28-Way vertebrate alignment and conservation track in the UCSC Genome Browser

Webb Miller,[1,11] Kate Rosenbloom,[2] Ross C. Hardison,[1] Minmei Hou,[1] James Taylor,[3] Brian Raney,[2] Richard Burhans,[1] David C. King,[1] Robert Baertsch,[2] Daniel Blankenberg,[1] Sergei L. Kosakovsky Pond,[4] Anton Nekrutenko,[1] Belinda Giardine,[1] Robert S. Harris,[1] Svitlana Tyekucheva,[1] Mark Diekhans,[2] Thomas H. Pringle,[5] William J. Murphy,[6] Arthur Lesk,[1] George M. Weinstock,[7] Kerstin Lindblad-Toh,[8] Richard A. Gibbs,[7] Eric S. Lander,[8] Adam Siepel,[9] David Haussler,[2,10] and W. James Kent[2]

[1]Center for Comparative Genomics and Bioinformatics, Penn State University, University Park, Pennsylvania 16802, USA; [2]Center for Biomolecular Science and Engineering, University of California, Santa Cruz, California 95064, USA; [3]Courant Institute, New York University, New York, New York 10012, USA; [4]Antiviral Research Center, University of California at San Diego, San Diego, California 92103, USA; [5]Sperling Foundation, Eugene, Oregon 97405, USA; [6]Department of Veterinary Integrative Biosciences, Texas A&M University, College Station, Texas 77843, USA; [7]Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas 77030, USA; [8]Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA; [9]Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, New York 14853, USA; [10]Howard Hughes Medical Institute, Santa Cruz, California 95060, USA

This article describes a set of alignments of 28 vertebrate genome sequences that is provided by the UCSC Genome Browser. The alignments can be viewed on the Human Genome Browser (March 2006 assembly) at http://genome.ucsc.edu, downloaded in bulk by anonymous FTP from http://hgdownload.cse.ucsc.edu/goldenPath/hg18/multiz28way, or analyzed with the Galaxy server at http://g2.bx.psu.edu. This article illustrates the power of this resource for exploring vertebrate and mammalian evolution, using three examples. First, we present several vignettes involving insertions and deletions within protein-coding regions, including a look at some human-specific indels. Then we study the extent to which start codons and stop codons in the human sequence are conserved in other species, showing that start codons are in general more poorly conserved than stop codons. Finally, an investigation of the phylogenetic depth of conservation for several classes of functional elements in the human genome reveals striking differences in the rates and modes of decay in alignability. Each functional class has a distinctive period of stringent constraint, followed by decays that allow (for the case of regulatory regions) or reject (for coding regions and ultraconserved elements) insertions and deletions.

[Supplemental material is available online at www.genome.org.]

The National Human Genome Research Institute is funding a number of vertebrate genome sequencing projects, primarily at the Broad Institute of MIT (Massachusetts Institute of Technology) and Harvard, the Human Genome Sequencing Center at the Baylor College of Medicine, and the Genome Sequencing Center at Washington University. Additional genome sequence data are being provided by other organizations, including the Sanger Center, the Department of Energy's (DOE's) Joint Genome Institute, and the National Institute of Genetics in Japan. (Supplemental Table S1 identifies the producers of the individual genome sequences.) The sequences are being immediately and freely released to the public to allow scientists to use the genomic information in their own research. (The scientific community is expected to postpone publications of large-scale analyses of the data until consortia organized by the data producers can publish their initial large-scale analyses, in the spirit of the Ft. Lauderdale agreement; http://www.wellcome.ac.uk/doc_wtd003208.html.)

While a few sophisticated users can mine the raw sequences, many prefer to wait until the data have been assembled into chromosomes, annotated, and aligned to other sequences. The sequencing centers work closely with several groups, including the UCSC Browser team, Ensembl, and the National Center for Biotechnology Information, who maximize data availability and in some cases align the genome sequences to each other so as to facilitate their direct comparison.

Alignments of genomic sequences have long been used as guides to help locate certain kinds of functional noncoding regions (e.g., Hardison 2000), and have more recently been used for finding protein-coding genes (Siepel and Haussler 2004; Gross and Brent 2006) and noncoding RNA genes (Pedersen et al. 2006). Indeed, the primary justification for much of the recent effort to sequence mammalian genomes was to more reliably identify functional elements via sequence alignments (Margulies et al. 2005). The alignments can also reveal similarities and dif-

[11]Corresponding author.
E-mail webb@bx.psu.edu; fax (814) 865-3176.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.6761107.

ferences between the sequences of humans and those of disease-model species, and thereby enhance the effectiveness of the model species for experiments aimed at improving human health (Rhesus Macaque Genome Sequencing and Analysis Consortium 2007). Moreover, the alignments provide critical data for determining the course of evolution and for the computational reconstruction of ancestral genome sequences.

However, producing whole-genome alignments requires expertise and computational resources that are not easily available within a typical research group. Moreover, reproducibility of published results is facilitated if the alignments used by one group can be utilized directly by others, providing further impetus for creating a reliable, comprehensive, and up-to-date set of freely available alignments.

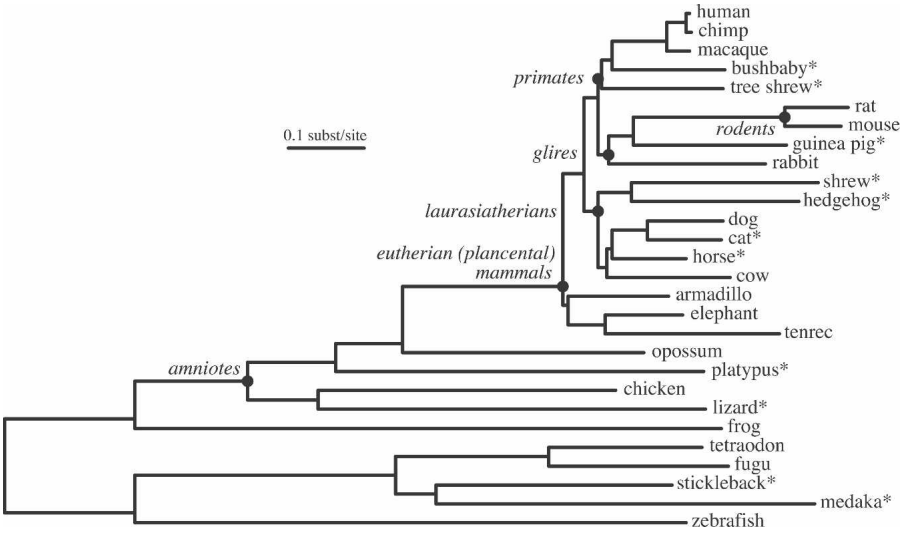

Figure 1. A tree indicating assumed evolutionary relationships among the sequences in the 28-way alignment. Branch lengths are proportional to average number of substitutions per site. Species not previously available in our whole-genome alignments are indicated with an asterisk. Filled circles indicate named clades, such as amniotes and eutherians, that are mentioned in the text. A tree labeled with branch lengths is given in the Supplemental material.

To meet these needs, the UCSC Genome Bioinformatics group produces browsers for a number of species, each containing genome-wide multiple alignments with conservation scoring ("conservation tracks") in addition to other types of annotations. Alignments at the Human Genome Browser are constructed using all vertebrate species that have their own UCSC browser, plus a number of additional mammalian species for which low-coverage sequence is available (Margulies et al. 2005). For instance, although there is currently no browser devoted to the rabbit genome, the alignments at the Human Genome Browser include rabbit sequences.

We have recently performed a major update to the Conservation track on the latest UCSC Human Genome Browser. The purpose of this article is to describe the new alignments and to illustrate their use by exploring a few interesting biological issues. This article does not provide an overall analysis of the sequence data sets themselves, which will be the subject of various consortium papers in preparation.

Results and Discussion

The 28-way alignment

In April 2007, the UCSC Conservation track for the NCBI Build 36 (UCSC hg18) human genome sequence was updated from a 17-way to a 28-way vertebrate alignment (see Methods). The expanded conservation track now includes 11 new species and incorporates updated sequence assemblies for six of the old ones. Figure 1 shows the evolutionary relationships among the sequences that were assumed for the purpose of aligning them (see Methods). Table 1 provides data about the extent to which the individual sequences are represented in the alignment. Note that even for vertebrate species that are quite distant from humans and for mammals with only twofold shotgun sequence coverage, at least 79% of human protein-coding intervals are aligned, though the overall fraction of the human sequence that aligns is <2% for some species.

The 28 genome sequences (Table 1) form a heterogeneous mix. They include two finished sequences, human and mouse, with an estimated coverage of >99% of the euchromatin and an error rate of one in 100,000 (International Human Genome Sequencing Consortium 2004). In addition, there are 16 high-quality draft sequences based on whole-genome shotgun assemblies with coverage from 5.1× to 8.5×. Finally, there are 10 whole-genome shotgun assemblies with coverage ~2×. In theory, according to the Lander and Waterman (1988) formula, a 2× assembly should include 87.5% of the bases in the genome, and a 5× assembly 99.4%. In practice the base inclusion is somewhat less, due to cloning bias and related issues, and this bias is difficult to estimate precisely. Table 1 also shows the percentage of the human genome that aligns to the various other genomes. Because of coverage issues, the true total aligning portions once the genomes are finished will be likely be 10%–15% above what we currently see for the 2× assemblies, and 1%–4% above the current values for the higher coverage draft sequences.

Application I: Three vignettes about insertions and deletions in protein-coding regions

To illustrate applications of the 28-way alignment for exploring biological issues and hypotheses, we look briefly at three uses of data about insertions and deletions (collectively called *indels*) in protein-coding regions. First, we test the hypothesis that coding indels have accumulated at a uniform rate during evolution of placental mammals. Second, we look for human-specific coding indels with phenotypic consequences. Finally, we consider the hypothesis that coding deletions are more likely to have adverse phenotypic consequences if the affected amino acids are well conserved over evolutionary time. These brief explorations illustrate the utility of the 28-way alignment for addressing a wide range of interesting questions, while highlighting several Browser features.

Do coding indels accumulate at a uniform rate?

Coding indels have been used for inferring phylogenetic relationships among species. For instance, Poux et al. (2002) identified a

Table 1. Species-by-species information

Scientific name	Common name	Seq. cov.	Filter	Total align	Start aligns	Stop aligns	Total %ID	Start %ID	Stop %ID	ORF cover
Homo sapiens	Human	Fin	n/a	100%	100%	100%	100%	100%	100%	100%
Pan troglodytes	Chimp	6.0×	Syn	93.9%	94.0%	97.1%	98.7%	99.0%	99.1%	96.58%
Macaca mulatta	Rhesus	5.1×	Syn	85.1%	93.5%	95.5%	93.7%	97.6%	97.3%	96.31%
Otolemur garnettii	Bushbaby	2.0×	Rec	44.3%	67.3%	78.2%	77.6%	92.2%	90.3%	79.10%
Tupaia belangeri	Tree shrew	1.5×	Rec	37.7%	63.8%	75.0%	75.5%	91.4%	89.9%	81.47%
Rattus norvegicus	Rat	7.0×	Syn	35.7%	94.5%	95.1%	66.6%	91.8%	88.2%	94.47%
Mus musculus	Mouse	Fin	Syn	37.6%	97.7%	96.8%	66.8%	92.2%	88.3%	95.36%
Cavia porcellus	Guinea pig	2.0×	Rec	30.9%	60.3%	69.5%	70.0%	89.4%	86.5%	80.12%
Oryctolagus cuniculus	Rabbit	2.0×	Rec	34.0%	67.8%	73.3%	72.3%	91.2%	87.7%	83.43%
Sorex araneus	Shrew	2.0×	Rec	20.7%	55.0%	67.5%	69.0%	90.0%	87.1%	85.23%
Erinaceus europaeus	Hedgehog	2.0×	Rec	20.1%	65.6%	74.0%	69.7%	91.4%	88.3%	83.82%
Canis familiaris	Dog	7.6×	Syn	55.4%	89.4%	97.1%	74.3%	92.6%	89.6%	95.18%
Felis catus	Cat	2.0×	Rec	36.1%	63.0%	76.1%	74.7%	91.0%	89.3%	86.76%
Equus caballus	Horse	6.8×	Syn	58.8%	85.9%	96.8%	77.0%	91.7%	91.1%	92.70%
Bos taurus	Cow	7.1×	Syn	48.2%	94.4%	96.6%	73.8%	93.2%	89.7%	94.78%
Dasypus novemcinctus	Armadillo	2.0×	Rec	32.5%	61.0%	67.9%	73.3%	89.4%	88.6%	82.33%
Loxodonta africana	Elephant	2.0×	Rec	34.1%	65.3%	71.0%	74.5%	92.3%	89.8%	81.59%
Echinops telfairi	Tenrec	2.0×	Rec	24.4%	67.7%	75.4%	69.7%	90.6%	86.4%	81.09%
Monodelphis domestica	Opossum	6.5×	Syn	11.1%	81.8%	88.4%	64.2%	88.8%	86.0%	91.43%
Ornithorhynchus anatinus	Platypus	6.0×	No	8.2%	66.5%	77.9%	63.8%	68.4%	76.8%	86.43%
Gallus gallus	Chicken	6.6×	No	3.8%	52.8%	73.3%	65.7%	78.2%	79.7%	88.61%
Anolis carolinensis	Lizard	6.8×	No	4.7%	57.8%	72.1%	63.5%	77.3%	77.7%	88.65%
Xenopus tropicalis	Frog	7.9×	No	2.6%	45.2%	64.6%	64.8%	80.1%	76.1%	87.44%
Tetraodon nigroviridis	Tetraodon	7.9×	No	2.0%	48.4%	61.5%	61.5%	62.0%	60.0%	79.12%
Takifugu rubripes	Fugu	8.5×	No	1.8%	41.6%	56.8%	62.5%	71.2%	62.5%	82.66%
Gasterosteus aculeatus	Stickleback	6.0×	No	1.9%	45.1%	61.2%	62.2%	68.7%	60.9%	82.22%
Oryzias latipes	Medaka	6.7×	No	2.0%	40.4%	56.4%	61.7%	72.9%	61.7%	82.92%
Danio rerio	Zebrafish	6.5×	No	2.0%	40.6%	58.8%	62.3%	74.2%	63.7%	82.38%

The "Seq. Cov." column describes the shotgun sequencing average read depth except for the human and mouse sequences, which are finished (Fin). "Filter" indicates what sort of filtering, either syntenic, reciprocal best, or no filtering, was applied to the pairwise alignments with human before they were brought into the multiple alignment (see Methods). "Total Align" is the percentage of human bases covered by alignments with that species (including alignments to the "-" gap character). "Start Aligns" and "Stop Aligns" show what percentage of human start and stop codons align in the other species. (The coverage of start and stop codons is discussed in greater depth later in the article.) "Total %ID" shows the percentage identity between human and the other species within aligning regions excluding gaps. "Start %ID" shows what percentage of aligning human start codons are also start codons in the other species, and similarly "Stop %ID" shows what percentage of aligning human stop codons are stop codons in the other species. "ORF Cover" measures the percent of human ORFs (region between start and stop codon) that are covered by the largest ORF in the same reading frame in the other species. The genes used for the start and stop codons and ORFs are the reviewed subset of human RefSeq genes (11,729 genes total).

6-bp deletion near the start of the prion protein gene, *PRNP*. The indel supports the claim that humans are more closely related to mice (and other glires) than to dogs (and other laurasiatherians). In brief, the reasoning is that the indel can be explained as resulting from a single evolutionary event if and only if primates and glires form a monophyletic clade; any other tree topology requires the unlikely situation that two or more evolutionary events combined to appear like a single event. Figure 2 shows this region of the 28-way alignment. Another use of coding indels to help understand the course of mammalian evolution is discussed by Murphy et al. (2007), supporting the assertion that elephants and armadillos separated from the human lineage in the same speciation event. The relationship between the alignments at the UCSC Browser and this use of coding indels has bidirectional synergy. Determination of the most likely history of mammalian evolution is needed to produce the most appropriate multiple alignments, and conversely, the pairwise and multiple alignments available at UCSC provide an excellent substrate for studying the mechanisms and history of vertebrate evolution.

Given that coding indels provide evidence concerning the topology (shape) of the phylogenetic tree, are they introduced at such a uniform rate that they can be used to estimate the time separating speciation events? To address this issue, we asked if accelerations or decelerations in the rate of insertions/deletions within protein-coding regions could be detected along certain branches of the phylogenetic tree. We started by searching the 28-way alignment for positions within annotated coding exons where humans have an insertion or deletion relative to both elephant and opossum, i.e., indels that appear to have been cre-

```
     Human  ATGGCGA------ACCTTGG
     Chimp  ATGGCAA------ACCTTGG
    Rhesus  ATGGCGA------ACCTTGG
  Bushbaby  ATGGCGA------GACTTGG
TreeShrew  ATGGCAC------AGCTGGG
     Mouse  ATGGCGA------ACCTTGG
       Rat  ATGGCGA------ACCTTGG
GuineaPig  ATGGCAA------ATGCCGG
    Rabbit  ATGGCGC------ACCTCGG
  Hedgehog  ATGGTGAAAAACCACGTGGG
       Dog  ATGGTGAAAAGCCACATAGG
       Cow  ATGGTGAAAAGCCACATAGG
 Armadillo  ATGGTGAGAAGCCGCGTAGG
  Elephant  ATGGTGAAAAGCAGCTTGGG
   Opossum  ATGGGGAAAATCCACTTGGG
    Lizard  ATGGGGAAGCACCAGATGAC
```

Figure 2. A 6-bp deletion near the start of *PRNP*. Species in the 28-way alignment lacking data for this region are not shown. This alignment can be seen in the current (hg18) UCSC Genome Browser at chr20:4,627,867–4,627,880. See Supplemental Figure S2 for an amino acid alignment of the deletion involving many more species.

ated during placental evolution along the human lineage. (The search used custom-built software; see Conclusions, below.) For each such indel and for each of the 16 other placental mammalian sequences in the alignment, we determined whether the species agrees with human, with elephant, or with neither. Finally, we retained the indels that could be unambiguously assigned to a currently sampled branch of the human lineage (relative to the phylogenetic tree that we used). For example, the deletion in Figure 2 would be assigned to the branch from the human–dog ancestor to the human–mouse ancestor (except that it fails a requirement for amount of flanking coding sequence imposed by our program). This process assigned 209 indels as shown in Figure 3.

We tested the hypothesis that coding indels were fixed in the human lineage at a uniform rate over time. A parametric bootstrap test showed that the observed indel frequencies differ significantly ($P < 10^{-3}$) from the hypothesis of uniform distribution. Specifically, the indels detected by our procedure seem to have occurred at a rate of about four per million years in early placental evolution, but at less than two per million years in recent times.

Human–specific protein indels

Among the assignments of coding indels to tree branches, described above, we identified 11 indels that arose after human–chimp divergence (Table 2). Much additional work is required before one can have any confidence that a particular mutation is related to an observable human trait; here we illustrate some appropriate kinds of bioinformatic analyses for two of these genes, *SULF1* and *GFM2*. Laboratory or clinical evidence is required to make a convincing case for any phenotypic consequences of an indel from this list.

SULF1 has a human-specific 3-bp insertion in exon 11. The insertion adds a GAA codon (E, in the amino acid sequence) to a run of four identical codons. The repetitive nature of this region immediately suggests a mechanism for expansion and/or contraction by replication slippage, which makes the change in the human protein perhaps seem unremarkable. The insertion appears to be fixed in humans (i.e., not polymorphic). No reliable

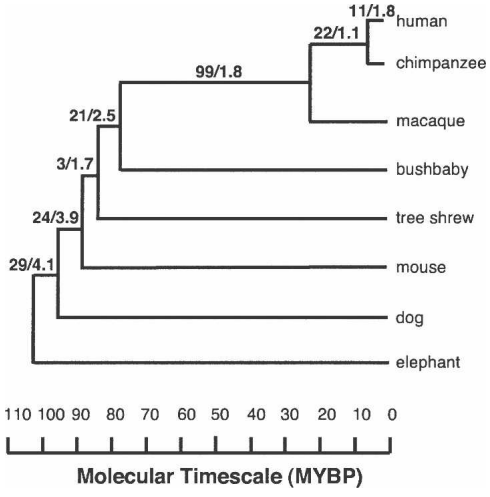

Figure 3. Number of inferred coding indels and number per million years on the placental branches leading to human. Estimated time elapsed on each branch is taken from Murphy et al. (2007).

Table 2. Eleven human-specific coding indels observed in the 28-way alignment

Gene	Near	Mutation	Description
ARHGEF10L	chr1:17,807,020	2 aa inserted	Rho guanine nucleotide exchange factor
C1orf131	chr1:229,441,270	1 aa deleted	
CHL1	chr3:41,4931	1 aa deleted	Cell adhesion molecule with homology with L1CAM
FAM83F	chr22:38,747,851	1 aa deleted	
GFM2	chr5:74,057,608	2 aa inserted	Mitochondrial elongation factor G2 isoform 1
LRIG1	chr3:66,514,655	2 aa inserted	Leucine-rich repeats and immunoglobulin-like
MAP7	chr6:136,723,965	2 aa inserted	Microtubule-associated protein 7
NPC1	chr18:19,370,803	1 aa inserted	Niemann-Pick disease, type C1
SULF1	chr8:70,698,813	1 aa inserted	Sulfatase 1
TCTA	chr3:49,424,929	3 aa deleted	T-cell leukemia translocation altered
ZCCHC6	chr9:88,127,675	1 aa inserted	Zinc finger, CCHC domain containing 6

Locations are in human genome assembly hg18.

information is available about the three-dimensional structure of this part of the protein sequence, and we could find no published association of genetic disease with exon 11.

On the other hand, conservation of the sequence around the insertion site is extremely high (Fig. 4). Moreover, conservation of the four Es and the surrounding sequence was also observed in orangutan, marmoset, tarsier, squirrel, microbat, dolphin, pig, sloth, hyrax, and wallaby, according to sequences we located by searching public databases. For some reason, nature has retained the pattern of precisely four successive Es (except for a loss of one E in the lizard sequence) over a total of 2 billion years of evolution along the various lineages. This evidence that the change was heavily resisted over evolutionary time, despite existence of a putative mechanism that could easily make the change (replication slippage), raises the question of whether the extra amino acid in humans is beneficial or deleterious.

A 6-bp insertion in human *GFM2* illustrates other aspects of the computational analysis of human-specific coding indels identified in the 28-way alignment. In this case, the region around the indel is not particularly well conserved. Indeed, the insertion appears to be found in only some humans (Fig. 5); i.e., some of the available human sequence data for this gene lack the extra 6 bp. We have some information about the three-dimensional conformation of this region of the *GFM2* product, based on the known structure of a protein (Protein Data Bank [PDB] entry 2bm0) with 42% identity. This model of the protein structure does not suggest that the insertion site is constrained by secondary or tertiary structural interactions that would make the sequence change difficult to accommodate (see Supplemental materials). That is, the structural model does not suggest that there is likely to be any phenotypic consequence for humans.

Have positions in potentially disease–associated deletions resisted substitution over evolutionary time?

The number of known amino acid differences in the human population is rapidly increasing. There is much interest in devel-

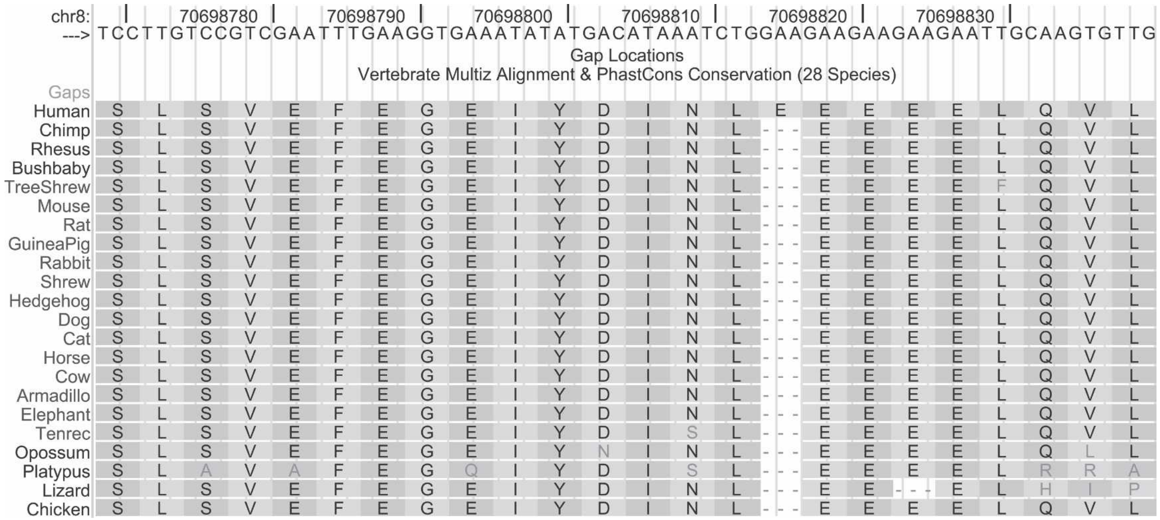

Figure 4. Extreme conservation of the region around the 3-bp insertion in human *SULF1*. The symbol "-" indicates that there is no base in the aligning species that aligns to this location. Placement of the gap at the first human E results from tie-breaking rules in the alignment software. The second gap in the lizard sequence was positioned using nucleotide content of codons. hg18.chr8:70,698,769–70,698,840.

oping computational tools to help predict which of these changes might have implications for human health, as summarized by Ng and Henikoff (2006). One type of clue that has been successfully employed comes from interspecies conservation. In particular, it is known that human replacement mutations resulting in disease are overabundant at amino acid positions that are most conserved throughout the long-term history of metazoans (Subramanian and Kumar 2006, and references cited therein).

One way to bring indels into the picture is to ask if disease-associated deletions in the human population tend to involve highly conserved amino acids. Our discussion differs somewhat from that of Subramanian and Kumar (2006) by focusing on potentially disease-associated deletions and by considering data from more species. We employed the data on human variants that are available in the PhenCode Locus Variants custom track (Giardine et al. 2007; www.bx.psu.edu/phencode) for the UCSC Human Genome Browser. In particular, we looked at the gene for PAH (phenylalanine hydroxylase); PAH deficiency causes PKU, the most common inborn error of amino acid metabolism in Caucasians. To keep everything simple, our measure of (lack of) conservation at each position was the number of distinct amino acids in that column of the 28-way alignment.

The distribution of these conservation scores at positions where a frame-preserving deletion (i.e., of length divisible by three) is annotated in the Locus Variants track was not strikingly different from the scores' distribution at all positions of PAH, so this small experiment failed to support our hypothesis. (The *P*-

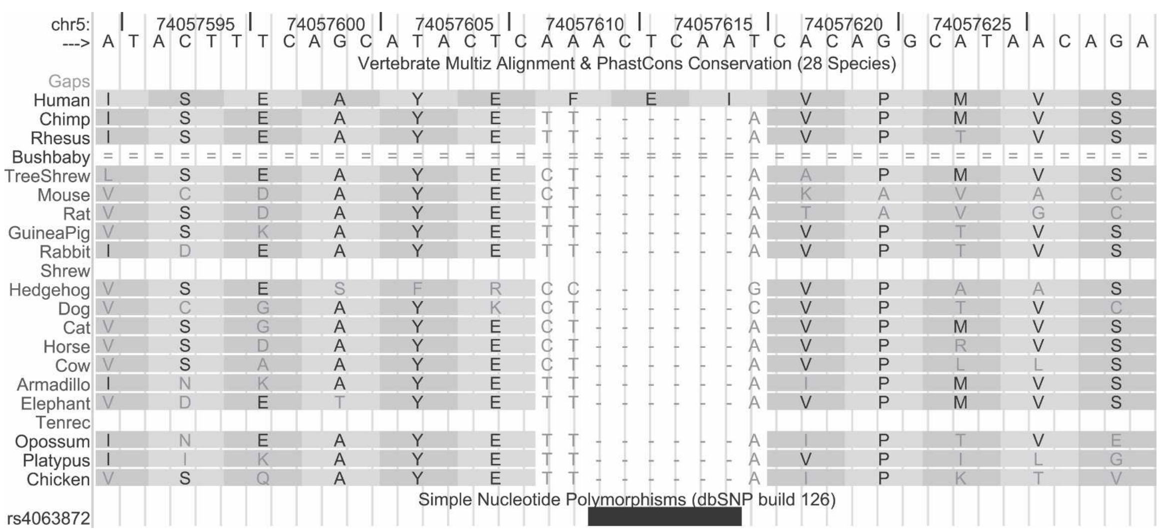

Figure 5. The 6-bp insertion in the human *GFM2* gene, showing the location of a 6-bp interval that is absent in some people. The symbol "-" indicates that there is no base in the aligning species that aligns to this location, and "=" indicates that at this location in the aligning species there is a sequence of bases of such different length and/or sequence composition that it cannot be reliably aligned. Sequence for this interval is currently not available for shrew or tenrec. chr5:74,057,590–74,057,630.

value for the two-tailed permutation test based on 1000 iterations was ~0.69.) However, the number of annotated PAH deletions is very small, so the result is heavily affected by one deletion where the amino acid has low conservation, shown in Figure 6.

The PAH mutations in the Locus Variants track were obtained from the PAHdb knowledgebase (Scriver et al. 2003; www.pahdb.mcgill.ca), which documents known mutations in this gene. The majority of these deletions result in deficient enzyme activity but some do not, so one cannot assume that the mere presence of a particular mutation in the track implies that it causes disease. In this case, one can follow links from the Locus Variants track to the publication cited for the association of PKU with this deletion (namely, Svensson et al. 1991). That and earlier work referenced in the cited article present evidence from Swedish families carrying this mutation that the alteration likely results in a "profound loss of enzymatic activity" (Svensson et al. 1990). In any case, as this small example suggests, the Browser, together with the Locus Variants custom track, is a valuable resource for this kind of study, provided that appropriate caution is taken.

Application 2: Conservation of start and stop codons

In general, as one would expect, the sequence within human coding regions is much more likely to align, and when aligning much more likely to be identical to other species, than the human genome overall. This becomes more and more apparent the more distantly related the other organism is to human. While the overall amount of the coding region that aligns to human typically remains >80%, even in fish, the stop and particularly the start codons drift away at a much faster rate (Table 1). By the time we reach zebrafish, only 40.6% of start codons and 58.8% of stop codons align. The trend of more stop than start codons aligning is present in every sequence except for the one other finished sequence, mouse, where 97.7% of starts and 96.8% of stops align (P-value for one-tailed test for proportions $< 10^{-3}$). This level of drift certainly makes the job of identifying precise gene boundaries based on comparative genomics data more difficult.

The greater rate of start codon drift compared with stop codon drift is puzzling. We advanced three hypotheses to explain this. Since the one finished sequence, mouse, was an exception to this rule, we hypothesized that it might be due to the CpG

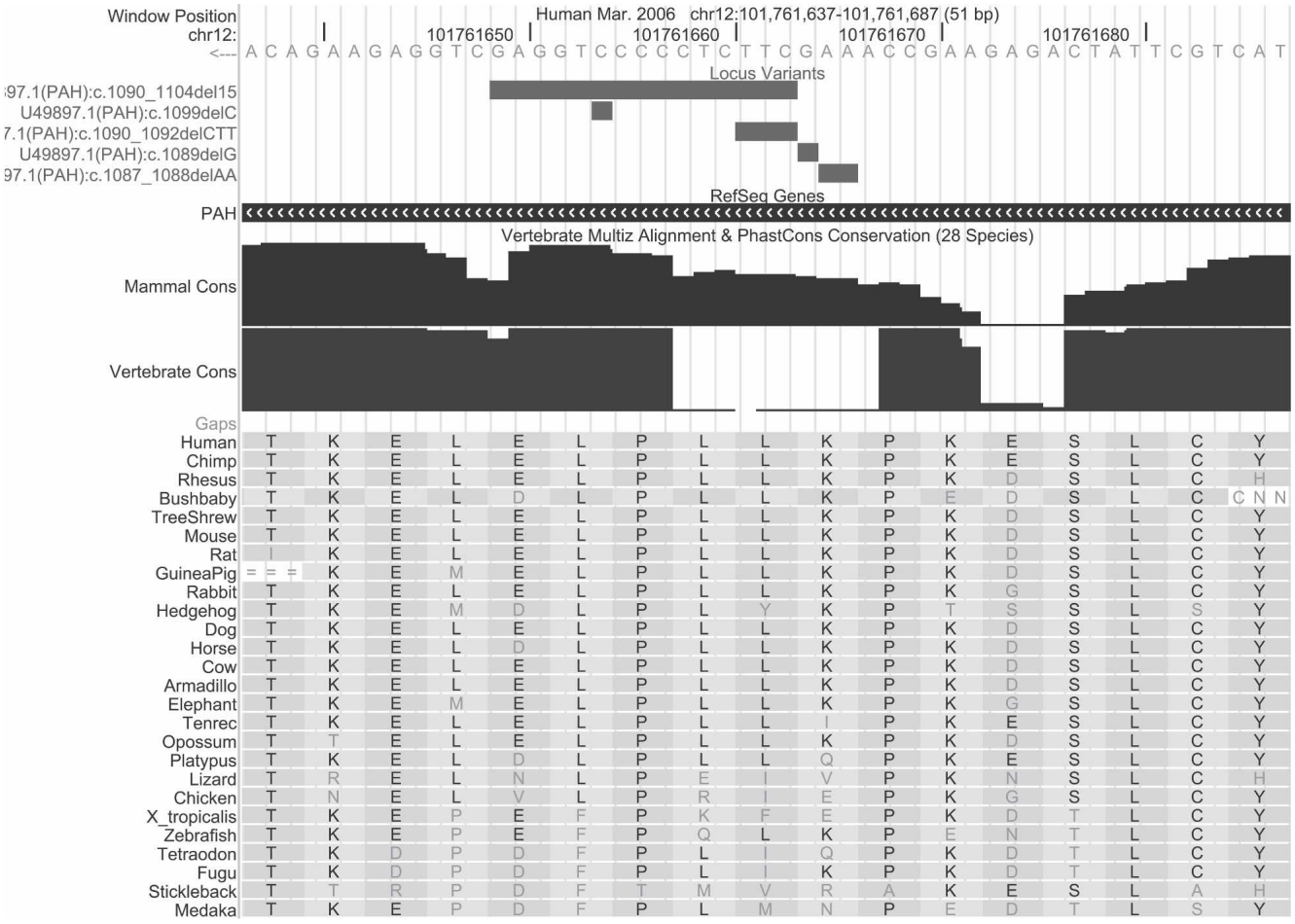

Figure 6. A segment of the gene for PAH, showing positions of some deletions that may be associated with the disease PKU. The CTT deletion (shown in reverse orientation) removes an amino acid whose column has six distinct letters and in that sense is not well conserved. The nucleotide symbol "N" represents an unsequenced base, and "=" indicates that at this location in the aligning species there is a sequence of bases of such different length and/or sequence composition that it cannot be reliably aligned. The two conservation tracks indicate that the deleted position is not well conserved among all vertebrates, but is fairly well conserved within mammals. chr12:101,761,637–101,761,687.

islands that are common near gene starts being more difficult to sequence. If we look at genes that do not start in a CpG island, indeed there is no statistically significant difference between the start and stop codon drift within placental mammals (*P*-value for the two-tailed test for proportions is 0.16). However at further evolutionary distances, the difference in drift is actually more pronounced in genes lacking CpG islands. For instance, for Tetraodon the difference between stop and start codon conservation is 58.48% − 50.79% = 7.71% for genes with a CpG island and 52.65% − 42.64% = 10.01% for those without (Table 3).

Another hypothesis was that selection at the start codon might be more relaxed in genes with multiple promoters, and this could account for some of the difference between the start and stop codon drift. However, this does not seem to be the case. For instance, the average difference for placental mammals was 96.38% − 94.73% = 1.65% for genes with alternate promoters, and 96.41% − 92.41% = 4.00% for other genes. The third hypothesis was that since the initial coding exon is often small, particularly at large evolutionary distances, our programs might not have enough surrounding conserved sequence to reliably align the region around the start codon. We did find that between humans and species more distant than placental mammals, there is a large increase in alignability of start codons that are part of exons with at least 100 coding bases (*P*-value < 10^{-3}), though still some disparity remained with 67% of stop codons but only 63% of start codons aligning. For additional details on these experiments, see Table 3.

Overall, a bias against CpG islands in the draft sequence combined with difficulty in aligning small initial coding exons does explain a great deal of the observed unalignability of start codons compared with stop codons. Another hypothesis, which is harder to test, is that the start of a protein, since it is very often trimmed by proteases, is subject to less selective pressure than protein ends. Related to this, part of the reason that the smaller initial coding exons are particularly vulnerable to this drift may be that there is less selection against deletion of the entire exon when the protein-coding portion of it is small.

Regardless of the actual cause of the drift, in practical terms it is something that investigators building gene models based on an analysis of multiple genomic alignments need to be well aware of, as was noted in regards to start codons based on CAGE tag data (Frith et al. 2006). This suggests that the total branch length provided by several placental species may be superior for gene prediction to the branch length provided by a single non-placental, and that finishing additional placental genomes may be worthwhile.

Application 3: Phylogenetic extent of alignment (alignability) of functional regions

A major observation of the recent article (Mikkelsen et al. 2007) on the genome sequence of *Monodelphis domestica* (short-tailed opossum) is that conservation in noncoding regions is much more subject to evolutionary turnover than in protein-coding regions. More specifically, conservation turnover refers to cases where an interval of human sequence shows signs of purifying selection compared with some species but not with others. The 28-way alignment provides an ideal resource for investigating conservation turnover in greater depth, as we now show.

Effective use of comparative genomics to find and better understand functional regions of genomes remains a challenge. Most coding exons show a strong signature of evolutionary constraint, but others are under positive selection for adaptive changes. A small fraction of *cis*-regulatory modules (CRMs) are deeply conserved from humans to fish (Woolfe et al. 2005), but others are present only in particular clades (Valverde-Garduno et al. 2004; King et al. 2007). The new multiple alignment provides the opportunity to examine the phylogenetic depth of conserva-

Table 3. Start/stop codon drift for genes with big and small initial protein-coding exons, with and without CpG islands, and with and without alternative promoters

Species	Big start	Big end	Small start	Small end	CpG start	CpG stop	No CpG start	No CpG stop	Alt start	Alt stop	No Alt start	No Alt stop
Chimp	94.23%	96.88%	93.99%	97.60%	92.99%	97.36%	96.11%	96.75%	95.41%	97.95%	93.76%	96.97%
Rhesus	93.15%	95.55%	93.61%	95.31%	92.49%	95.44%	95.07%	95.34%	95.36%	95.91%	93.12%	95.36%
Rat	95.24%	95.79%	93.63%	94.43%	94.67%	95.64%	94.25%	93.96%	94.69%	95.16%	94.53%	95.09%
Mouse	98.21%	97.19%	97.17%	96.18%	98.16%	97.35%	96.51%	95.58%	96.66%	96.88%	97.87%	96.71%
Dog	90.47%	97.52%	88.54%	96.55%	86.71%	97.46%	94.53%	96.06%	94.39%	96.02%	88.51%	97.21%
Horse	86.05%	96.99%	86.40%	96.42%	81.59%	96.83%	94.18%	96.48%	92.86%	96.96%	84.67%	96.71%
Cow	94.97%	96.70%	93.95%	96.49%	94.36%	97.20%	94.31%	95.50%	93.77%	95.75%	94.44%	96.84%
Opossum	88.19%	89.11%	75.04%	87.41%	82.65%	90.05%	79.45%	84.89%	80.17%	86.84%	82.10%	88.55%
Platypus	75.91%	79.25%	56.26%	76.45%	66.42%	79.52%	65.76%	74.64%	69.18%	76.27%	65.93%	78.09%
Chicken	64.72%	72.94%	38.96%	73.65%	52.37%	74.86%	53.35%	69.23%	55.33%	72.74%	52.24%	73.22%
Lizard	71.21%	73.35%	42.99%	70.32%	59.99%	74.82%	52.92%	65.79%	56.59%	71.32%	57.95%	72.13%
Frog	58.51%	64.48%	29.46%	64.31%	44.52%	66.83%	46.00%	58.88%	48.49%	63.91%	44.53%	64.60%
Tetraodon	60.53%	56.12%	34.92%	57.19%	50.79%	58.48%	42.64%	52.65%	43.75%	54.69%	49.30%	57.26%
Fugu	53.82%	61.32%	27.82%	60.76%	42.82%	62.71%	38.93%	58.28%	39.56%	61.66%	42.08%	61.47%
Stickleback	57.59%	61.11%	31.03%	61.03%	47.08%	63.19%	40.40%	56.69%	41.14%	58.30%	45.86%	61.80%
Medaka	52.13%	56.78%	27.11%	55.45%	41.12%	57.43%	38.37%	53.61%	37.97%	57.37%	40.90%	56.20%
Zebrafish	51.08%	57.20%	28.87%	59.74%	41.48%	60.79%	38.14%	53.77%	41.15%	55.51%	40.52%	59.42%
Placentals	93.19%	96.66%	92.47%	96.14%	91.57%	96.75%	94.99%	95.66%	94.73%	96.38%	92.41%	96.41%
Non-placentals	63.37%	67.17%	39.25%	66.31%	52.92%	68.89%	49.60%	62.84%	51.33%	65.86%	52.14%	67.27%
Average	75.65%	79.31%	61.16%	78.78%	68.84%	80.35%	68.29%	76.36%	69.20%	78.43%	68.72%	79.27%

The columns show the percentage of human start and stop codons that align (at the chosen thresholds) for various subsets of the RefSeq reviewed gene set from Table 1. The Big columns are from genes where (in human) there are at least 100 coding bases in the exon containing the start codon. The Small columns are from genes where there are less than 100 coding bases in that exon. The CpG columns are defined by whether the first 200 bases of the gene's transcript overlap a CpG island as defined by the corresponding track at the UCSC Genome Browser. The Alt columns are defined by whether the gene's transcription start site overlaps an AltPromoter item in the Alt Events track at UCSC.

tion of different functional classes at a higher resolution, because of the larger number of species included. In this study, we examined 251,000 coding exons of RefSeq genes (Pruitt and Maglott 2001), 481 ultraconserved elements (UCEs) (Bejerano et al. 2004), and 94,000 predicted regulatory regions, which we call PRPs, characterized by both clusters of conserved transcription factor binding sites (PReMods) (Blanchette et al. 2006) and a strong signal for alignment patterns that discriminate between regulatory regions and neutral DNA (high RP; see Methods) (Taylor et al. 2006). We also included a collection of 3900 putative transcriptional regulatory regions (pTRRs) discovered by chromatin immunoprecipitation followed by hybridization to high-density microarray chips (ChIP-chip experiments) from the ENCODE pilot project (The ENCODE Project Consortium 2007; King et al. 2007). The latter data set is composed of functional regions identified by methods that do not depend on sequences or their alignments.

Determining the range of comparison species in which homologs of a DNA sequence are present is fundamental to studying its evolution. We define the *alignability* of a particular DNA segment (e.g., in human) as the fraction that aligns with a designated comparison species. In the current study, the 28-way alignment is the basis for computing alignability (see Methods). In Figure 7C, the alignability of the human functional classes with each comparison species is plotted as a function of phylogenetic distance estimated by substitutions per fourfold degenerate (4D) site in coding regions. The 4D sites were the ones established by detailed examination of protein-coding segments in ENCODE regions (Harrow et al. 2006; The ENCODE Project Consortium 2007). The substitutions per 4D site were determined using the REV model (Yang 1997). A set of tools have been added to the Galaxy workspace (Blankenberg et al. 2007) to facilitate similar analyses of other data sets and other phylogenetic ranges.

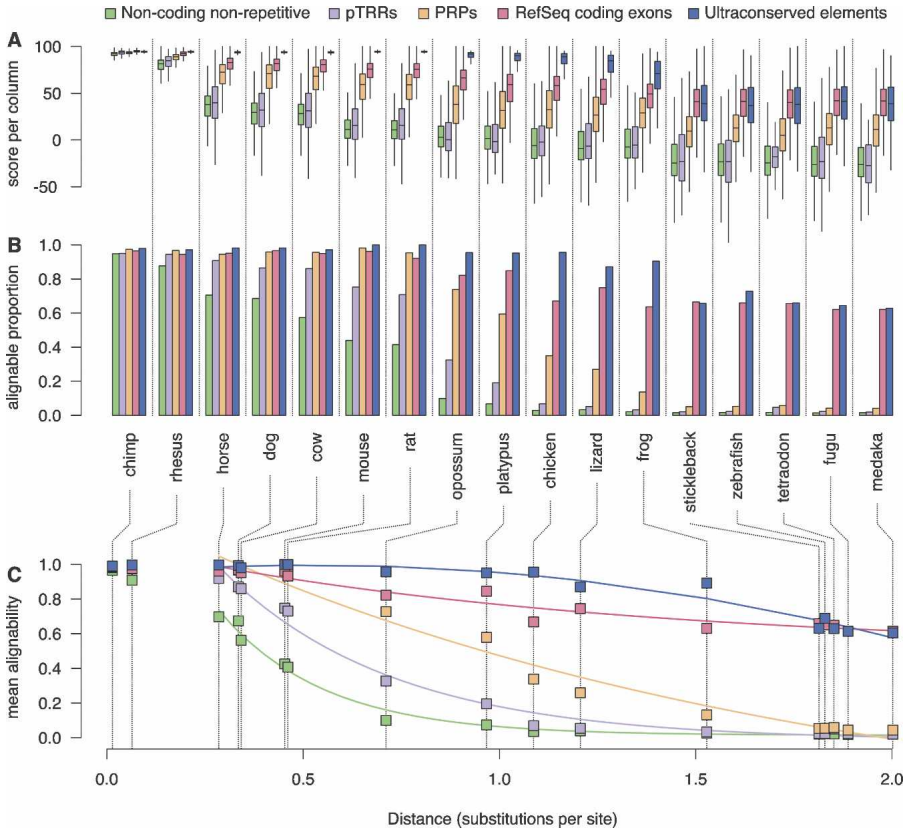

Figure 7. Phylogenetic extent of the alignment of functional features. (*A*) The distributions of alignment scores per column for the subset of intervals in each feature set (coding exons, UCEs, putative transcriptional regulatory regions, and PRPs) and the background human genome (nonrepetitive, noncoding) that align with each comparison species. For these box plots, the center line of each box is the median, the box extends from the 25th to 75th percentiles, and the feathers extend to 1.5 times the interquartile distance. The boxes are colored by feature set according to the legend along the *top*. (*B*) Barplots showing the fraction of intervals with >50% alignability for each feature set and for the background. (*C*) Decay of mean alignability as a function of phylogenetic distance. The mean alignabilities of the background human genome and intervals in each feature set are plotted against the distance from human to each comparison species. The distance is measured as the total substitutions per 4D site on each of the branches connecting human to the comparison species. The common name for each comparison species is given *below* the barplots in *B* and is connected to the phylogenetic distance in *C* by dotted lines. The data are best fit by two decay curves, one for primates with a slow rate of change and the other for horse to medaka. The curves shown are the fits to the data points from horse to medaka. (Statistics and coefficients for these fits are in Supplemental Table S4.)

In order to establish a baseline for evaluation of the alignability of various feature sets, we examined the alignability of the nonrepetitive, noncoding portion of the human genome, which we will refer to as the *background*. Loss of alignability is likely to be driven by deletions in the lineage to the comparison species and insertions in the human lineage, with some contribution from substitutions. If these events occur independently, then the alignability should decay exponentially with increasing phylogenetic distance (Kimura 1969). This is indeed observed for the background alignability, especially for comparisons outside primates (Fig. 7C; Supplemental Table S4).

This predictable decline of the background alignability with phylogenetic distance provides a basis for comparison of the decay (if any) in alignability of functional regions. We examined several feature sets to determine the range of comparison species over which high alignability is maintained, the mode and rate of decay, the fraction of intervals that continue to align at a given distance, and the quality of alignments, and we found striking differences.

Almost all coding exons align in all placental mammals, followed by a decay that is much slower than that of the background (Fig. 7B,C). Almost all UCEs align with species as distant as chicken, but a decline is seen with more distant species (Fig. 7B,C). Indeed, alignability of UCEs appears to decay as a function of the square of the phylogenetic distance. The decay equation (Supplemental material) predicts a substantial loss of recognizable UCEs at distances of 2.5 to 3 substitutions per site, which is consistent with the failure to detect UCEs outside vertebrates (Bejerano et al. 2004). The 60%–70% of UCEs conserved between humans and fish are distinct from the ones limited to mammals

(see Supplemental materials). The pan-vertebrate UCEs are further from genes on average than are the pan-mammalian ones, reminiscent of the jungles of noncoding conserved sequences observed in human–chicken comparisons (Hillier et al. 2004). As expected, the deeply conserved UCEs are near genes enriched in GO categories for transcription factors and developmental regulatory genes (Woolfe et al. 2005). Both the coding exons and UCEs maintain a high conservation level in all comparisons, shown by the high values in the distributions of alignment scores per column (Fig. 7A).

The predicted CRMs (PReMods with high regulatory potential, or PRPs) show substantially elevated alignability above that of the background DNA (Fig. 7B,C). The PRPs were selected at least in part by their ability to align among mammalian species, and as expected, virtually all of them align to other eutherian species. In addition, a substantial fraction aligns to marsupials and monotremes. The alignability for comparisons outside primates decays exponentially but substantially more slowly than the background. The distributions of conservation-level scores are considerably higher than the background but substantially less than those for coding exons (Fig. 7A). Similar results are obtained for the decay in alignability of a small set of 93 curated known regulatory regions (Elnitski et al. 2003; data not shown).

Another set of genomic regions (pTRRs), which is implicated in transcriptional regulation by ChIP-chip biochemical analyses, shows a different pattern. The loss of alignability with phylogenetic distance for the ENCODE pTRRs is more rapid than the decay seen for PRPs (Fig. 7C). The shape of the decay curve is similar to that of the background, but shifted to the right. The rightward shift indicates a slower decay, implying that the pTRRs are constrained; i.e., as a class they are eroding more slowly than the background. This is also seen in the higher fraction of the pTRR intervals that pass an alignability threshold compared with that of the background (Fig. 7B). Interestingly, the similarity level of the alignments in the pTRR intervals overlaps substantially with that of the background, whereas the alignments of UCEs and coding exons have strikingly higher scores (Fig. 7A). Results similar to those for pTRRs are obtained for a genome-wide set of 13,000 segments occupied by the transcription factor CTCF (Kim et al. 2007; data not shown).

While the pTRRs as a class do have a higher alignability than the background for many comparisons, the number of pTRRs that are conserved drops dramatically between eutherian mammals (horse, dog, cow, and rodents) and the marsupial opossum (Fig. 7.B). This shows that 70% of pTRRs are conserved within eutherian mammals, while 70% are not conserved outside of eutherians. Smaller fractions are conserved in more distant comparisons. The pTRRs conserved out to chicken and fish are likely to have been under constraint over this phylogenetic span. Given the estimates for rates of neutral substitutions, sequences that still align between human and chicken, and more distant species, are all likely to be under constraint (Hillier et al. 2004).

This study illustrates the range and complexity of the relationships between the functionality of genomic sequences and their phylogenetic extent of alignability. The decays of alignability for some strongly constrained functional classes differ markedly from that of the genome background. Coding exons are largely conserved through eutherians, after which they decay much more slowly than the background. Almost all UCEs are conserved in amniotes, after which their alignability decays more rapidly than do coding exons. Alignability of functional classes associated with gene regulation presents exponential decay curves, but with lag times and slopes characteristic of the feature sets and distinctly different from the background. The alignability of PRPs begins to decline outside most eutherians whereas the alignability of pTRRs and CTCF binding sites begins to decay outside the available primates. This can be interpreted as stringent selection being exerted over different phylogenetic spans for the distinct types of regulatory regions, followed by decay in sequence similarity. The mode of decay in alignability is exponential both for the features associated with regulation and for the background. This could mean that the processes of sequence change are similar, with deletions and insertions allowed as in the background genome (albeit at a slower rate), whereas these events tend to be rejected in coding exons and UCEs.

Conclusion

To use alignments of placental-mammal genome sequences to identify small intervals (say, the size of a transcription-factor binding site) that are under strong negative selection, one needs sequence data from perhaps 40 well-chosen species (Eddy 2005; Margulies et al. 2005). Of course, one could obtain an equivalent total phylogenetic branch length by using a smaller number of more distant species, but as shown in Figure 7, many functional sites will then be lost to conservation turnover. The 28-way alignment comes the closest to date to attaining this 40-species goal. Here we illustrate the use of those alignments to explore hypotheses about vertebrate evolution. The individual results are interesting in their own right and worthy of further study, but they are merely a sampling of what we and others will discover with those alignments.

Some of the phenomena described here can be directly observed in the UCSC Browser. Others require the use of tools such as the UCSC Table Browser (Karolchik et al. 2004) or the Galaxy server (Giardine et al. 2005; Blankenberg et al. 2007) to help identify genome-wide trends. The most generally applicable approach is to write special purpose programs to analyze the alignments. For instance, C-language programs used for analyzing the rate of coding indels can be downloaded from http://www.bx.psu.edu/miller_lab/publications/. Other programs for manipulating the alignments are mentioned in the Methods section.

The analyses presented here were necessarily quite brief, and in each case, it would be possible to strengthen the analysis by using more complex approaches. For instance, our modeling of coding indels was quite simplistic, and much more sophisticated methods (e.g., Diallo et al. 2007) could be applied. Similarly, we frequently assumed that regions aligned to a functional human gene are themselves functional genes in the other species, though some of them may have become inactive pseudo-genes. To help ameliorate the (presumed minor) effects of such cases, one could incorporate gene-model data from the other species, such as that available from Ensembl (Hubbard et al. 2007).

While in practice we use the 28-way alignment to explore many issues, there are times when other alignments might be more appropriate. For instance, the 28-way alignment is based on a phylogenetic tree (Fig. 1) that is not universally accepted (e.g., see Wible et al. 2007). While this may not affect most conclusions, there are cases, such as the use of the alignment to support or reject certain evolutionary hypotheses, where the tree matters (Kumar and Filipski 2007). In such cases, it might be better to use a set of pairwise alignments that are agnostic about phylogenetic hypotheses (e.g., Rosenbloom et al. 2007). Also, our alignments

are made with programs that implement only one approach and that use particular settings of a large number of thresholds and other parameters; others may prefer to use different software, such as the tools discussed by Margulies et al. (2007). In any case, users need to remain aware that there are genomic regions in which the 28-way alignment, or indeed an alignment computed by any means, is unreliable because evolutionary changes have saturated the sequences (Prakash and Tompa 2007).

A general theme running through our observations, and one that continues to enthrall many investigators, is to identify the best ways to use multispecies alignments to help predict the locations of functional genomic elements. In particular, are these elements more frequently revealed by reduced levels of substitutions, by particular patterns of substitutions, by reduced levels of small insertions and deletions, or by a reduced level of complete loss? Data summarized in Figure 7 can be interpreted as suggesting that resistance to complete loss is sometimes more informative than low substitution rates. Another, very preliminary, study reported here raises the question of whether substitution frequency can be used to predict whether deletion of an amino acid might cause human disease. Whole-genome alignments provide a valuable resource for investigating these and many other fascinating issues.

Methods

Phylogenetic tree

We used the tree topology (Fig. 1) that seemed in best agreement with our interpretation of the published literature. Branch lengths, which are used for quantifying sequence conservation (see below) but not for computing the alignments, were computed by phyloFit as described below.

Alignments

Pairwise alignments with the human genome were generated for each species (BLASTZ; Schwartz et al. 2003) from repeat-masked genomic sequence (RepeatMasker, by A. Smit and R. Hubley [Institute for Systems Biology, Seattle, WA], or WindowMasker, by Morgulis et al. 2006), with lineage-specific repeats removed prior to alignment, then reinserted. The pairwise alignment coverage for all species is listed in Table 1. Pairwise alignments were then linked into chains using a dynamic programming algorithm that finds maximally scoring chains of gapless subsections of the alignments organized in a kd-tree (axtChain; Kent et al. 2003). The scoring matrix and parameters for pairwise alignment and chaining were tuned for each species based on its phylogenetic distance from humans. High-scoring chains were then placed along the genome, with gaps filled by lower-scoring chains (axtNet, by J. Kent), to produce an alignment net. Filtering of the component pairwise alignments was performed to reduce paralogs, pseudogenes, and suspect alignments from the 2× species. The alignments of high-quality mammalian sequences (placental and marsupial) were filtered based on synteny, while those for 2× mammalian genomes were filtered to retain only alignments that were best quality in both species ("reciprocal best"). The resulting best-in-genome pairwise alignments were progressively aligned following the tree topology of Figure 1, using the MULTIZ program (Blanchette et al. 2004). Alignments were post-processed to add annotations for alignment gaps and genomic breaks, indications of base quality in the component sequences (see Supplemental material), and information that permits prediction of amino acid sequences in all species.

Conservation scoring was performed using the phastCons package (Siepel et al. 2005), which computes conservation based on a two-state phylogenetic hidden Markov model (HMM). These measurements rely on a tree model containing the tree topology, branch lengths representing evolutionary distance at neutrally evolving sites, the background distribution of nucleotides, and a substitution rate matrix. The vertebrate tree model for this track was generated using the phyloFit program from the phastCons package (REV model, EM algorithm, medium precision) using multiple alignments of fourfold degenerate sites extracted from the 28-way alignment (msa_view). The 4D sites were derived from the October 2005 Gencode Reference Gene set (Harrow et al. 2006), which was filtered to select single-coverage long transcripts. A second tree model, including only placental mammals, was used to generate the placental mammal conservation scoring. The phastCons parameters were tuned to produce 5% conserved elements in the human genome for the vertebrate conservation measurement; this parameter set (expected-length, 45; target-coverage, 0.3; rho, 0.31) was then used to generate the placental mammal conservation scoring.

The phastCons program computes conservation scores based on a phylo-HMM, a type of probabilistic model that describes both the process of DNA substitution at each site in a genome and the way this process changes from one site to the next (Yang 1995; Felsenstein and Churchill 1996; Siepel and Haussler 2005). PhastCons uses a two-state phylo-HMM, with a state for conserved regions and a state for nonconserved regions. The value plotted at each site is the posterior probability that the corresponding alignment column was "generated" by the conserved state of the phylo-HMM. These scores reflect the phylogeny (including branch lengths) of the species in question, a continuous-time Markov model of the nucleotide substitution process, and a tendency for conservation levels to be autocorrelated along the genome (i.e., to be similar at adjacent sites). Unlike many conservation-scoring programs, phastCons does not rely on a sliding window of fixed size; therefore, short highly conserved regions and long moderately conserved regions can both obtain high scores. More information about phastCons can be found in work by Siepel et al. (2005).

PhastCons currently treats alignment gaps as missing data, which sometimes has the effect of producing undesirably high conservation scores in gappy regions of the alignment. We are looking at several possible ways of improving the handling of alignment gaps.

Extraction and analysis of MAF blocks

Local portions of the 28-way alignment are stored as so-called MAF (Multiple Alignment Format) blocks. In Galaxy, MAF-block extraction is implemented using an on-disk variation of the positional binning scheme described by Kent et al. (2002) to allow fast extraction of alignment blocks overlapping specific regions of the human sequence. Due to the size of these alignments, compression is essential; here we used LZO compression (http://www.lzop.org/). As a result of compression, it is no longer possible to seek directly to a particular alignment block in these files. However, because the LZO format compresses the data in independent chunks, accessing a particular alignment block only requires decompressing the containing chunk(s) of compressed data (i.e., semi-random access). This is a substantial benefit over the common GZIP format, for example, which requires decompressing all preceding data to access a particular location. We have implemented an indexing scheme for (1) identifying the locations of all MAF blocks that overlap a particular interval in some aligned species and (2) mapping those locations to their containing chunks of compressed data. Combined, these meth-

ods make working with alignments of this scale substantially easier. Command-line programs and Python modules for generating and using these indexes are available as part of the "bx-python" package (http://bx-python.trac.bx.psu.edu/).

PRPs

PRPs are a set of predicted CRMs that have both the properties of (1) clusters of conserved transcription factor binding sites (PReMods; Blanchette et al. 2006) and (2) high regulatory potential (high RP intervals; Taylor et al. 2006). PRP is a brief name denoting the combination of PReMod and RP.

The set of 118,402 PReMods (human genome assembly hg17) was collected from the Web server discussed by Blanchette et al. (2006). The RP scores were determined for a seven-way alignment of human (hg17), chimpanzee (panTro2), macaque (rheMac2), mouse (mm8), rat (rn4), dog (canFam2), and cow (bosTau2). Human DNA intervals with an RP score of at least 0.05 for at least 200 bp (i.e., the minimum score never goes below 0.05) were selected to obtain 314,020 high-RP intervals.

The next series of operations was conducted using online tools in Galaxy. KnownGenes exons obtained from the UCSC Table Browser (extended 15 bp on each side) were subtracted from the high RP intervals, and intervals shorter than 200 bp were removed to obtain a total of 282,639 nonexonic high-RP intervals. These were intersected with the PReMods, requiring an overlap of at least 10 bp, to produce 106,383 pieces present in both sets. Intervals within 100 bp of each other were then merged, the merged intervals were combined with the other intervals from the intersection, and those shorter than 50 bp were removed. This series of operations produced a set of 92,269 PRPs. Their average length is 350 bp, ranging from 50 bp to 3793 bp.

The ENCODE regions cover 1% of the human genome and have been examined experimentally for chromatin alterations, occupancy by a set of transcription factors, and transcription start sites (The ENCODE Project Consortium 2007; Margulies et al. 2007). Examination of the 1389 PRPs that lie within ENCODE regions supports the hypothesis that the PRPs are good predictors of regulatory function. About half are within 100 bp of an interval associated with transcriptional regulation (using the ENCODE pilot phase data available now at the UCSC Web site), which is about a twofold enrichment over the bulk DNA in the ENCODE regions. In some well-studied ENCODE regions, the support level for PRPs reaches 100%.

Alignability

The alignability of an interval in the human sequence with another species is the proportion of bases in that interval that are covered by any local pairwise alignment with the other species. Positions in the human interval are conservatively classified as aligned, not aligned, or potentially missing using the local alignments and alignment gap annotation, with the goal of excluding from analysis all positions in the human sequence that are not aligned due to missing data in the other species sequence. Specifically, (1) all positions covered by a local alignment are marked aligned; (2) if there is no local alignment or gap annotation covering the position, or the gap annotation indicates contiguity across the position but with missing data in the other species sequence, the position is marked potentially missing; and (3) if there is no local alignment but the gap annotation indicates contiguity in both species, the position is marked not aligned. Alignability is then the number of aligned positions divided by the number of aligned plus not aligned positions. For 10 of the 28 species, the alignability measures were markedly different from those seen for other species at a similar distance from human. This is likely a result of incomplete coverage, since all had ~2× coverage or less. The program used to compute alignability, "maf_interval_alignability.py," is available as part of the "bx-python" package (http://bx-python.trac.bx.psu.edu/).

Acknowledgments

This project has been funded in part by grants from NHGRI (HG002238 to W.M., and 1P41HG02371 to W.J.K.), NIDDK (DK65806 to R.C.H.), and NCI (Contract no. NO1-CO-12400). Sudhir Kumar and the reviewers provided many useful suggestions. Additional help was provided by Cathy Riemer and Heather Lawson.

References

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W.J., Mattick, J.S., and Haussler, D. 2004. Ultraconserved elements in the human genome. *Science* **304:** 1321–1325.

Blanchette, M., Kent, W.J., Riemer, C., Elnitski, L., Smit, A.F.A., Roskin, K.M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E.D., et al. 2004. Aligning multiple genomic sequences with the Threaded Blockset Aligner. *Genome Res.* **14:** 708–715.

Blanchette, M., Bataille, A.R., Chen, X., Poitras, C., Laganiere, J., Lefebvre, C., Deblois, G., Giguere, V., Ferretti, V., Bergeron, D., et al. 2006. Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. *Genome Res.* **16:** 656–668.

Blankenberg, D., Taylor, J., Schenck, I., He, J., Zhang, Y., Ghent, M., Veeraaghavan, N., Albert, I., Miller, W., Makova, K., et al. 2007. A framework for collaborative analysis of ENCODE data: Making large-scale analyses biologist-friendly. *Genome Res.* **17:** 960–964.

Diallo, A.B., Makarenkov, V., and Blanchette, M. 2007. Exact and heuristic algorithms for the Indel Maximum Likelihood problem. *J. Comput. Biol.* **14:** 446–461.

Eddy, S.R. 2005. A model of the statistical power of comparative genome sequence analysis. *PLoS Biol.* **3:** e10. doi: 10.1371/journal.pbio.0030010.

Elnitski, L., Hardison, R.C., Li, J., Yang, S., Kolbe, D., Eswara, P., O'Connor, M.J., Schwartz, S., Miller, W., and Chiaromonte, F. 2003. Distinguishing regulatory DNA from neutral sites. *Genome Res.* **13:** 64–72.

The ENCODE Project Consortium. 2007. The ENCODE pilot project: Identification and analysis of functional elements in 1% of the human genome. *Nature* **447:** 799–816.

Felsenstein, J. and Churchill, G.A. 1996. A hidden Markov model approach to variation among sites in rate of evolution. *Mol. Biol. Evol.* **13:** 93–104.

Frith, M.C., Ponjavic, J., Fredman, D., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y., and Sandelin, A. 2006. Evolutionary turnover of mammalian transcription start sites. *Genome Res.* **16:** 713–722.

Giardine, B., Riemer, C., Hardison, R.C., Burhans, R., Elnitski, L., Shah, P., Zhang, Y., Blankenberg, D., Albert, I., Taylor, J., et al. 2005. Galaxy: A platform for interactive large-scale genome analysis. *Genome Res.* **15:** 1451–1455.

Giardine, B., Riemer, C., Hefferon, T., Thomas, D., Hsu, F., Zielenski, J., Sang, Y., Elnitski, L., Cutting, G., Trumbower, H., et al. 2007. PhenCode: Connecting ENCODE data with mutations and phenotype. *Hum. Mutat.* **28:** 554–562.

Gross, S.S. and Brent, M.R. 2006. Using multiple alignments to improve gene prediction. *J. Comput. Biol.* **13:** 379–393.

Hardison, R.C. 2000. Conserved noncoding sequences are reliable guides to regulatory elements. *Trends Genet.* **16:** 369–372.

Harrow, J., Denoeud, F., Frankish, A., Reymond, A., Chen, C.K., Chrast, J., Lagarde, J., Gilbert, J.G., Storey, R., Swarbreck, D., et al. 2006. GENCODE: Producing a reference annotation for ENCODE. *Genome Biol.* **7:** S4.1–S4.9. doi: 10.1186/gb-2006-7-s1-s4.

Hillier, L.W., Miller, W., Birney, E., Warren, W., Hardison, R.C., Ponting, C.P., Bork, P., Burt, D.W., Groenen, M.A., Delany, M.E., et al. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432:** 695–716.

Hubbard, T.J., Aken, B.L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T., et al. 2007. Ensembl 2007. *Nucleic Acids Res.* **35:** D610–D617. doi: 10.1093/nar/gkl996.

International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. *Nature* **431:** 931–945.
Karolchik, D., Hinrichs, A.S., Furey, T.S., Roskin, K.M., Sugnet, C.W., Haussler, D., and Kent, W.J. 2004. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* **32:** D493–D496. doi: 10.1093/nar/gkh103.
Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. 2002. The human genome browser at UCSC. *Genome Res.* **12:** 996–1006.
Kent, W.J., Baertsch, R., Hinrichs, A., Miller, W., and Haussler, D. 2003. Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl. Acad. Sci.* **100:** 11484–11489.
Kim, T.H., Abdullaev, Z.K., Smith, A.D., Ching, K.A., Loukinov, D.I., Green, R.D., Zhang, M.Q., Lobanenkov, V.V., and Ren, B. 2007. Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell* **128:** 1231–1245.
Kimura, M. 1969. The rate of molecular evolution considered from the standpoint of population genetics. *Proc. Natl. Acad. Sci.* **63:** 1181–1188.
King, D.C., Taylor, J., Zhang, Y., Cheng, Y., Martin, J., ENCODE groups for Transcriptional Regulation and Multispecies Sequence Analysis, Chiaromonte, F., Miller, W., and Hardison, R.C. 2007. Finding *cis*-regulatory elements using comparative genomics: Some lessons from ENCODE data. *Genome Res.* **17:** 775–786.
Kumar, S. and Filipski, A. 2007. Multiple sequence alignment: In pursuit of homologous DNA positions. *Genome Res.* **17:** 127–135.
Lander, E.S. and Waterman, M.S. 1988. Genomic mapping by fingerprinting random clones: A mathematical analysis. *Genomics* **2:** 231–239.
Margulies, E.H., Vinson, J.P., NISC Comparative Sequencing Program, Miller, W., Jaffe, D.B., Lindblad-Toh, K., Chang, J.L., Green, E.D., Lander, E.S., Mullikin, J.C., et al. 2005. An initial strategy for the systematic identification of functional elements in the human genome by low-redundancy comparative sequencing. *Proc. Natl. Acad. Sci.* **102:** 4795–4800.
Margulies, E.H., Cooper, G.M., Asimenos, G., Thomas, D.J., Dewey, C.N., Siepel, A., Birney, E., Keefe, D., Schwartz, A.S., Hou, M., et al. 2007. Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. *Genome Res.* **17:** 760–774.
Mikkelsen, T.S., Wakefield, M.J., Aken, B., Amemiya, C.T., Chang, J.L., Duke, S., Garber, M., Gentles, A.J., Goodstadt, L., Heger, A., et al. 2007. Genome of the marsupial *Monodelphis domestica* reveals innovation in non-coding sequence. *Nature* **447:** 167–178.
Murphy, W.J., Pringle, T.H., Crider, T.A., Springer, M.S., and Miller, W. 2007. Using genomic data to unravel the root of the placental mammal phylogeny. *Genome Res.* **17:** 413–421.
Ng, P.C. and Henikoff, S. 2006. Predicting the effects of amino acid substitutions on protein function. *Annu. Rev. Genomics Hum. Genet.* **7:** 61–80.
Pedersen, J.S., Bejerano, G., Siepel, A., Rosenbloom, K., Lindblad-Toh, K., Lander, E.S., Kent, W.J., Miller, W., and Haussler, D. 2006. Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput. Biol.* **2:** e33. doi: 10.1371/journal.pcbi.0020033.
Poux, C., van Rheede, T., Madsen, O., and de Jong, W.W. 2002. Sequence gaps join mice and men: phylogenetic evidence from deletions in two proteins. *Mol. Biol. Evol.* **19:** 2035–2037.
Prakash, A. and Tompa, M. 2007. Measuring the accuracy of genome-size multiple alignments. *Genome Biol.* **8:** R24. doi: 10.1186/gb-2007-8-6-r124.
Pruitt, K.D. and Maglott, D.R. 2001. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.* **29:** 137–140.
Rhesus Macaque Genome Sequencing and Analysis Consortium. 2007. Evolutionary and biomedical insights from the rhesus macaque genome. *Science* **316:** 222–234.
Rosenbloom, K., Taylor, J., Schaeffer, S., Kent, W.J., Haussler, D., and Miller, W. 2007. Phylogenomic resources at the UCSC Genome Browser. In *Methods and molecular biology: Phylogenomics* (ed. W.J. Murphy) Humana Press, Totowa, NJ. (in press).
Schwartz, S., Kent, W.J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R.C., Haussler, D., and Miller, W. 2003. Human-mouse alignments with BLASTZ. *Genome Res.* **13:** 103–107.
Scriver, C.R., Hurtubise, M., Konecki, D., Phommarinh, M., Prevost, L., Erlandsen, H., Stevens, R., Waters, P.J., Ryan, S., McDonald, D., et al. 2003. PAHdb 2003: What a locus-specific knowledgebase can do. *Hum. Mutat.* **21:** 333–344.
Siepel, A. and Haussler, D. 2004. Computational identification of evolutionarily conserved exons. In *Proceedings of the 8th annual international conference on research in computational molecular biology*, pp. 177–186. ACM Press, New York.
Siepel, A. and Haussler, D. 2005. Phylogenetic hidden Markov models. In *Statistical methods in molecular evolution* (ed. R. Nielsen), pp. 325–351. Springer, New York.
Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15:** 1034–1050.
Subramanian, S. and Kumar, S. 2006. Evolutionary anatomies of positions and types of disease-associated and neutral amino acid mutations in the human genome. *BMC Genomics* **7:** 306. doi: 10.1186/1471-2164-7-306.
Svensson, E., Andersson, B., and Hagenfeldt, L. 1990. Two mutations within the coding sequence of the phenylalanine hydroxylase gene. *Hum. Genet.* **85:** 300–304.
Svensson, E., von Döbeln, U., and Hagenfeldt, L. 1991. Polymorphic DNA haplotypes at the phenylalanine hydroxylase locus and their relation to phenotype in Swedish phenylketonuria families. *Hum. Genet.* **87:** 11–17.
Taylor, J., Tyekucheva, S., King, D.C., Hardison, R.C., Miller, W., and Chiaromonte, F. 2006. ESPERR: Learning strong and weak signals in genomic sequence alignments to identify functional elements. *Genome Res.* **16:** 1596–1604.
Valverde-Garduno, V., Guyot, B., Anguita, E., Hamlett, I., Porcher, C., and Vyas, P. 2004. Differences in the chromatin structure and *cis*-element organization of the human and mouse GATA1 loci: Implications for *cis*-element identification. *Blood* **104:** 3106–3116.
Wible, J.R., Rougier, G.W., Novacek, M.J., and Asher, R.J. 2007. Cretaceous eutherians and Laurasian origin for placental mammals near the K/T boundary. *Nature* **447:** 1003–1006.
Woolfe, A., Goodson, M., Goode, D.K., Snell, P., McEwen, G.K., Vavouri, T., Smith, S.F., North, P., Callaway, H., Kelly, K., et al. 2005. Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol.* **3:** e7. doi: 10.1371/journal.pbio.0030007.
Yang, Z. 1995. A space-time process model for the evolution of DNA sequences. *Genetics* **139:** 993–1005.
Yang, Z. 1997. PAML: A program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13:** 555–556.

Received June 4, 2007; accepted in revised form August 30, 2007.