolution than four-fold degenerate non-CpG sites or even introns. The specific situation at particular loci of interest should of course be assessed separately.

We performed a genome-wide analyses of diversity and divergence in the vicinity of genes. To this end we examined the distribution and properties of SNPs in and near genes and in sequence elements of particular interest for association studies of human diseases. Our results on a genome-wide scale suggest that strongest evidence of stabilizing selection is observed for coding sequences but is also clearly present for putative regulatory elements detected as phylogenetic footprints. This in turn suggests that a significant proportion of these footprints may indeed be functionally relevant. Our data also suggest that non-coding sequences near genes are on average a better model for neutral evolution than four-fold degenerate non-CpG sites in coding sequences or even introns.

Acknowledgments. We thank Holger Kirsten for helpful discussions and Jörg Reichardt for support with calculating GO-annotation statistics. Funding for this research is gratefully acknowledged: DFG Bioinformatics Initiative grant no. BIZ-6/1-2 to CF and PFS, Sächsische Aufbaubank/F"orderbank grant no. 7692/1187, European Fund for Regional Development (EFRE) grant no. 4212/04-04, and Hochschul- und Wissenschaftsprogramm of the German Federal Ministry for Education and Research to PA.

References

Altschul S.F., Gish W., Miller W., Myers E.W., and Lipman D.J., 1990. Basic local alignment search tool. *J. Mol. Biol.* **215**:403–410.

Barkur S.S., 2002. SNP alleles in human disease and evolution. *J. Hum. Genet.* **47**:561–566.

Blanchette M. and Tompa M., 2002. Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Research* **12**:739–748.

Cargill M., Altshuler D., Ireland J., Sklar P., Ardlie K., Patil N., Shaw N., CR L., Lim E., Kalyanaraman N., *et al.*, 1999. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat Genet.* **22**:231–238.

Dermitzakis E.T. and Clark A.G., 2002. Evolution of transcription factor binding sites in mammalian gene regulatory regions: conservation and turnover. *Mol. Biol. Evol.* **19**:1114–1121.

Duret L. and Bucher P., 1997. Searching for regulatory elements in human noncoding sequences. *Curr. Opin. Struct. Biol.* **7**:399–406.

Fay J.C., Wyckoff G.J., and Chung-I W., 2001. Positive and negative selection on the human genome. *Genetics* **158**:1227–1234.

Fickett J.W. and Wasserman W.W., 2000. Discovery and modeling of transcriptional regulatory regions. *Curr. Opin. Biot.* **11**:19–24.

Fisher R.A., 1935. The logic of inductive inference. *Journal of the Royal Statistical Society Series A* **98**:39–54.

Fisher R.A., 1962. Confidence limits for a cross-product ratio. *Australian Journal of Statistics* **4**:41.

Freudenberg-Hua Y., Freudenberg J., Kluck N., Cichon S., Propping P., and Nöthen M.M., 2003. Single nucleotide variation analysis in 65 candidate genes for CNS disorders in a representative sample of the european population. *Genome Research* **13**:2271–2276.

Guo Y. and Jamison D.C., 2005. The distribution of SNPs in human gene regulatory regions. *BMC Genomics* **6**:140.

Heinemeyer T., Wingender E., Reuter I., Hermjakob H., Kel A.E., Kel O.V., Ignatieva E.V., Ananko E.A., Podkolodnaya O.A., Kolpakov F.A., *et al.*, 1998. Databases on transcriptional regulation: TRANSFAC, TRRD, and COMPEL. *Nucl. Acids Res.* **26**:364–370.

Hellmann I., 2005. *Mutation and Selection as Inferred by the Comparison of the Human and Chimpanzee Genomes*. Ph.D. thesis, University of Leipzig,.

Hughes A.L., Packer B., Welch R., Bergen A.W., Chanock S.J., and Yeager M., 2005. Effects of natural selection on interpopulation divergence at polymorphic sites in human protein-coding loci. *Genetics* **170**:1181–1187,.

Jaillon O., Aury J.M., Brunet F., Petit J.L., Stange-Thomann N., Mauceli E., Bouneau L., Fischer C., Ozouf-Costaz C., Bernot A., *et al.*, 2004. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate protokaryotype. *Nature* **431**:946–957.

Khaitovich P., Hellmann I., Enard W., Nowick K., Leinweber M., Franz H., Weiss G., Lachmann M., and Pääbo S., 2005. Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. *Science* **309**:1850–1854.

Kimura M., 1983. *The neutral theory of molecular evolution*. Cambridge University Press, Cambridge.

Knight J.C., Keating B.J., Rockett K.A., and Kwiatowski D.P., 2003. In vivo characterization of regulatory polymorphisms by alelle-specific quantification of RNA polymerase loading. *Nature Genetics* **33**:469–475.

Kumar S. and Hedges S., 1998. A molecular timescale for vertebrate evolution. *Nature* **392**:917–20.

Ludwig M.Z., Bergman C., Patel N.H., and Kreitman M., 2000. Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature* **403**:564–567.

Majewski J. and Ott J., 2002. Distribution and characterization of regulatory elements in the human genome. *Genome Res.* **12**:1827–36.

Mattick J.S., 2003. Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. *Bioessays* **25**:930–939.

Miller R., Taillon-Miller P., and Kwok P., 2001. Regions of low single-nucleotide polymorphism incidence in human and orangutan xq: deserts and recent coalescences. *Genomics* **71**:78–88.

Prohaska S.J., Fried C., Flamm C., Wagner G.P., and Stadler P.F., 2004. Surveying phylogenetic footprints in large gene clusters: Applications to *Hox* cluster duplications. *Mol. Phyl. Evol.* **31**:581–604.

Sandelin A., Pär Engström W.A., Wasserman W., and Lenhard B., 2004. Jaspar: an open access database for eukaryotic transcription factor binding profiles. *Nucl. Acids Res.* **32**.

Sherry S., Ward M., Kholodov M., Baker J., Phan L., Smigielski E., and Sirotkin K., 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**:308–311.

Subramanian S. and Kumar S., 2003. Neutral substitutions occur at a faster rate in exons than in noncoding dna in primate genomes. *Genome Research* **13**:838–844.

Sunyaev S.R., Lathe W.C., Ramensky V.E., and Bork P., 2000. SNP frequencies in human genes an excess of rare alleles and differing modes of selection. *TIG* **16**:335–337.

Tagle D.A., Koop B.F., Goodman M., Slightom J.L., Hess D.L., and Jones R.T., 1988. Embryonic ϵ and γ globin genes of a prosimian primate (galago crassicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *J. Mol. Biol.* **203**:439–455.

Tautz D., 2000. Evolution of transcriptional regulation. *Curr. Opin. Genet. Dev.* **10**:575–579.

The Human Genome Sequencing Consortium, 2004. Finishing the euchromatic sequence of the human genome. *Nature* **431**:931–945.

The International Human Genome Sequencing Consortium, 2001. Initial sequencing and analysis of the human genome. *Nature* **409**:860–921.

Wagner G.P., Takahashi K., Lynch V., Prohaska S.J., Fried C., Stadler P.F., and Amemiya C.T., 2005. Molecular evolution of duplicated ray finned fisch HoxA clusters: Increased synonymous substitution rate and asymmetrical co-divergence of coding and non-coding sequences. *J. Mol. Evol.* **60**:665–676.

Wang X., Tomso D.J., Liu X., and Bell D.A., 2005. Single nucleotide polymorphism in transcriptional regulatory regions and expression of environmentally responsive genes. *Toxicol. Appl. Pharmacol.* **207**:84–90.

Washietl S., Hofacker I.L., Lukasser M., Hüttenhofer A., and Stadler P.F., 2005. Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. *Nature Biotech* **23**:1383–1390.

Waterston R.H., Lindblad-Toh K., Birney E., Rogers J., Abril J.F., Agarwal P., Agarwala R., Ainscough R., Alexandersson M., An P., *et al.*, 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**:520–562.

Xie X., Lu J., Kulbokas E.J., Golub T.R., Mootha V., Lindblad-Toh K., Lander E.S., and Kellis M., 2005. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature* **434**:338–345.

Zhao Z., Fu Y., Hewett-Emmett D., and Boerwinkle E., 2003. Investigating sinlge nucleotide polymorphsim (SNP) density in the human genome and its implications for molecular evolution. *Gene* **312**:207–213.