GBE

Noncoding Sequences Near Duplicated Genes Evolve Rapidly

Dennis Kostka*,[1], Matthew W. Hahn[2], and Katherine S. Pollard[1,3,4]

[1]Gladstone Institute for Cardiovascular Disease, Gladstone Institutes, University of California–San Francisco

[2]Department of Biology and School of Informatics and Computing, Indiana University

[3]Institute for Human Genetics, University of California–San Francisco

[4]Division of Biostatistics, University of California–San Francisco

*Corresponding author: E-mail: dennis.kostka@gladstone.ucsf.edu.

Accepted: 25 June 2010

Abstract

Gene expression divergence and chromosomal rearrangements have been put forward as major contributors to phenotypic differences between closely related species. It has also been established that duplicated genes show enhanced rates of positive selection in their amino acid sequences. If functional divergence is largely due to changes in gene expression, it follows that regulatory sequences in duplicated loci should also evolve rapidly. To investigate this hypothesis, we performed likelihood ratio tests (LRTs) on all noncoding loci within 5 kb of every transcript in the human genome and identified sequences with increased substitution rates in the human lineage since divergence from Old World Monkeys. The fraction of rapidly evolving loci is significantly higher nearby genes that duplicated in the common ancestor of humans and chimps compared with nonduplicated genes. We also conducted a genome-wide scan for nucleotide substitutions predicted to affect transcription factor binding. Rates of binding site divergence are elevated in noncoding sequences of duplicated loci with accelerated substitution rates. Many of the genes associated with these fast-evolving genomic elements belong to functional categories identified in previous studies of positive selection on amino acid sequences. In addition, we find enrichment for accelerated evolution nearby genes involved in establishment and maintenance of pregnancy, processes that differ significantly between humans and monkeys. Our findings support the hypothesis that adaptive evolution of the regulation of duplicated genes has played a significant role in human evolution.

Key words: accelerated substitution, noncoding sequence, gene duplication.

Introduction

The genetic basis for human-specific traits is of great interest. Despite striking phenotypic divergence in the Hominini (members the of human–chimp lineage), genome sequence data suggest a slowdown in the rate of nucleotide substitutions in humans and our primate relatives (Wu and Li 1985; Huttley et al. 2007). Furthermore, orthologous human and chimpanzee proteins differ by only two amino acid substitutions on average, and nearly a third of proteins are identical between the two species (The Chimpanzee Sequencing and Analysis Consortium 2005). Low amino acid divergence in the "hominin" (human–chimp) lineage lends support for the hypothesis that divergence between closely related species is accompanied by evolution of the gene regulatory network (King and Wilson 1975; Levine and Tjian 2003; Carroll 2005).

Structural variation in the genome is another mutational mechanism that contributes significantly to genomic divergence (Wilson et al. 1974; Kent et al. 2003). An increased rate of structural genomic rearrangements (such as gene duplications) has been observed in primates (Cheng et al. 2005; She et al. 2006; Hahn et al. 2007; Marques-Bonet et al. 2009). Structural variation has been recognized as a major contributor to genomic diversity, with gene duplication serving as an evolutionary mechanism for functional innovation (Ohno 1999; Zhang 2003). Also, gene turnover in the form of rapid expansion or contraction of gene families has been put forward as a possible explanation of phenotypic divergence (Fortna et al. 2004; Marques et al. 2005; Demuth et al. 2006; Dumas et al. 2007; Zhu et al. 2007; Perry et al. 2008). Additionally, evidence for excess positive selection on the coding sequences of genes in families that

expanded rapidly in primates corroborates the hypothesis that gene duplication can lead to functional innovation (Hahn et al. 2007).

The beta subunit of the glycoprotein hormone chorionic gonadatropin (*CGB*), for example, is believed to have arisen by duplication of luteinizing hormone beta (*LHB*) about 35–50 million years ago. This duplication was followed by one deletion and two insertions in the coding sequence that lead to the appearance of a carboxy-terminal peptide in *CGB*. Additional mutations in the promoter induced an expression shift from pituitary gland to placenta (Maston and Ruvolo 2002; Henke and Gromoll 2008). The evolution of *CGB* illustrates the process of gene duplication followed by neofunctionalization. Notably, the emergence of new function involved changes in both the coding sequence and nearby noncoding sequence, which provided a new regulatory context. This example highlights a particular type of gene evolution whereby adaptation occurs in regulatory noncoding sequences after a duplication event. We hypothesize that this form of divergence played a significant role in human evolution.

The ever-increasing richness of genomic sequence and functional data provides a foundation for empirical studies of duplication-mediated divergence. Although there have been some large-scale studies on the evolution of duplicated loci in human (Lynch and Conery 2000; Kondrashov et al. 2002; Lynch and Conery 2003; Zhang et al. 2003; Shiu et al. 2006; Hahn et al. 2007; Marques-Bonet et al. 2008; De et al. 2009), these analyses have typically focused on coding rather than noncoding sequences. Bioinformatic challenges related to sequence assembly and alignment, as well as interpretation of evolutionary analyses, are largely responsible for the paucity of genome-wide studies of noncoding sequences to date. In light of the mounting evidence that regulatory divergence has played a major role in hominin evolution, however, it is desirable to understand how noncoding sequences in and near duplicated loci evolve.

These considerations motivated us to perform a systematic genome-wide search for signatures of accelerated sequence evolution and functional innovation in noncoding sequences associated with duplicated genes. We focus on the hominin lineage since divergence from the common ancestor with Old World Monkeys (represented by the macaque genome). The hominin lineage is very relevant to our understanding of human evolution, and statistical tests on this lineage have greater power than tests on the much shorter human lineage since divergence from the chimp–human ancestor. For this analysis, we identified all noncoding sequences within 5 kb of a human Ensembl transcript. We then performed molecular evolutionary tests to highlight regions of unusually high substitution rates in the Hominini. Next, we assessed the likely impact of human–macaque sequence differences on transcription factor (TF) binding, thereby identifying noncoding regions likely to have affected transcriptional regulation. Finally, we asked whether noncoding sequences associated with genes that duplicated in the hominin lineage show stronger evidence of divergence than noncoding sequences nearby nonduplicated genes. We find strong enrichment for accelerated substitution rates and transcription factor–binding site (TFBS) divergence in noncoding sequences associated with duplicated genes.

Materials and Methods

Sequence Data and Orthologous Sequence Blocks

We downloaded 28-way alignments in multiple alignment format (MAF) from the genome browser maintained by the University of California–Santa Cruz (UCSC) (hg18, NCBI assembly version 36). MAF-formatted alignments are partitioned into consecutively aligned sequence blocks (Blanchette et al. 2004) that can be viewed as orthologous units. A MAF sequence block is a local alignment where each row represents consecutive (though potentially gapped) sequence from one species, and there are no gap-only columns. Moving along the human chromosomes, a new block starts when there is a change in orthology (e.g., a species drops in or out of the alignment). We used this "natural" partition of whole-genome multiple sequence alignments in our analyses. In cases of duplications, both of the duplicated sequences in the Hominini are considered orthologous to single-copy regions in the outgroups. Thus, the orthologous outgroup sequences may appear in more than one MAF block. We included only the genomes with data-use policies allowing genome-wide analysis (human, chimp, macaque, mouse, rat, dog, opossum, platypus, chicken, zebrafish, fugu, and medaka) in our analyses. Not all species are present in all MAF blocks.

We delineated all noncoding MAF blocks that are located within 5 kb of a human Ensembl transcript (version 41) (Flicek et al. 2008), that is, a region spanned by the transcriptional unit plus 5 kb of upstream and downstream flanking sequence. We trimmed off coding sequence from any block overlapping a coding exon.

Block Delineation

We functionally annotated each block's "genic location" based on Ensembl gene models with the following categories:

- Flanking region (5′, 3′, or ambiguous, see below)
- Exonic 5′ untranslated regions (UTR)
- Intronic 5′ UTR
- First intron, early (≤500 bp 3′ from intron start)
- First intron, late (>500 bp 3′ from intron start)
- Intron (all other intronic sequence)
- Intronic 3′ UTR
- Exonic 3′ UTR.

We defined exonic 5′ UTRs as the sequences between the transcription start site (TSS) and the coding start site (CSS) that are annotated as exons; intronic 5′ UTRs contain all other sequence between the TSS and CSS. Analogously, we defined exonic 3′ UTRs as sequences between coding stop and transcription stop that are annotated as exons. Intronic 3′ UTRs are all other sequences between coding stop and transcription stop. We clustered overlapping transcripts and annotated each sequence block with a unique genic location category using the following hierarchy: exonic UTR > intronic UTR > 5′ > 3′ > first intron early > first intron late > intron > flanking sequence. For example, if a block overlaps both exonic 5′ UTR and first early intron sequence (due to overlapping transcripts), we annotated it as exonic 5′ UTR. Flanking regions were annotated as 5′ (upstream of TSS), 3′ (downstream of transcription stop site), or ambiguous (if not uniquely 5′ or 3′ due to overlapping or nearby transcripts). We refer to contiguously transcribed genomic regions on either strand as "transcript clusters."

Alignment Quality Filters

To produce a data set with high-quality syntenic alignments, we used the following filtering criteria to exclude certain alignment blocks from further analysis:

- Blocks not containing chimp and macaque plus at least one other placental mammal with no more than 50% nongap characters were excluded.
- Blocks with more than 1/3 of bases (chimp or macaque) inserted or deleted with respect to human were excluded.
- Blocks with more than 1/2 gap characters in human, chimp, and macaque were excluded.
- Blocks with more than 25% of all nongap bases differing between human and chimp (or 35% between human and macaque) were excluded.
- Blocks were masked if more than 1/2 of their bases were repeat masked in human.
- Blocks with more than 1/2 of their bases overlapping annotated pseudogenes (Ensembl version 54) were excluded.
- Blocks that were not syntenic between human and chimp were excluded.

Additionally, quality scores for chimp and macaque were taken from the UCSC genome browser databases rheMac2 and panTro2 (table "quality") and bases with a score less than 40 were masked in both species. Synteny was derived from human–chimp alignments: Syntenic net alignment files were downloaded from UCSC, and syntenic regions were defined as top-level chain alignments of at least 5-Mb length; gaps in this chain were filled with syntenically aligned chains from lower levels. Repeat masking was performed on the basis of the rmskRM327 track downloaded from UCSC.

This quality filtering produced a data set of 4,699,477 high-quality MAF blocks with median length of 64 bp (range 10–2,580 bp). These blocks cover 410,274,564 bp of the human genome. Alignment filtering might introduce various biases in patterns of sequence composition and conservation, such as biased retention of more conserved blocks over less conserved blocks. We addressed this issue by using calibrated chromosome-specific models for unconstrained sequence in our LRTs, which were then rescaled using the data from each block to produce a local null model (see below). We believe that the benefits of more trustworthy alignments outweigh the potential drawbacks of a higher false-negative rate (i.e., filtered-out blocks that are not tested) and a bias toward analysis of more conserved blocks.

Likelihood Ratio Test for Accelerated Substitution Rates

To test for acceleration in the rate of nucleotide substitutions in the Hominini, we subjected each alignment block to a one-sided LRT using phylogenetic models (Yang et al. 1994). Phylogenetic continuous time Markov models are parameterized by a tree T (topology and branch lengths), a rate matrix Q, and equilibrium base frequencies π. The tree T can be viewed as consisting of a tree $T-$, spanning the hominin lineage, and a tree $T+$, spanning the other species: $T = (T+, T-)$. We used a general time-reversible parameterization (REV) of the rate matrix Q. All model fitting was performed using maximum likelihood estimation with the "phyloFit" program from the PHAST package (http://compgen.bscb.cornell.edu/phast/). To assess the robustness of our results, we also calculated test statistics after deleting chimp from T, so that $T-$ consisted of the human lineage alone (from the macaque–human ancestor to modern humans). To keep the two analyses comparable, we performed this analysis on the same set of filtered MAF blocks as before. Leaving chimp in the filtering rules creates a bias toward well-aligned blocks (conservative with respect to identifying accelerated substitution rates), whereas excluding chimp from the LRT analysis guards for false positives presumably due to misalignment and/or erroneous assembly of the low-coverage shotgun sequenced chimp genome.

Test Statistics. The LRT statistic for an alignment block B is based on the ratio of the likelihood of the sequence alignment under two different models: 1) a null model and 2) an alternative model with acceleration in the hominin lineage that allows a faster rate of substitutions on the human and chimp branches, as described in Pollard et al. (2009). In more detail, the null and the alternative models were estimated starting from a chromosome-specific model $M = (Q, T, \pi)$ for unconstrained sequence. The null model was then obtained by rescaling T by a constant γ to allow for a faster or slower overall rate of substitutions across the whole tree

(i.e., in all species), thereby maintaining the same relative branch lengths (i.e., substitution rates) across lineages. This rescaling step accounts for local substitution rate variation. The alternative model also adjusts for local rate variation but includes an additional parameter $\rho > 1$, which allows the hominin lineage to have a faster rate of substitutions relative to the rest of the tree. Thus, the rate of substitutions in the hominin lineage compared with the rest of the tree is increased, whereas relative rates of substitutions across lineages in the rest of the tree are maintained. The models can be represented as follows:

$$\text{Null model } (M0): \; T = \gamma(T+, T-)$$

$$\text{Alternative model } (M+): \; T' = \gamma(T+, \rho \times T-), \rho > 1$$

The LRT statistic for block B is

$$S = \log[P(B|M+)/P(B|M0)],$$

where the parameters γ and ρ are estimated for each model by maximum likelihood, using the alignment data for block B. The null hypothesis being tested is H_0: $\rho = 1$ (no acceleration in Hominini) versus the alternative hypothesis H_a: $\rho > 1$ (acceleration in Hominini).

Chromosome-Specific Initial Models. We detected lineage-specific increases in substitution rates by comparing an alternative model $M+$ with a null model $M0$ (see previous paragraph). Because both of these models are rescaled versions of a model M for unconstrained sequence, it is important to pick M such that rescaling its tree T provides a reasonable estimate of the local substitution process at block B. On the one hand, it is desirable that the base frequencies π and relative rates of various types of substitutions in the rate matrix of M are as appropriate for B as possible. On the other hand, the better the initial model M fits the data in block B, the more difficult it is to reject the null hypothesis. This provides a dilemma as to how "local" (with respect to B) the initial model should be. We examined a range of choices, from a single genome-wide model to a rescaled local model fit on several megabases of nearby sequence. Based on this analysis, we chose to estimate separate initial models for each chromosome (see supplementary section SR3, Supplementary Material online). In this way, we attempt to accommodate chromosome-specific biases (e.g., sequence composition and substitution patterns) and condition on the fact that chromosomes, as a whole, are not accelerated. Scaling the null model $M0$ by γ ensures that more local rate variation is also accounted for in the test.

To estimate chromosome-wide models, we performed maximum likelihood estimation with starting values for parameter optimization obtained from a genome-wide REV model fit to 4-fold degenerate sites from 28-way alignments (obtained from UCSC). We estimated γ and ρ, as well as equilibrium frequencies, using all blocks on a given chromosome, by maximum likelihood. These chromosome-specific models were then subsequently used as the initial models (M) for the LRT analysis described above.

Statistical Significance. To determine significantly accelerated blocks, we calculated P values for the observed LRT scores using the asymptotic distribution of a 50:50 mixture of a point mass at zero and a χ^2 distribution (Self and Liang 1987). As our data contain LRT scores from blocks of different lengths, we checked the correlation between block length and LRT score. We found only a minimal association ($r = 0.051$). From the P values, we derived an LRT score cutoff controlling the false discovery rate (FDR) at 10% using the Benjamini-Hochberg procedure (R package multtest; http://www.bioconductor.org). Blocks with an LRT score larger than the cutoff were termed as "accelerated."

Duplication Data

Duplicated Genes. We used gene tree reconciliation to determine a set of Ensembl (version 41) peptides that underwent duplication on the lineage between the macaque–hominin ancestor (common ancestor of human, chimp, and macaque) and the chimp–human ancestor. We constructed neighbor-joining trees for each family defined by Ensembl using the peptide sequences from human, chimpanzee, macaque, rat, mouse, and dog. We reconciled the resulting gene trees with the species tree of the six species using NOTUNG (Durand et al. 2006). Duplications that occurred on the lineage leading to Hominini after the split with macaque but before the human–chimpanzee split were identified as Hominini specific. These duplicated genes were required to have synonymous divergence less than 0.064 (twice the average distance back to the human–macaque ancestor). This approach yielded 716 unique Hominini-specific Ensembl (version 41) peptides. We refer to these as duplication peptides. After alignment filtering (see above), we retain noncoding blocks in 5-kb neighborhoods of 449 duplication peptides. We call these as duplication-associated blocks or "DA blocks."

Duplicated sequences are often located a long distance from the locus they are copied from, especially in mammals (She et al. 2006; McGrath et al. 2009). When paralogous genes are located far apart, we can define one as the "parent" and one as the "daughter"; these correspond to the paralog in the original location and the paralog in a new location, respectively (Han and Hahn 2009). We used the likelihood method of Han and Hahn (2009) to define parent and daughter duplicates (when possible) based on the length of shared synteny between the human copies of the hominin-specific duplicates and the single-copy genes in macaque. In total, we were able to uniquely polarize 95 peptides. These

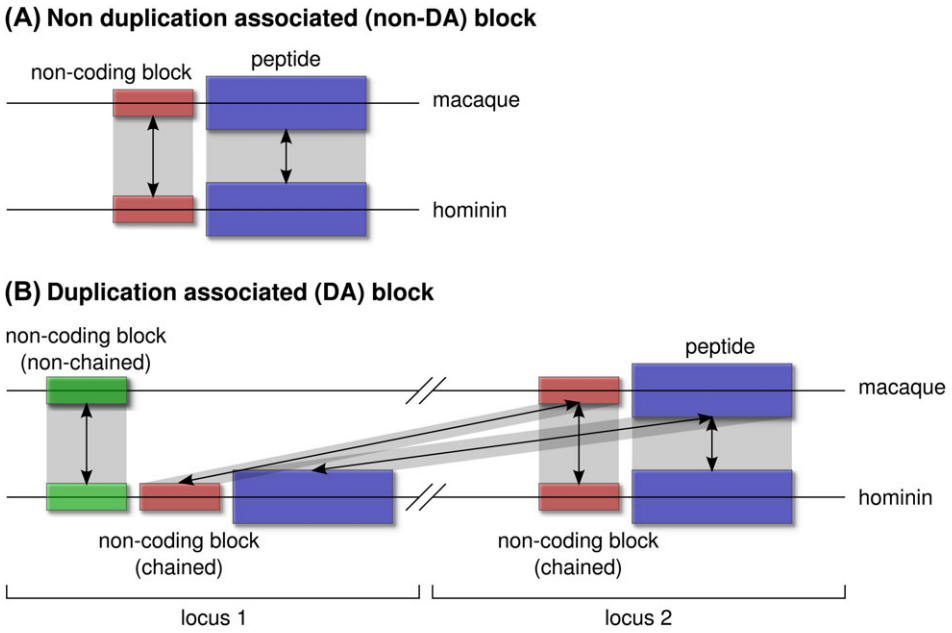

Fig. 1.—Examples of two types of alignment blocks. Panel (*A*) shows a non-DA noncoding block (red) that is associated with a nonduplicated peptide (blue). Panel (*B*) shows examples of two different types of DA noncoding blocks (red and green). The peptide (blue) is duplicated in the Hominini, and paralogous copies are present at two loci. The two red noncoding blocks and the peptide align to the same genomic regions in macaque (chained blocks, see Materials and Methods for definition). In contrast, the green noncoding block aligns to a different region in macaque compared with the peptide (nonchained block, see Materials and Methods). There are other scenarios that generate both types of DA blocks, but these are common examples. Panel (*B*) depicts the presence of both types of DA blocks (chained and nonchained) nearby a single peptide. This is the case in 17% of the 459 duplicated genes (transcript clusters, see Materials and Methods) in our study; 49% have exclusively chained blocks nearby, whereas 34% have only nonchained blocks in their proximity. Note that panel (*A*) depicts a chained non-DA block. This is for illustration purposes, and nonchained non-DA blocks are also possible, although they do not play an explicit role in our analysis.

correspond to 56 parents and 39 daughters (multiple parents may exist when the original locus is duplicated both in tandem and to a distant location).

Duplication Status of DA Blocks. Not all DA blocks are themselves duplicated (fig. 1). To delineate the duplication status of noncoding blocks, we used human–macaque alignment nets (UCSC hg18.netRheMac2). We annotated noncoding blocks based on the alignment chain they reside on compared with the exons of duplication peptides within 5 kb. We find that the majority of transcripts of duplication peptides have exons that align to a single unique macaque alignment net (66%). DA blocks within 5 kb of those peptides and on the same chain as the exons were annotated as "chained." Specifically, a "chained DA block" is (A) a DA block that is part of a multiple sequence alignment (MAF block) that meets our quality filters and is located within 5 kb of a recently duplicated peptide (human coordinates and annotation), and (*B*) in the syntenic net alignment files, it is on the same chain as the duplicated peptide. DA blocks that are not chained are called nonchained. We refer to other noncoding blocks within 5 kb of duplication peptides as nonchained DA blocks (see, e.g., fig. 1). This annotation does not directly reflect the duplication status of noncoding blocks, but chained blocks are much more likely to share the evolutionary history of the exons of a nearby transcript compared with nonchained blocks. Chained DA blocks near daughter peptides are likely to have been duplicated along with the duplicated coding sequence, whereas nonchained DA blocks are more likely to be nonduplicated or to have duplicated separately from the nearby duplicated peptide. This approach enables us to roughly estimate the duplication status of noncoding blocks, even though reliable phylogeny-based duplication information is currently only available on genome-wide scale for coding sequences.

Because parent loci are less likely than daughter loci to have been affected by genomic rearrangements, we expect to see an enrichment of chained blocks near parent peptides compared with daughter peptides. This is indeed the case: 49 of the 56 parent peptides (88%) have chained noncoding blocks nearby, whereas this is only the case for 6 of the 39 daughter peptides (15%).

Genes Near Accelerated DA Blocks. To determine genes nearby DA blocks, we mapped the Ensembl version 41 peptides near each accelerated (i.e., with a significant LRT; see above) DA block to Ensembl version 56 using the BioMart data management system (http://www.ensembl.org/biomart).

Testing for Association with the Fisher's Exact Test

From our analyses, we have a variety of block-level annotations (accelerated, duplication status, genic location, etc.). We test for association between pairs of annotations by calculating the corresponding contingency table (either on all or on a subset of blocks) and applying the Fisher's exact test (FET) for independence. The FET is significant if there is evidence that an annotation is enriched within one category of another annotation, for example, acceleration is more common among duplicated versus nonduplicated loci. We also compute the corresponding odds ratio (OR), which measures the direction and magnitude of the association. An OR greater than 1 reflects a positive correlation, whereas an OR less than 1 indicates a negative association.

Transcription Factor–Binding Site Turnover

We developed a method to estimate the number of TFBS lost and gained on the hominin lineage since the hominin–macaque ancestor. Our approach scores human and macaque sequence variants in an MAF block for TF-binding potential and compares predicted binding sites between the two species (Kostka D, Holloway AK, Pollard KS, manuscript in preparation).

Briefly, we downloaded binding motifs and annotation for 11 TF families from the JASPAR FAM database (Sandelin and Wasserman 2004; Wasserman and Sandelin 2004) and regularized their weight matrices by adding 0.01 to each entry. These matrices can be used to scan genome sequences for matches to the TF-binding motif. For each family, a significance threshold for matches (i.e., predicted TFBS) was computed using a method that balances Type I and Type II errors (Rahmann et al. 2003).

For each alignment block and TF family, we predicted TFBS in the human and macaque sequences, after removing gaps from the alignment. We then calculated *P* values for the difference in binding site predictions in the two species under a model of two correlated Bernoulli processes. Specifically, for each block and TF, we model the prediction of TFBS in a single sequence as a Bernoulli trial. These trials are correlated between human and macaque because their sequences are related through homology. The table below presents the probabilities for the different outcomes. The probability of a prediction in one species but not the other, denoted *c*, is constrained to be smaller than min(p, $1 - p$), where p is the probability of a match (i.e., TFBS prediction for that TF). The number of trials is the average gap-free sequence length of human and macaque minus the length of the TF motif plus 1.

		Human		
		Prediction	No prediction	
Macaque	Prediction	$p - c$	c	p
	No prediction	c	$1 - p - c$	$1 - p$
		p	$1 - p$	

We estimate p for each TF from genome-wide data. It is the fraction of all predictions divided by the total number of trials. Conditional on this estimate, we obtain the maximum likelihood estimate for *c*, which is based on the observed number of differences between human and macaque in TFBS predictions per block. Then, for each block and TF, we calculate a *P* value based on the estimates of p and *c*. More details can be found in (Kostka D, Holloway AK, Pollard KS, in preparation). Finally, we correct the *P* values for multiple testing using the FDR controlling procedure of Benjamini and Hochberg (1995).

This method allowed us to identify alignment blocks with TFBS turnover (gain or loss) while controlling the FDR at 10%. Our approach naturally takes the different block lengths into account, which allows for a meaningful comparison between different alignment blocks. We expect that two facts help to mitigate common problems (like high false-positive rates) generally encountered in single species TFBS prediction: 1) JASPAR FAM family motifs are of high quality and 2) we focus on differential predictions between human and macaque.

Gene Ontology Term Enrichment and Depletion Tests

We performed gene ontology (GO) enrichment and depletion analyses to determine if different block annotation groups (e.g., duplicated, accelerated) were significantly enriched for any GO functional categories. GO terms were mapped from transcripts to all blocks within 5 kb. Thus, a block may receive multiple GO terms from each of multiple transcripts. For each enrichment test, we first defined a reference "universe" of blocks from which the "target" annotation group is drawn. For example, the target group of syntenic blocks is drawn from the universe of all DA blocks. The role of the universe is to provide an appropriate null distribution of GO term frequencies. Enrichment of each GO term in the target set compared with the universe was assessed using standard one-tailed hypergeometric tests. Note that both the universe and target sets are groups of MAF blocks, not groups of peptides. Enrichment testing on the block level corrects for the fact that the number of MAF blocks per transcript is variable, which can create bias in gene-level enrichment tests (Taher and Ovcharenko 2009).

When comparing the target group of DA blocks to the universe of all blocks, we restrict ourselves to report the GO Slim subset (downloaded from http://www.ebi.ac.uk/GOA) of GO terms (see tables 1 and 2).

Results

Accelerated Substitution Rates in the Hominini

To identify the fastest evolving regulatory sequences in the Hominini, we scored all noncoding regions associated with a human gene for evidence of accelerated substitution rates

Table 1
Significantly Enriched GO Slim Categories in Duplicated Loci

GO Slim Term	*P* Value	Term Name
Biological process		
GO:0008152	0.00	Metabolism
GO:0050896	1.90×10^{-59}	Response to stimulus
GO:0006118	2.49×10^{-55}	Electron transport
GO:0007154	5.45×10^{-27}	Cell communication
GO:0006810	3.28×10^{-07}	Transport
Cellular compartment		
GO:0005576	0.00	Extracellular region
GO:0005737	1.96×10^{-220}	Cytoplasm
GO:0005622	6.95×10^{-71}	Intracellular
GO:0005634	6.73×10^{-06}	Nucleus
Molecular function		
GO:0016787	0.00	Hydrolase activity
GO:0016853	9.11×10^{-141}	Isomerase activity
GO:0003824	3.59×10^{-138}	Catalytic activity
GO:0015075	6.01×10^{-125}	Ion transporter activity
GO:0016491	2.48×10^{-108}	Oxidoreductase activity
GO:0004871	4.60×10^{-106}	Signal transducer activity
GO:0030188	9.41×10^{-67}	Chaperone regulator activity
GO:0004872	1.19×10^{-28}	Receptor activity
GO:0005488	3.39×10^{-24}	Binding
GO:0005515	1.73×10^{-17}	Protein binding
GO:0003774	1.25×10^{-07}	Motor activity
GO:0005215	1.79×10^{-02}	Transporter activity

Table 2
Significantly Depleted GO Slim Categories in Duplicated Loci

GO-Slim Term	*P* Value	Term Name
Biological process		
GO:0007275	4.84×10^{-63}	Development
GO:0030154	6.53×10^{-47}	Cell differentiation
GO:0008219	3.36×10^{-20}	Cell death
GO:0006139	2.05×10^{-18}	Nucleobase, nucleoside, nucleotide, and nucleic acid metabolism
GO:0006944	6.33×10^{-10}	Membrane fusion
GO:0007610	1.73×10^{-06}	Behavior
GO:0046903	6.84×10^{-04}	Secretion
Cellular compartment		
GO:0005578	3.60×10^{-50}	Extracellular matrix (sensu Metazoa)
GO:0005694	1.45×10^{-24}	Chromosome
GO:0009986	2.62×10^{-12}	Cell surface
GO:0005615	3.02×10^{-10}	Extracellular space
Molecular function		
GO:0016301	1.19×10^{-92}	Kinase activity
GO:0005198	1.02×10^{-50}	Structural molecule activity
GO:0016874	1.12×10^{-48}	Ligase activity
GO:0030528	1.03×10^{-34}	Transcription regulator activity
GO:0004386	6.05×10^{-34}	Helicase activity
GO:0003676	3.43×10^{-21}	Nucleic acid binding
GO:0008565	2.01×10^{-12}	Protein transporter activity

since divergence from the macaque–hominin ancestor. Specifically, we used whole-genome multiple sequence alignments of up to 12 vertebrates to identify short alignments of orthologous sequence within 5 kb of all human Ensembl transcripts (Flicek et al. 2008). After strict filtering to ensure high-quality syntenic alignments (see Materials and Methods), we obtained a set of ~4.7 million alignments covering ~410 Mb of the human genome (median length 64 bp). We call these regions "blocks" because they are derived from the putatively orthologous sequence alignment blocks in MAF files. This approach to identifying candidate regions for evolutionary analysis allows alignability and conservation to define orthologous regions of variable length, in contrast to windows of arbitrary fixed size or restriction to a predefined set of genomic elements. Also, candidate regions are not inferred to be evolutionarily conserved in (a subset of) the species in our analysis; this is different form previous work focusing on conserved noncoding elements (Pollard, Salama, Lambert et al. 2006; Prabhakar et al. 2006; Bird et al. 2007; Kim and Pritchard 2007). Next, we performed an LRT on each block to detect lineage-specific acceleration in substitution rate in the Hominini since the macaque–hominin ancestor (see Materials and Methods).

Controlling the FDR at 10% (see Materials and Methods), we found 3,805 blocks (~0.081%) with significant evidence of accelerated substitution rate in the Hominini. These accelerated blocks cover 611,318 bp of the human genome (0.15% of the noncoding bp analyzed). Accelerated blocks tend to be slightly longer than nonaccelerated blocks, as expected because the power of our test is higher in longer blocks. But this trend does not translate into a strong correlation between block length and LRT statistic (see Materials and Methods). Accelerated blocks have roughly the same GC content as the average MAF block (43% accelerated vs. 40% average), but they tend to be in gene-rich regions. Although the average noncoding block in our data set is within 100 kb of 2.4 genes (i.e., transcript clusters; see Materials and Methods), accelerated blocks are within 100 kb of 3.1 genes on average.

Noncoding Regions Near Duplicated Genes Evolve Rapidly. We hypothesized that adaptive evolution favoring gene expression divergence after duplication may have generated an excess of accelerated blocks nearby duplicated genes. To explore this idea, we employed gene tree to species tree reconciliation (Durand et al. 2006) based on Ensembl peptide and gene family annotations to identify duplication events in the mammalian phylogeny. Using these duplication histories, we defined "expanded" gene families as sets of homologs with more members in the Hominini than in macaque (see Materials and Methods). We refer to the noncoding regions within 5 kb of a peptide in an expanded gene family as "DA loci" and we call the 26,283 blocks in these loci "DA blocks." Note that proximity to a peptide in an expanded family does not necessarily

Table 3
Fisher Exact Tests

No.	Blocks	Comparison	*P* Value	OR	95% CI
1	All	DA? Accelerated	$<1 \times 10^{-15}$*	8.41	7.18–9.82
1a	All	Chained DA? Accelerated	$<1 \times 10^{-15}$*	4.05	3.1–5.21
2	All	Polarized? Accelerated	0.29	1.34	0.67–2.39
3	DA and polarized	Daughter? Accelerated	0.001*	7.82	1.98–36.43
4	DA	Nonchained? Accelerated	$<1 \times 10^{-15}$*	4.8	3.48–6.67
5	DA and chained and polarized	Daughter? Accelerated	0.002*	23.69	2.71–282.99
6	DA	Nearby Accelerated Protein? Accelerated	0.45	0.46	0.055–1.71
7	All	Accelerated? TFBS	$<1 \times 10^{-15}$*	8.04	6.39–10
8	All	DA? TFBS	0.03*	1.26	1.02–1.55
9	DA	Accelerated? TFBS	3.12×10^{-05}*	11	3.87–25.46
10	Accelerated	DA? TFBS turnover	0.27	1.68	0.59–3.90

[a] Significant tests.

imply the noncoding block itself was duplicated (fig. 1 and see below).

To functionally characterize this set of DA blocks, we conducted GO (Ashburner et al. 2000) enrichment and depletion analyses. GO analyses were performed using a novel method that maps GO terms to noncoding elements and performs statistical analysis on the elements themselves, rather than the genes (see Materials and Methods). This approach adjusts for the different distributions of noncoding elements around different categories of genes (Taher and Ovcharenko 2009). We found that terms related to signal transduction, response to stimulus, and metabolic processes are enriched among DA blocks compared with all MAF blocks (tables 1 and 2).

Using FET and ORs, we investigated evidence of acceleration in DA blocks compared with non-DA blocks. Table 3 presents an overview of all the tests we conducted. Each row corresponds to a test for association, and the columns contain the type of blocks considered in the comparison, the attributes compared, FET *P* value, and OR with a 95% confidence interval (CI). Contingency tables for each comparison can be found in the supplement (supplementary results SR5, Supplementary Material online). We identified 171 significantly accelerated DA blocks (see table S2 in the supplement). Thus, acceleration is roughly ten times more common in DA blocks compared with non-DA blocks (171/26,283 $\approx$ 0.65% compared with 0.078% in non-DA blocks). This enrichment of accelerated blocks near duplicated genes is highly significant (FET: $P < 1 \times 10^{-15}$, OR = 8.41 [95% CI, 7.18–9.82]; see row one in table 3). The 48 genes with accelerated DA blocks nearby are an interesting set of candidates for functional divergence in Hominini (table 4). Many of these genes are paralogous members of the same families and/or belong to related pathways (see below).

Performing GO enrichment tests to compare accelerated with nonaccelerated DA blocks, we found that accelerated DA blocks are enriched in GO terms related to transferase activity (glycosyl and hexolyl groups), metabolism (steroid and estrogen), G-protein–coupled receptor (GPCR) activity (including olfactory receptors), and visual perception (table 5). Notably, enriched terms also include female gamete generation, whereas depleted terms include spermatogenesis (table 6).

Noncoding Regions Near Daughter Genes Are More Accelerated than Their Parents. Having established a strong association between duplicated loci and accelerated substitution rates, we next attempted to delineate patterns of accelerated evolution under different duplication scenarios (fig. 2). A subset of DA blocks can be polarized to be associated uniquely with either a daughter peptide (new genomic location) or a parent peptide (preduplication genomic location; see Materials and Methods). Both types of polarized DA blocks show roughly the same length distribution (see supplementary fig. S1, Supplementary Material online). Although the set of polarized DA blocks is not enriched for acceleration compared with all other MAF blocks (FET: $P = 0.29$, OR = 1.34 [95% CI, 0.67–2.39,]; see row two in table 3), we see a clear pattern among polarized DA blocks themselves. DA blocks near daughters are significantly enriched for acceleration compared with DA blocks near parents (FET: $P = 0.001$, OR = 7.82 [95% CI, 1.98–36.43]; see row three in table 3). Thus, for the subset of duplicated peptides where we can infer a parent–daughter relationship, we find much faster noncoding evolution in the regulatory regions of the newly formed daughter gene. This asymmetry parallels the pattern seen in the protein sequences of parent and daughter duplicates (Han et al. 2009).

To assess if acceleration of DA blocks happened before or after the duplication of peptides on the hominin lineage, we compared LRT statistics from polarized blocks near parent and daughter peptides. We find minimal

Table 4
Genes Near Accelerated DA Blocks

Ensembl Gene ID	Symbol	Chromosome	Ensembl Gene ID	Symbol	chr
ENSG00000162365	CYP4A22	1	ENSG00000187134	AKR1C1	10
ENSG00000186160	CYP4Z1	1	ENSG00000120563	LYZL1	10
ENSG00000072694	FCGR2B	1	ENSG00000086205	FOLH1	11
ENSG00000181773	GPR3	1	ENSG00000205046	OR4A47, OR4A4P	11
ENSG00000162849	KIF26B	1	ENSG00000205496	OR51A2	11
ENSG00000204513	PRAMEF11	1	ENSG00000205497	OR51A4	11
ENSG00000204502	PRAMEF5	1	ENSG00000226288	OR52I2	11
ENSG00000232423	PRAMEF6	1	ENSG00000204450	TRIM64	11
ENSG00000187010	RHD	1	ENSG00000134551	PRH2	12
ENSG00000136682	CBWD2	2	ENSG00000132341	RAN	12
ENSG00000163040	CCDC74A	2	ENSG00000173262	SLC2A14	12
ENSG00000152076	CCDC74B	2	ENSG00000184227	ACOT1	14
ENSG00000173272	FAM128A	2	ENSG00000206149	HERC2P2, HERC2P3	15
ENSG00000185304	RGPD2	2	ENSG00000197711	HP,HPR	16
ENSG00000232382	OR5K1	3	ENSG00000140992	PDPK1	16
ENSG00000184203	PPP1R2	3	ENSG00000204414	CSHL1	17
ENSG00000213759	UGT2B15, UGT2B11	4	ENSG00000170832	USP32	17
ENSG00000135226	UGT2B28	4	ENSG00000189052	CGB5	19
ENSG00000122194	PLG	6	ENSG00000196337	CGB7	19
ENSG00000105835	NAMPT	7	ENSG00000174667	OR7D4	19
ENSG00000177076	ACER2	9	ENSG00000161643	SIGLECP16	19
ENSG00000230453	ANKRD18B	9	ENSG00000165583	SSX5	X
ENSG00000187559	FOXD4L3	9	ENSG00000204648	SSX9	X
ENSG00000137080	IFNA21	9	ENSG00000198205	ZXDA	X

correlation between LRTs in parents and daughters on average (supplementary fig. S4, Supplementary Material online), suggesting that accelerated substitutions happened independently in one or both copies after duplication (supplementary results SR4, Supplementary Material online).

The Fastest Evolving Noncoding Blocks Did Not Duplicate with the Associated Gene. Based on our finding that daughter peptides are enriched with accelerated noncoding blocks, we further investigated the relationship between DA blocks and the genes with which they are associated. Using human–macaque alignments, we annotated each DA block as either chained (same alignment chain as the gene's exons) or nonchained (all other DA blocks; see Materials and Methods and fig. 1). Comparing rates of acceleration in both sets, we found that nonchained DA blocks are significantly enriched for acceleration compared with chained ones (FET: $P < 1 \times 10^{-15}$, OR = 4.8 [95% CI, 3.48–6.67]; see row four in table 3). This suggests that noncoding sequences that are close to—but not included in—duplication events evolve more rapidly than noncoding sequences that are either 1) far away from duplication events or 2) duplicated alongside a gene. In light of these results, we asked whether our previous results that acceleration is enriched in DA blocks still holds for chained DA blocks. We find that this is indeed the case (FET: $P < 1 \times 10^{-15}$, OR = 4.05 [95% CI, 3.1–5.1]; see row two in table 3).

We note that accelerated substitutions in nonchained DA blocks have a distinct interpretation from the same phenomenon in chained DA blocks. Although the latter induce changes in duplicated sequences, the former affect "ancestral" sequences close to duplicated loci. That is, noncoding sequence that did not previously regulate any gene has been co-opted to (presumably) regulate a newly duplicated locus placed nearby. This type of change has been associated with the gain of transcriptional regulation of retrotransposed duplicated genes, which are not copied with any flanking noncoding sequences (Bai et al. 2008; Fablet et al. 2009; Kaessmann et al. 2009).

The division of DA blocks into chained and nonchained sets also allows us to explore whether the association we observed between acceleration and daughter peptides (see above) is driven by nonchained blocks. Focusing only on chained DA blocks, we still find significantly faster evolution in daughter compared with parent loci (FET: $P = 0.002$, OR = 23.69 [95% CI, 2.71–282.99]; see row five in table 3). Assuming co-occurrence on an alignment chain indicates duplication of the noncoding sequence with the gene, this finding suggests that the derived (i.e., daughter) noncoding sequence is more likely to diverge from the

Table 5
Enriched GO Terms in Accelerated DA Blocks Compared with All DA Blocks

GO Term	*P* Value	Term Name
Biological process		
GO:0006805	4.89×10^{-37}	Xenobiotic metabolism
GO:0008202	2.67×10^{-26}	Steroid metabolism
GO:0008152	1.16×10^{-14}	Metabolism
GO:0008210	1.39×10^{-11}	Estrogen metabolism
GO:0007292	3.27×10^{-11}	Female gamete generation
GO:0006915	2.09×10^{-07}	Apoptosis
GO:0050896	1.54×10^{-06}	Response to stimulus
GO:0007608	1.81×10^{-05}	Sensory perception of smell
GO:0050909	2.09×10^{-05}	Sensory perception of taste
GO:0006118	6.03×10^{-05}	Electron transport
GO:0007582	1.46×10^{-04}	Physiological process
GO:0007267	1.78×10^{-04}	Cell–cell signaling
GO:0007165	2.49×10^{-04}	Signal transduction
GO:0007186	2.08×10^{-03}	GPCR protein signaling pathway
GO:0007601	3.66×10^{-02}	Visual perception
Cellular compartment		
GO:0005792	2.11×10^{-36}	Microsome
GO:0005783	5.62×10^{-19}	Endoplasmic reticulum
GO:0016020	3.02×10^{-18}	Membrane
GO:0016021	8.33×10^{-12}	Integral to membrane
GO:0005625	1.06×10^{-09}	Soluble fraction
GO:0005576	1.12×10^{-05}	Extracellular region
Molecular function		
GO:0015020	4.30×10^{-41}	Glucuronosyltransferase activity
GO:0016758	1.47×10^{-35}	Transferase activity, transferring hexosyl groups
GO:0016757	1.62×10^{-27}	Transferase activity, transferring glycosyl groups
GO:0016740	1.23×10^{-16}	Transferase activity
GO:0004497	2.61×10^{-09}	Monooxygenase activity
GO:0005179	8.24×10^{-09}	Hormone activity
GO:0020037	2.14×10^{-07}	Heme binding
GO:0005506	5.90×10^{-06}	Ironion binding
GO:0008527	2.09×10^{-05}	Taste receptor activity
GO:0004984	5.76×10^{-05}	Olfactory receptor activity
GO:0001584	2.42×10^{-04}	Rhodopsin-like receptor activity
GO:0016491	9.27×10^{-04}	Oxidoreductase activity
GO:0004930	1.20×10^{-03}	GPCR activity
GO:0004871	3.28×10^{-03}	Signal transducer activity
GO:0050381	1.59×10^{-02}	Unspecific monooxygenase activity
GO:0004872	3.01×10^{-02}	Receptor activity

ancestral version than the copy that remains at the parent locus.

Correlating Elevated Substitution Rates of Coding and Noncoding Sequence in Duplicated Loci. Han et al. (2009) reported 27 duplicated peptides that exhibit accelerated substitution rates in their coding sequence

Table 6
Depleted GO Terms in Accelerated DA Blocks Compared with All DA Blocks

GO Term	*P* value	Term Name
Biological process		
GO:0006512	3.29×10^{-03}	Ubiquitin cycle
GO:0006810	6.58×10^{-03}	Transport
GO:0006464	3.28×10^{-02}	Protein modification
GO:0006629	4.20×10^{-02}	Lipid metabolism
GO:0006350	4.32×10^{-02}	Transcription
GO:0007283	4.58×10^{-02}	Spermatogenesis
Cellular compartment		
GO:0005622	1.77×10^{-03}	Intracellular
GO:0005737	2.07×10^{-03}	Cytoplasm
GO:0005634	2.72×10^{-03}	Nucleus
GO:0005739	3.52×10^{-02}	Mitochondrion
Molecular function		
GO:0005515	1.41×10^{-05}	Protein binding
GO:0016787	2.31×10^{-03}	Hydrolas activity
GO:0004197	4.76×10^{-02}	Cysteine-type endopeptidase activity

along the hominin lineage. From our filtered set of noncoding blocks, 661 blocks are within 5 kb of one of those accelerated peptides (12 peptides; supplementary table S1, Supplementary Material online). We asked whether these blocks were evolving differently than other DA blocks. Interestingly, noncoding blocks nearby duplicated peptides that have evidence of acceleration in their coding sequences are less likely to be accelerated than other DA blocks (FET: $P = 0.45$, OR = 0.46 [95% CI, 0.055–1.71]; see row six in table 3). Although the power of this test is low, it suggests a possible negative correlation between protein evolution and regulatory evolution in duplicated loci and hints that different categories of duplicated genes may diverge through structural protein changes versus regulatory mechanisms (see Discussion).

Fast-Evolving Blocks Are Enriched in Flanking Regions and Exonic 5′ UTRs. We developed a bioinformatics pipeline to annotate each block with respect to human gene structure (e.g., UTRs, introns, flanking sequences; see Materials and Methods). Using these transcript-based annotations, we investigated whether or not acceleration occurs uniformly across different noncoding genic location categories. Figure 3 shows the log odds score for each annotation category compared with its complement together with a 95% CI. A positive log odds score indicates an enrichment of accelerated blocks in the respective category, whereas a negative score indicates depletion. We find enrichment for acceleration in 5′- and 3′-flanking regions, as well as in exonic 5′ UTRs. In contrast, introns are relatively depleted of accelerated blocks. To further investigate these results, we performed an equivalent analysis using log-linear models. The main advantages of this approach are that 1) we account for all pair-wise correlations between variables

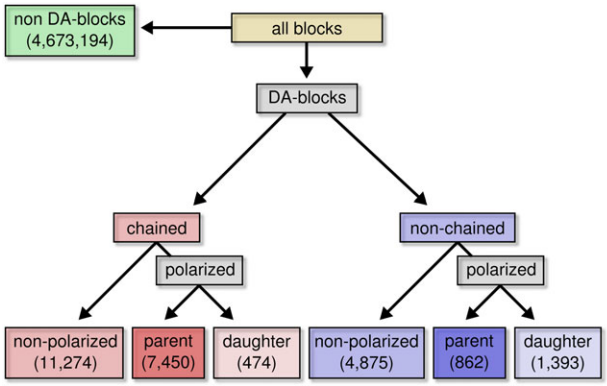

Fig. 2.—Categories of noncoding blocks. First, noncoding blocks are divided into whether they are DA (i.e., within 5 kb of a duplicated gene, DA blocks) or not. DA blocks are then further split into chained and nonchained sets. Additionally, a subset of each of these sets is said to be polarized, that is, the peptides close to the blocks can be classified as either daughter or parent with respect to the duplication event on the hominin lineage. The number of blocks in each category is given in parentheses. Overall, there is an abundance of chained compared with nonchained blocks. Polarized parent blocks tend to be chained, whereas polarized daughter blocks tend to be nonchained. See Materials and Methods section for details regarding definitions.

simultaneously and 2) we correct for possibly confounding factors, such as GC content, alignment length, and alignment depth. This analysis yielded qualitatively similar results to the log odds scores in figure 3 (supplementary results SR1 and supplementary figure S3, Supplementary Material online).

Next, we investigated whether these distributions of accelerated blocks are similar in DA blocks. Although the enrichment patterns are not as clear-cut (supplementary fig. S2, Supplementary Material online), we do find acceleration to be enriched in 3′-flanking regions and weakly in exonic 5′ UTRs but not in 5′-flanking regions.

Patterns of Acceleration Are Not Driven by Changes on the Chimpanzee Lineage. To account for potential false positives introduced by sequencing, assembly, or alignment errors in the 6X chimp genome, we repeated all of the above tests involving acceleration without chimp sequence in the alignments. This filtering did not qualitatively change our results (supplementary results SR2, Supplementary Material online).

Turnover of TFBS

To provide a complementary and functionally oriented analysis of divergence, we assessed the predicted impact of substitutions on TF-binding affinity in our data set of ~4.7 million noncoding blocks within 5 kb of a human gene. Specifically, we scored human and macaque sequences using motifs for 11 families of TFs from the JASPAR database (Sandelin and Wasserman 2004; Wasserman and Sandelin 2004) to identify predicted binding sites in each species. Using a novel approach (see Materials and Methods), we assessed the statistical significance of total binding site gain and loss ("TFBS turnover") in each block. At an FDR of 10%, we identified 13,067 blocks (~0.3%) with significant TFBS turnover between human and macaque. We refer to these as "turnover blocks." Using FETs and ORs, we examined the association between TFBS turnover and 1) accelerated substitution rates (significant LRTs) and 2) duplication status.

Accelerated Blocks Exhibit High TFBS Turnover. First, we asked whether TFBS turnover occurs at a higher rate in blocks that show accelerated substitution rates in the Hominini. We find that accelerated blocks have much higher rates of TFBS turnover compared with nonaccelerated blocks (FET: $P < 1 \times 10^{-15}$, OR = 8.04 [95% CI, 6.39–10]; see row seven in table 3). To some extent such a correlation is expected because accelerated blocks have, on average, higher substitution rates than nonaccelerated blocks (supplementary fig. S5, Supplementary Material online). Higher substitution rates, in turn, mean a higher probability of destroying or creating a TFBS. On the other hand, higher substitution rates are not sufficient to explain TFBS turnover. This is illustrated by the fact that the vast majority of turnover blocks (12,984) are not accelerated. Nevertheless, associations between TFBS turnover and accelerated blocks are to some degree inherent and FET P values have to be taken with a large grain of salt. Keeping the above in mind, we find that DA blocks are enriched for TFBS turnover compared with non-DA blocks ($P = 0.03$, OR = 1.26 [95% CI, 1.02–1.55]; see row eight in table 3) and that the association of acceleration and turnover remains significant if we focus on DA blocks exclusively (FET: $P = 3.12 \times 10^{-5}$, OR = 11.00 [95% CI, 3.87–25.46]; see row nine in table 3). In fact, among accelerated blocks, DA blocks have higher odds of TFBS turnover than non-DA blocks, although this trend is not significant (FET: $P = 0.27$, OR = 1.68 [95% CI, 0.59–3.90]; see row ten in table 3). We note that the reported associations are largely descriptive. More sophisticated analyses are needed to unambiguously disentangle the correlation between acceleration and TFBS turnover arising purely because of accelerated substitution rates from a biological signal.

TFBS turnover DA blocks are enriched in many of the same GO terms as accelerated DA blocks, but lack the GPCR and olfactory receptor–related terms (table 7). Also, turnover blocks are enriched in RNA-related GO terms (binding and transport), as well as in terms related to regulation of development (including embryonic and mammary gland). The only GO term we find depleted in TFBS turnover DA blocks is signal transduction.

Overall, our results concerning TFBS turnover are in line with our findings on accelerated substitutions. Our data

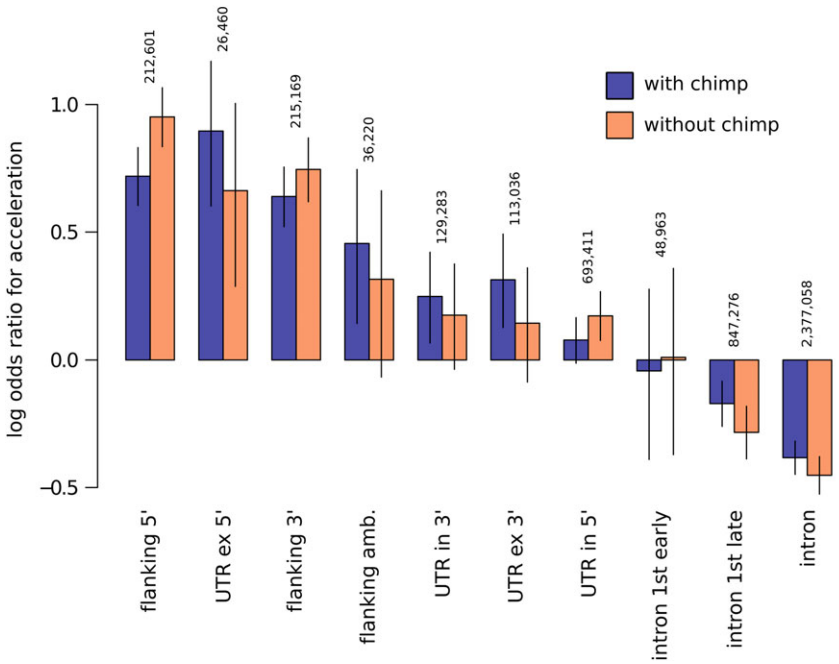

Fig. 3.—Enrichment and depletion for acceleration in different genic locations. The panel shows log ORs for acceleration for each genic location category (compared with its complement). A positive log OR indicates enrichment for accelerated blocks in that category, whereas a negative log OR indicates depletion of acceleration. Bars correspond to 95% CIs. The number of blocks in each category is given above each pair of bars. Analyses excluding the chimp genome sequence are shown in orange. Nonambiguous 5′- and 3′-flanking regions and 5′ exonic UTRs are enriched for accelerated blocks, whereas intronic sequences are depleted for acceleration.

support the hypothesis that higher rates of substitution result in more binding site turnover, potentially contributing to changes in the transcriptional regulation of nearby genes.

Fast-Evolving Noncoding Sequences Are Associated with Pregnancy-Related Genes

Because GO enrichment analysis has a known set of limitations (Khatri and Drăghici 2005; Rhee et al. 2008; Taher and Ovcharenko 2009), we manually analyzed the genes near accelerated DA blocks (table 4) with respect to their functions as annotated in public databases and the literature. We find three PRAME genes and five olfactory receptor genes. Both families experienced positive selection on the protein level in the human lineage (Birtle et al. 2005; Han et al. 2009). Additionally, table 4 contains three genes from the UDP glycosyltransferase superfamily, which is known to exhibit copy number variations in humans (Guillemette et al. 2010). Our results suggest that functional changes in these families may have occurred through divergence in both protein structure and gene regulation. We also find genes related to immunity and to metabolism, both of which are functional categories that have been identified in the context of positive selection and duplicated genes (Haygood et al. 2007; Studer et al. 2008).

Additionally, we find two chorionic gonadatropins (CGs: *CGB5* and *CGB7*) and a chorionic somatomammotropin (*CSHL1*). Both CGs and *CSHL1* are expressed in the placenta and play a crucial role in pregnancy. Motivated by this observation, we asked whether other genes in table 4 might also be associated with pregnancy by looking for placental expression and/or pregnancy-related functional annotation.

CGs regulate endometrial functions by influencing progesterone (Szmidt et al. 2008), a hormone that is catalyzed to its inactive form by another gene in table 4, *AKR1C1*, which encodes an aldo–keto reductase. *AKR1C1* utilizes NAD and/or NAD(P)H as cofactors. *NAMPT* (also in table 4) is an NAD(P) biosynthetic enzyme (Garten et al. 2008); NAD(P)H is active in the placenta, and there is evidence that it is a modulator of antioxidant stress response in early pregnancy (Raijmakers et al. 2006). Another gene we find, *CYP4A22*, a cytochrome P450 superfamily member, is part of the *PPAR-gamma* signaling pathway (Kanehisa et al. 2008). *PPAR-gamma*, in turn, is essential for placental development (Fournier et al. 2008). *UGT2B15* and *UGT2B28* (both in table 4) are part of the androgen and estrogen metabolism pathway, and estrogen is prominently involved in regulation of the menstrual cycle and pregnancy. Also, copy number variations in *UGT2B28* may influence fetal development and gestation length (Guillemette et al. 2010).

In addition, there is an abundance of potentially pregnancy-related terms among GO terms enriched for accelerated DA blocks (e.g., female gamete generation, estrogen metabolism, hormone activity; table 5) and TFBS turnover

Table 7
GO terms Enriched in TFBS Turnover DA Blocks Compared with All DA Blocks

GO Term	*P* Value	Term Name
Biological process		
GO:0006805	2.30×10^{-04}	Xenobiotic metabolism
GO:0008202	4.20×10^{-04}	Steroid metabolism
GO:0006406	2.80×10^{-02}	mRNA export from nucleus
GO:0050658	2.80×10^{-02}	RNA transport
GO:0001569	4.62×10^{-02}	Patterning of blood vessels
GO:0001709	4.62×10^{-02}	Cell fate determination
GO:0001763	4.62×10^{-02}	Morphogenesis of a branching structure
GO:0007219	4.62×10^{-02}	Notch signaling pathway
GO:0009790	4.62×10^{-02}	Embryonic development
GO:0030097	4.62×10^{-02}	Hemopoiesis
GO:0030879	4.62×10^{-02}	Mammary gland development
GO:0045596	4.62×10^{-02}	Negative regulation of cell differentiation
GO:0045602	4.62×10^{-02}	Negative regulation of endothelial cell differentiation
GO:0050793	4.62×10^{-02}	Regulation of development
Cellular compartment		
GO:0005792	6.01×10^{-04}	Microsome
GO:0016020	7.52×10^{-03}	Membrane
GO:0005842	7.97×10^{-03}	Cytosolic large ribosomal subunit (sensu Eukaryota)
GO:0042272	2.80×10^{-02}	Nuclear RNA export factor complex
GO:0016021	3.28×10^{-02}	Integral to membrane
GO:0005622	3.38×10^{-02}	Intracellular
Molecular function		
GO:0015020	4.74×10^{-05}	Glucuronosyltransferase activity
GO:0016758	4.09×10^{-04}	Transferase activity, transferring hexosyl groups
GO:0005488	2.98×10^{-03}	Binding
GO:0016740	7.96×10^{-03}	Transferase activity
GO:0016757	8.51×10^{-03}	Transferase activity, transferring glycosyl groups
GO:0004759	1.12×10^{-02}	Serine esterase activity
GO:0016290	1.12×10^{-02}	Palmitoyl-CoA hydrolase activity
GO:0004497	1.49×10^{-02}	Monooxygenase activity
GO:0019843	2.59×10^{-02}	rRNA binding
GO:0005506	3.97×10^{-02}	Iron ion binding

DA blocks (embryonic development, mammary development, and cell fate determination; see table 7). Together, these findings constitute compelling if circumstantial evidence that noncoding sequence evolution near duplicated loci played a role in the lineage-specific evolution of pregnancy and reproduction.

Discussion

We conducted a high-resolution genome-wide scan for accelerated substitution rates in noncoding sequences within 5 kb of all human genes. Genes belonging to families that expanded through gene duplication in the hominin lineage show enrichment for accelerated evolution in associated noncoding sequences. Noncoding elements that most likely duplicated along with the coding sequence of the associated gene (i.e., chained blocks of daughter genes) are particularly enriched for acceleration. Flanking sequence and exonic 5′ UTRs are enriched for elevated substitution rates, especially compared with introns, which are relatively depleted of accelerated elements. Rapid evolution of 5′ UTR elements could affect transcription and is consistent with a recent study that correlates changes in the TSSs of recently duplicated genes with expression changes (Park and Makova 2009). Because 5′ UTR and flanking regions are enriched for regulatory elements, their particularly rapid divergence suggests the possible action of positive selection to modify the expression patterns of duplicate genes. However, we emphasize that our analyses cannot distinguish positive selection from neutral mutational processes that might affect substitution rates in a lineage-specific manner.

To further pursue the link between noncoding sequence evolution and gene expression, we investigated noncoding elements associated with human genes for the effects of substitutions on predicted TFBS. We found that duplicated loci have more noncoding elements in which sequence differences between human and macaque are predicted to affect TF binding. Together, our findings are consistent with the hypothesis that modification of the regulation of duplicated genes is an important mechanism for the evolution of hominin-specific traits.

We took several precautions to control for false positives and ensure the quality of our data. Because duplicated noncoding regions are particularly difficult to align and incorrect alignment can lead to false inference about evolutionary events, conservative quality filtering of sequence alignments was an essential component of our analysis. It is nonetheless possible that alignment errors contributed to our estimates of substitution rates in some regions. However, we do not expect that such bias would lead to an inference of accelerated evolution in the Hominini in particular. For instance, we performed our analyses twice, once with the chimp sequence and once without it. Although it could be hypothesized that the lower coverage shotgun sequenced chimp genome would leads to false signals of acceleration, we find qualitative agreement between the two analyses (supplementary results SR2, Supplementary Material online). Also, although we cannot rule out that the enrichment of accelerated blocks in exonic 5′ UTRs is due to hypervariable-methylated CpG dinucleotides decaying to CA or TG, we find that accelerated exonic 5′ UTRs on average have roughly the same GC content as nonaccelerated exonic 5′ UTRS (58.2% accelerated vs. 58.0% nonaccelerated).

We focus on the Hominini because previous studies found accelerated gene duplication, sometimes accompanied by amino acid divergence, in the ape lineage

(Hahn et al. 2007; Marques-Bonet et al. 2009). From the point of view of understanding human evolution, many biologically important human traits are shared with chimpanzee and other great apes. Hence, the fast-evolving noncoding sequences that we identified are candidates for understanding the genetic basis of human-specific biology. Furthermore, by studying evolution over tens of millions of years, we have more power to detect changes in substitution rates than we would if we focused on events that took place in the ~6 million years since the human–chimp ancestor. Our approach could of course be used to study noncoding sequence evolution in loci that duplicated on the human lineage or other lineages of interest. Further studies will determine if a propensity toward accelerated evolution in noncoding sequences is a universal characteristic of duplicated loci.

Several previous publications have focused on predicted functional noncoding sequences with accelerated substitution rates in the human branch (Pollard, Salama, King et al. 2006; Pollard, Salama, Lambert et al. 2006; Prabhakar et al. 2006; Bird et al. 2007; Haygood et al. 2007). In this study, we extended that approach in two ways. First, we expanded the set of candidate regions by considering all noncoding sequence in the vicinity of all known genes, not just deeply conserved elements. Noncoding sequences nearby genes are known to harbor regulatory elements, and sequence changes in these regions have the potential to modify the expression of the associated gene. Second, by including information about gene duplication, our method aims to identify regions that are able to take on new functions by two complementary evolutionary mechanisms: gene expression divergence and protein sequence divergence. On one hand, we find some agreement between these two levels of evolution in the sense that both appear to occur more often in the daughter copies of recently duplicated genomic loci. Interestingly, however, our data show a nonsignificant negative correlation between accelerated rates of protein and regulatory sequence evolution. This observation suggests the hypothesis that relatively disjoint subsets of proteins have evolved at the regulatory versus protein-coding level in the hominin lineage. But, exceptions to this rule are known (e.g., CGB, see Introduction).

We performed GO term enrichment analysis of genes with fast-evolving regulatory regions. We note that enrichment analysis of noncoding sequences using GO is prone to ascertainment bias (Taher and Ovcharenko 2009). In this study, we account for ascertainment bias by performing enrichment tests on the block level, mapping GO terms from genes to the associated noncoding sequences and performing tests on the set of noncoding blocks. Using this approach, we highlight functional categories, such as reproduction, host defense, and metabolism. Many of these terms have been mentioned before in the context of positive selection at the protein level, but our analysis also highlights several processes and pathways that have not been emphasized in studies of single-copy genes.

For instance, some of the genes we identified with hominin-specific acceleration in their regulatory regions are connected to placentation. Although there are multiple differences between human and macaque pregnancies (de Rijk and van Esch 2008), it has been argued that some of these differences are not very large, especially when factors such as body size are taken into account (Martin 2007). However, one particularly notable difference between the pregnancies of humans and other primates involves the formation of the trophoblastic shell by cytotrophoblasts. In macaques and baboons, the shell is continuous and sharply delineated from the endometrium. In humans, on the other hand, extravillous trophoblast cells invade the uterine stroma (Carter 2007). CG is necessary for the invasion of cytotrophoblasts into the endometrium during embryo implantation (Henke and Gromoll 2008). Interestingly, CG genes are highlighted by our genomic approach, making them and other genes in our list excellent targets for functional studies of human–macaque differences in pregnancy. Unfortunately, it is challenging to contrast commonalities of human and chimp pregnancies to those of macaques, as placentation in chimpanzees remains poorly studied (Carter 2007).

This is the first genome-wide study to address the question of whether genetic divergence in noncoding sequences might contribute to functional divergence of duplicated genes in the hominin lineage. Consistent with the hypotheses that 1) divergence between closely related species occurs through changes in gene regulation and 2) duplicated regions are enriched for genetic and functional divergence, we find a strong propensity for rapid sequence evolution in noncoding elements near duplicated genes. We quantify this rapid evolution in terms of substitution rates and predicted TFBS turnover. Using both metrics, we find an excess of fast-evolving elements associated with duplicated genes. Together with evidence of accelerated evolution in the coding sequence of young duplicates (Han et al. 2009), our results support the view that two sources of genetic variation—structural rearrangements and point mutations—synergistically contribute to the evolution of new traits.

Supplementary Material

Supplementary results SR1–SR5, figures S1–S5, and tables S1–S3 are available at *Genome Biology and Evolution* online (http://www.oxfordjournals.org/our_journals/gbe/).

Acknowledgments

We thank Genevieve Erwin for feedback throughout the project; David Williamson for suggestions about identification of binding site turnover; Joseph Shieh for a helpful

discussion about pregnancy-related genes; Ralph Haygood for code and ideas about transcript clustering and annotation; Evan Eichler for suggestions regarding alignment quality; Charlyn Suarez for contributing code to an early version of the LRT analysis; and Mira Han for sharing her data. K.S.P. and D.K. were supported by National Institutes of Health (NIGMS) grant GM82901. M.W.H. was supported by the National Science Foundation grant DBI-0543586.

Literature Cited

Ashburner M, et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 25:25–29.

Bai Y, Casola C, Betrán E. 2008. Evolutionary origin of regulatory regions of retrogenes in Drosophila. BMC Genomics. 9:241.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Statist Soc Ser B Meth. 57:289–300.

Bird C, et al. 2007. Fast-evolving noncoding sequences in the human genome. Genome Biol. 8:R118.

Birtle Z, Goodstadt L, Ponting C. 2005. Duplication and positive selection among hominin-specific PRAME genes. BMC Genomics. 6:120.

Blanchette M, et al. 2004. Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res. 14:708–715.

Carroll SB. 2005. Evolution at two levels: on genes and form. PLoS Biol. 3:e245.

Carter AM. 2007. Animal models of human placentation—a review. Placenta. 28:S41–S47.

Cheng Z, et al. 2005. A genome-wide comparison of recent chimpanzee and human segmental duplications. Nature. 437:88–93.

De S, Teichmann SA, Babu MM. 2009. The impact of genomic neighborhood on the evolution of human and chimpanzee transcriptome. Genome Res. 19:785–794.

de Rijk EPCT, van Esch E. 2008. The macaque placenta: a mini-review. Toxicol Pathol. 36:108S–118S. [cited 2008 October 20]. Available from http://tpx.sagepub.com/cgi/content/abstract/0192623308326095v1.

Demuth JP, De Bie T, Stajich JE, Cristianin N, Hahn MW. 2006. The evolution of mammalian gene families. PLoS ONE. 1:e85.

Dumas L, et al. 2007. Gene copy number variation spanning 60 million years of human and primate evolution. Genome Res. 17:1266–1277.

Durand D, Halldórsson BV, Vernot B. 2006. A hybrid micro-macroevolutionary approach to gene tree reconstruction. J Comput Biol. 13:320–335.

Fablet M, Bueno M, Potrzebowski L, Kaessmann H. 2009. Evolutionary origin and functions of retrogene introns. Mol Biol Evol. 26:2147–2156.

Flicek P, et al. 2008. Ensembl 2008. Nucleic Acids Res. 36:D707–D714.

Fortna A, et al. 2004. Lineage-specific gene duplication and loss in human and great ape evolution. PLoS Biol. 2:e207.

Fournier T, Thérond P, Handschuh K, Tsatsaris V, Evain-Brion D. 2008. PPARgamma and early human placental development. Curr Med Chem. 15:3011–3024.

Garten A, Petzold S, Körner A, Imai SI, Kiess W. 2008. Nampt: linking NAD biology, metabolism and cancer. Trends Endocrinol Metab. 20(3):130–138. [cited 2009 April 1]. Available from: http://dx.doi.org/10.1016/j.tem.2008.10.004.

Guillemette C, Lévesque E, Harvey M, Bellemare J, Menard V. 2010. UGT genomic diversity: beyond gene duplication. Drug Metab Rev. 42(1):24–44.

Hahn MW, Demuth JP, Han SG. 2007. Accelerated rate of gene gain and loss in primates. Genetics. 177:1941–1949.

Han MV, Demuth JP, McGrath CL, Casola C, Hahn MW. 2009. Adaptive evolution of young gene duplicates in mammals. Genome Res. 19:859–867.

Han MV, Hahn MW. 2009. Identifying parent-daughter relationships among duplicated genes. Proc Pac Symp Biocomput. 14:114–125.

Haygood R, Fedrigo O, Hanson B, Yokoyama KD, Wray GA. 2007. Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution. Nat Genet. 1140–1144.

Henke A, Gromoll J. 2008. New insights into the evolution of chorionic gonadotrophin. Mol Cell Endocrinol. 291:11–19.

Huttley GA, Wakefield MJ, Easteal S. 2007. Rates of genome evolution and branching order from whole genome analysis. Mol Biol Evol. 24:1722–1730.

Kaessmann H, Vinckenbosch N, Long M. 2009. RNA-based gene duplication: mechanistic and evolutionary insights. Nat Rev Genet. 10:19–31.

Kanehisa M, et al. 2008. KEGG for linking genomes to life and the environment. Nucleic Acids Res. 36:D480–D484.

Kent JW, Baertsch R, Hinrichs A, Miller W, Haussler D. 2003. Evolution's cauldron: duplication, deletion and rearrangement in the mouse and human genomes. Proc Natl Acad Sci U S A. 100:11484–11489.

Khatri P, Drăghici S. 2005. Ontological analysis of gene expression data: current tools, limitations and open problems. Bioinformatics. 21:3587–3595.

Kim YS, Pritchard SK. 2007. Adaptive evolution of conserved noncoding elements in mammals. PLoS Genetics. 3:e147.

King MC, Wilson AC. 1975. Evolution at two levels in humans and chimpanzees. Science. 188:107–116.

Kondrashov F, Rogozin I, Wolf Y, Koonin E. 2002. Selection in the evolution of gene duplications. Genome Biol. 3:research0008.

Levine M, Tjian R. 2003. Transcription regulation and animal diversity. Nature. 424:147–151.

Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. Science. 290:1151–1155.

Lynch M, Conery JS. 2003. The evolutionary demography of duplicate genes. J Struct Funct Genomics. 3:35–44.

Marques AC, Dupanloup I, Vinckenbosch N, Reymond A, Kaessmann H. 2005. Emergence of young human genes after a burst of retroposition in primates. PLoS Biol. 3:e357.

Marques-Bonet T, Cheng Z, She X, Eichler EE, Navarro A. 2008. The genomic distribution of intraspecific and interspecific sequence divergence of human segmental duplications relative to human/chimpanzee chromosomal rearrangements. BMC Genomics. 9:384.

Marques-Bonet T, et al. 2009. A burst of segmental duplications in the genome of the African great ape ancestor. Nature. 457:877–881.

Martin RD. 2007. The evolution of human reproduction: a primatological perspective. Am J Phys Anthropol. (Suppl 45):59–84.

Maston GA, Ruvolo M. 2002. Chorionic gonadotropin has a recent origin within primates and an evolutionary history of selection. Mol Biol Evol. 19:320–335.

McGrath CL, Casola C, Hahn MW. 2009. Minimal effect of ectopic gene conversion among recent duplicates in four mammalian genomes. Genetics. 182:615–622.

Ohno S. 1999. Gene duplication and the uniqueness of vertebrate genomes circa 1970–1999. Semin Cell Dev Biol. 10:517–522.

Park C, Makova K. 2009. Coding region structural heterogeneity and turnover of transcription start sites contribute to divergence in expression between duplicate genes. Genome Biol. 10:R10.

Perry GH, et al. 2008. Copy number variation and evolution in humans and chimpanzees. Genome Res. 18:1698–1710.
Pollard KS, Hubisz M, Siepel A. 2009. Detection of non-neutral substitution rates on mammalian phylogenies. PLoS Genet. 2(10): e168.
Pollard KS, et al. 2006. Forces shaping the fastest evolving regions in the human genome. PLoS Genetics. 2:e168.
Pollard KS, et al. 2006. An RNA gene expressed during cortical development evolved rapidly in humans. Nature. 443:167–172.
Prabhakar S, Noonan JP, Pääbo S, Rubin EM. 2006. Accelerated evolution of conserved noncoding sequences in humans. Science. 314:786.
Rahmann S, Müller T, Vingron M. 2003. On the power of profiles for transcription factor binding site detection. Stat Appl Genet Mol Biol. 2:Article 7.
Raijmakers MTM, et al. 2006. Placental NAD(P)H oxidase mediated superoxide generation in early pregnancy. Placenta. 27:158–163.
Rhee SY, Wood V, Dolinski K, Draghici S. 2008. Use and misuse of the gene ontology annotations. Nat Rev Genet. 9:509–515.
Sandelin A, Wasserman WW. 2004. Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. J Mol Biol. 23:207–215.
Self SG, Liang K. 1987. Asymptotic properties of maximum likelihood, estimators and likelihood ratio tests under nonstandard conditions. J Am Stat Assoc. 82:605–610.
She X, et al. 2006. A preliminary comparative analysis of primate segmental duplications shows elevated substitution rates and a great-ape expansion of intrachromosomal duplications. Genome Res. 16:576–583.
Shiu SH, Byrnes JK, Pan R, Zhang P, Li WH. 2006. Role of positive selection in the retention of duplicate genes in mammalian genomes. Proc Natl Acad Sci U S A. 103:2232–2236.
Studer RA, Penel S, Duret L, Robinson-Rechavi M. 2008. Pervasive positive selection on duplicated and nonduplicated vertebrate protein coding genes. Genome Res. 18:1393–1402.
Szmidt M, Sysa P, Bartłomiej JB, Niemiec T. 2008. Chorionic gonadotropin as the key factor for embryo implantation. Ginekol. Polska. 79:692–696.
Taher L, Ovcharenko I. 2009. Variable locus length in the human genome leads to ascertainment bias in functional inference for noncoding elements. Bioinformatics. Available from http://bioinformatics.oxfordjournals.org/cgi/content/abstract/btp043v1.
The Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature. 437:69–87.
Wasserman WW, Sandelin A. 2004. Applied bioinformatics for the identification of regulatory elements. Nat Rev Genet. 5:276–287.
Wilson AC, Sarich VM, Maxson LR. 1974. The importance of gene rearrangement in evolution: evidence from studies on rates of chromosomal, protein, and anatomical evolution. Proc Natl Acad Sci U S A. 71:3028–3030.
Wu CI, Li WH. 1985. Evidence for higher rates of nucleotide substitution in rodents than in man. Proc Natl Acad Sci U S A. 82:1741–1745.
Yang Z, Goldman N, Friday A. 1994. Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. Mol Biol Evol. 11:316–324.
Zhang J. 2003. Evolution by gene duplication: an update. Trends Ecol Evol. 18:292–298.
Zhang P, Gu Z, Li WH. 2003. Different evolutionary patterns between young duplicate genes in the human genome. Genome Biol. 4:R56.
Zhu J, et al. 2007. Comparative genomics search for losses of long-established genes on the human lineage. PLoS Comput Biol. 3:e247.

Associate editor: Yoshihito Niimura