Annu. Rev. Genomics Hum. Genet. 2004. 5:15–56
doi: 10.1146/annurev.genom.5.061903.180057

First published online as a Review in Advance on April 15, 2004

COMPARATIVE GENOMICS

Webb Miller, Kateryna D. Makova, Anton Nekrutenko, and Ross C. Hardison

The Center for Comparative Genomics and Bioinformatics, The Huck Institutes of Life Sciences, and the Departments of Biology, Computer Science and Engineering, and Biochemistry and Molecular Biology, Pennsylvania State University, University Park, Pennsylvania; email: webb@bx.psu.edu, kdm16@psu.edu, anton@bx.psu.edu, rch8@psu.edu

Key Words whole-genome alignments, evolutionary rates, gene prediction, gene regulation, sequence conservation, Internet resources, genome browsers, genome databases, bioinformatics

■ **Abstract** The genomes from three mammals (human, mouse, and rat), two worms, and several yeasts have been sequenced, and more genomes will be completed in the near future for comparison with those of the major model organisms. Scientists have used various methods to align and compare the sequenced genomes to address critical issues in genome function and evolution. This review covers some of the major new insights about gene content, gene regulation, and the fraction of mammalian genomes that are under purifying selection and presumed functional. We review the evolutionary processes that shape genomes, with particular attention to variation in rates within genomes and along different lineages. Internet resources for accessing and analyzing the treasure trove of sequence alignments and annotations are reviewed, and we discuss critical problems to address in new bioinformatic developments in comparative genomics.

INTRODUCTION

Determining the genome sequences of humans and model organisms is a landmark achievement in the life sciences. Analysis of the individual genome sequences gives much insight into genome structure but less into genome function (84, 179). One grand challenge for the next phase of genomics research is to distinguish functional DNA and then assign a role to it (35). Comparative genomics is one of the major approaches used in the functional annotation of genomes. Functional sequences are subject to evolutionary selection, which can leave a signature in the aligned sequences. Purifying (negative) selection causes sequences to change more slowly than the bulk, nonfunctional sequences (Figure 1), and Darwinian (positive) selection causes sequences to change more rapidly. In principle, comparing genomic sequences can find these signatures of selection, and hence one

1527-8204/04/0922-0015$14.00

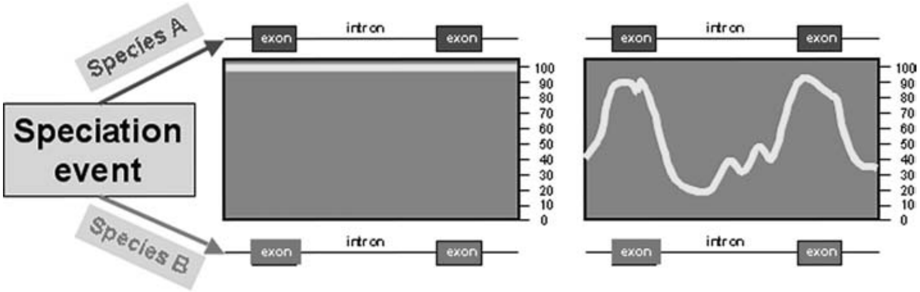


Figure 1 Evolution of functionally important regions over time. Immediately after a speciation event, the two copies of the genomic region are 100% identical (see graph on left). Over time, regions under little or no selective pressure, such as introns, are saturated with mutations, whereas regions under negative selection, such as most exons, retain a higher percent identity (see graph on right). Many sequences involved in regulating gene expression also maintain a higher percent identity than do sequences with no function.

can infer that such sequences are functional. Other analytical approaches, still in their infancy, are designed to predict the role of those sequences. However, both the inference of function and the assignment of a role to sequences are predictions. They need experimental tests, such as those currently being done on a large scale in the Encyclopedia of DNA Elements (ENCODE) project (35).

Whole-genome comparisons based on DNA-level alignments have greatly expanded the precision and depth of evolutionary analysis and functional inference. The first genomes sequenced were so phylogenetically distant that comparisons were limited to the encoded protein sequences. Thus, a large set of common proteins could be identified in yeast, worms, and flies (153), but noncoding sequences could not be meaningfully compared. Detailed information about the evolution and function of genomes can be gathered by comparing species that are more closely related. For example, comparing three genera of enteric bacteria (*Escherichia coli*, several species of *Salmonella*, and *Klebsiella pneumonia*) thought to have diverged about 150 mya showed that most of the genes are conserved and tend to be in the same order, and thus most of the genome was stable over this period (126). In the past two years, several genome comparisons at this closer phylogenetic distance have been published, beginning with comparison of the mouse and human genomes (132), followed by detailed comparisons of multiple species of budding yeasts (33, 34, 90) and of *Caenorhabditis elegans* with *C. briggsae* (167). A light-coverage sequence of the dog genome (96) provides additional insights into comparative genomics. Our review focuses on these recent results from whole-genome alignments, and those from a pilot study of deep phylogenetic sequencing of the *CFTR* locus (175). However, the biological issues are informed by results from earlier locus-specific studies, and some of these are also included.

The publicly available genome sequence data and related resources are a treasure trove for biologists seeking functional elements in a particular region or wishing to perform genome-wide studies. However, as biologists dive into this rich pool, it is appropriate to reflect on potential pitfalls for the endeavor; these are discussed throughout. Existing resources and tools for investigators to delve more deeply into such issues and to use comparative genomics in their experimental investigations are summarized. Finally, we discuss new directions for software development to improve the use of comparative genomics.

LESSONS LEARNED FROM COMPARATIVE GENOMICS

What Have We Learned About Genes by Comparing Genomic Sequences?

GENE PREDICTION IN SINGLE-GENOME ERA The evolution of gene prediction tools closely follows developments in genomic biology. Initial gene prediction tools such as Grail (178) were designed to locate protein-coding exons within the short fragments of genomic DNA that were available at the time. Advances in sequencing technology accelerated the assembly of longer segments of genomic DNA, which could contain multiple genes on both strands. This prompted the development of algorithms capable of predicting and assembling exons into multiple gene structures on both strands of a single genomic fragment. These can be called conventional gene prediction algorithms; here the word "conventional" is used to emphasize the fact that these tools do not use information from genome comparisons. These algorithms include Genscan, the most popular gene prediction tool (24), GenMark (117), Fgenesh (155), GeneID (144), and others (for an excellent overview of conventional gene prediction algorithms see Reference 190). These tools use sophisticated pattern recognition approaches to find protein-coding regions within genomic DNA and to assemble them into genes. However, such patterns may occur by chance, leading to a high false-positive rate characteristic of most current gene prediction algorithms (69, 150). For instance, the number of Genscan predictions in the human genome is more than twofold higher than the count of experimentally verified genes (65,010 versus 24,037, respectively; based on Ensembl release 18.34.1, http://www.ensembl.org). Although some of the ~41,000 additional predictions may represent true novel genes, many are likely false positives. This is why computationally predicted genes are routinely compared to experimental data or known sequences, most often to Expressed Sequence Tags (ESTs), to identify true predictions. A current release of the UniGene database (http://www.ncbi.nlm.nih.gov/UniGene) contains almost 4 million human ESTs. One may think this adequate for accurate gene annotation in humans, but even this large number of ESTs does not fully sample all tissues and all stages of development (190). Also, most methods of constructing the cDNA clones leave the 5′ terminal exons under-represented in EST databases. Most other species have substantially fewer ESTs, and thus the problem of fully annotating genes is exacerbated in those species. Thus, other approaches are needed.

COMPARISON OF MULTIPLE GENOMES OFFERS A RELIABLE SOLUTION The availability of multiple genomes, a scenario difficult to imagine a few years ago, offers the prospect of improved gene prediction by comparing genomic regions from multiple species and finding conserved (relatively unchanged) regions (Figure 1). For example, in closely related species such as human and mouse, exons of protein-coding genes tend to change substantially slower that the surrounding noncoding DNA. This is because a mutation that changes an amino acid is less likely to be retained in a functional protein-coding gene than a mutation that does not (109). The gene structure (the number and order of individual exons) is also well conserved among closely related species (12). Thus, genes can be predicted by comparing sequences from two (or more) genomes because we expect individual exons and their relative positions to be conserved.

Using comparative information in gene prediction results in a marked increase in sensitivity of predictions and substantially reduces the number of false positives (100). The improvement is dramatic, but it does not apply in all cases. In particular, the gain and loss of exons or entire genes in one species limits the effectiveness of this approach. For instance, although the protein complement in two mammals may differ by less than 1% (132), about 4% of the 19,500 protein-coding genes identified in the *C. briggsae* genome have no detectable matches in the *C. elegans* genome (167).

Current state-of-the-art methods for comparative prediction of protein-coding regions can be divided into three categories: (*a*) alignment-based algorithms that reconstruct gene structures by integrating a global alignment or a collection of local alignments with conventional gene prediction methods, (*b*) algorithms that perform alignment and gene recognition simultaneously, and (*c*) evolutionary algorithms that detect the signature of purifying selection within genomic alignments.

ALIGNMENT-BASED ALGORITHMS The first category includes methods such as ROSETTA (12), SGP1 (182), SGP2 (144), TWINSCAN (100), and DOUBLESCAN (128). ROSETTA reconstructs colinear gene structures from global alignments and defines exons as subsequences bounded by splice sites [modeled using the Maximum Dependence Decomposition Approach (23)]. It assumes that there is only one gene in each of the two input sequences. SGP1 reconstructs genes from a collection of local alignments between two sequences. Gene structures are reconstructed independently on each sequence. Exons are also defined as stretches of DNA between splice sites and/or start/stop codons. SGP2 assesses the reliability of gene models predicted by GENEID, a conventional gene predictor (68), using TBLASTX matches to another genome (2). Similarly, TWINSCAN represents a direct extension of the Genscan algorithm (23, 24) that integrates conservation information between two sequences into probabilities reported by the original Genscan model. TWINSCAN, like Genscan, models exons using frame-specific hexamer frequencies (53). DOUBLESCAN uses a Pair Hidden Markov Model (Pair HMM) to reconstruct gene structures from a series of local alignments created with BLAST (2).

SIMULTANEOUS ALIGNMENT/GENE FINDING ALGORITHMS One-step algorithms utilize a novel Generalized Pair HMM (GPHMM) approach for simultaneous alignment and prediction of genes from two unaligned, unannotated sequences. The GPHMM was implemented in the SLAM gene prediction program (1, 143). Because alignment and prediction are simultaneous, SLAM assumes that the order and the direction of genes and their exons are conserved between the two compared sequences.

EVOLUTIONARY ALGORITHMS Evolutionary algorithms do not rely on sequence similarity per se, rather they test whether the homologous sequences are truly protein-coding by using evolutionary signals within these sequences (137, 138). These signals are derived from the fact that nucleotide substitutions in coding regions can be classified as nonsynonymous (those which lead to amino acid replacements in the encoded polypeptide) or synonymous (or silent, those which do not change the encoded amino acid). The amount of each type of change is commonly given as the ratio K_A, which is the number of nonsynonymous substitutions per nonsynonymous site, and K_S, which is the number of synonymous substitutions per synonymous site.

Aligned genomic sequences from two species are deemed coding if they share a reading frame in which K_A is significantly less than K_S, i.e., the K_A/K_S ratio is much lower than expected for neutrally evolving DNA. This is because in the majority of true protein-coding regions nonsynonymous changes are subject to strong selective constraints (110). A server called ETOPE (136) is available for applying this evolutionary test to computationally predicted exons (Table 1).

CHALLENGES OF COMPARATIVE GENE PREDICTION Current methods for comparative prediction of protein-coding regions have several limitations. First, existing methods cannot be easily extended to three or more species, due to the dramatic increase in computational complexity (143). A high-quality draft sequence of the rat genome is available (148), and the dog genome is currently being actively sequenced. In a few years genomic sequences for dozens of vertebrate organisms will be available. Thus, it is necessary to develop tools that can use multiple sequences, which will dramatically improve the quality of genome annotation for these species.

Second, most current methods require training sets to derive parameters for underlying models (with the exception of evolutionary algorithms). Therefore, these programs are conservative, which may prevent detection of unusual features such as long exons, exons flanked by noncanonical splice sites, or single-exon genes. At present, it is virtually impossible to derive representative training sets for comparative methods because such sets must contain true orthologs from two species whose sequences are compared. In other words, to derive a reliable training set from a pair of genomes they must already be well annotated. For example, the DOUBLESCAN algorithm was trained using a set of only 36 genes (128). Dependence on training sets restricts current methods to a particular species; a program

TABLE 1 Internet resources for whole-genome comparative analysis and associated tools

Resource	URL
UCSC Genome4 Bioinformatics	http://genome.ucsc.edu/
Ensembl	http://www.ensembl.org/
MapViewer	http://www.ncbi.nlm.nih.gov/mapview/
VISTA Genome Browser	http://pipeline.lbl.gov/
K-BROWSER	http://hanuman.math.berkeley.edu/cgi-bin/kbrowser2
Comparative Regulatory Genomics	http://corg.molgen.mpg.de/
GALA	http://www.bx.psu.edu/
EnsMart	http://www.ensembl.org/EnsMart/
ETOPE	http://www.bx.psu.edu/
PipMaker and MultiPipMaker	http://www.bx.psu.edu/
VISTA server	http://www-gsd.lbl.gov/vista/
MAVID server	http://baboon.math.berkeley.edu/mavid/
zPicture server	http://zpicture.dcode.org/
rVISTA server	http://rvista.dcode.org/

trained on a set of human genes cannot perform well on *Drosophila* or *Caenorhabditis* genomes, for example. It may also be difficult to derive representative training sets from newly sequenced genomes where gene information is scarce.

Third, most current methods for comparative gene prediction assume that gene structures are colinear in both sequences. This limitation prevents these methods from predicting mosaic genes or genes containing duplicated or translocated exons. Although it is believed that such genes are not common, recent studies on segmental duplications in the human genome suggest that chromosome 22 alone contains 11 mosaic genes (10). Therefore, it is important to develop a tool that deals with such instances.

Fourth, although more reliable than conventional algorithms, predictions made by comparative gene-finding tools are often inconsistent (i.e., a prediction made by one algorithm is omitted by another). For example, initial analysis of the mouse genome reported a total of 22,011 protein-coding genes included in the "consensus data set" (132). However, the supplementary information section of the paper shows that the various methods predicted far more genes. For example, 48,451, 48,462, and 14,006 genes were predicted by SGP2, TWINSCAN, and SLAM, respectively. What is the biological significance of genes that are predicted but not included in the consensus data set? This question is difficult to answer with present methods, especially if these additional genes do not match existing ESTs or known protein sequences. However, such additional genes exist in great numbers. For example, if we assume that there is a complete overlap between the consensus data set of 22,011 genes and the set of 48,451 genes predicted by SGP2, then how can we classify the remaining 26,440 genes?

IMPLICATIONS FOR GENE NUMBER Comparing genomic sequences from multiple species has dramatically changed our understanding of how many genes it takes to make a human (or generic mammal) (Figure 2). The textbook figure of 100,000 genes began to crumble in 2000 with the publication of a comparative study suggesting that the lower bound of the human gene number may be only 28,000 (37). Comparing human genomic sequences with the genomic DNA of pufferfish (*Tetraodon nigroviridis*) produced this unexpectedly low number. Pufferfish was chosen because of its compact genome size and its substantial divergence from human, which eliminated virtually all noise due to nonfunctional conservation. Human and pufferfish are separated by ~400 MY. Any genome region preserved over such evolutionary distance will likely be functionally important. However, the high degree of divergence also implies that some mammal-specific genes are not counted in the total gene number estimate, making it too conservative. One way to put mammal-specific genes back into the picture is to compare the human genome with that of another mammal. Such an opportunity arose with the completion of the mouse genome project. In the case of human/mouse comparison, the use of experimentally derived data (such as cDNAs) with comparative gene predictors resulted in an estimate that is only slightly higher than the pufferfish-based count, i.e., about 30,000 genes (132).

The exact count of human (or mammalian) genes remains unknown, but it is likely not much higher than 30,000–35,000. A refined analysis of gene structure and ability to reliably identify unusual genome features such as overlapping or nested genes will contribute to the resolution of the gene count controversy. Comparative gene prediction methods will be essential to achieve this goal. For example, a recent application of evolutionary methods suggested that human genome alone may have as many as ~13,000 additional exons conserved with mouse and rat (137). The exact count of human genes may remain elusive for some time, but we expect the range of estimates to continue to tighten over the next few years.

What Have We Learned About Regulation?

DNA sequences that act in *cis* to regulate the timing and level of gene expression [*cis*-regulatory modules (CRMs)] include promoters, enhancers, silencers, and insulators/boundary elements. Promoters and enhancers are thought to be well conserved, and numerous examples support this conclusion. One of the first cellular enhancers discovered, which is in an intron of the kappa immunoglobulin light chain gene, was initially identified as a strongly conserved noncoding segment of the gene (49). Many studies (early examples include 44, 146) used alignments of promoter and enhancer sequences to find critical sequence motifs, which are generally binding sites for transcription factors. This approach, called phylogenetic footprinting (70, 173), works well for interspecies comparisons of noncoding DNA sequences (Figure 2). Fewer examples of silencers and insulators are available, although a binding site for the protein CTCF is a common feature of boundary elements (154). Critical studies of the interspecies conservation of these other types of CRMs are important topics for future studies.

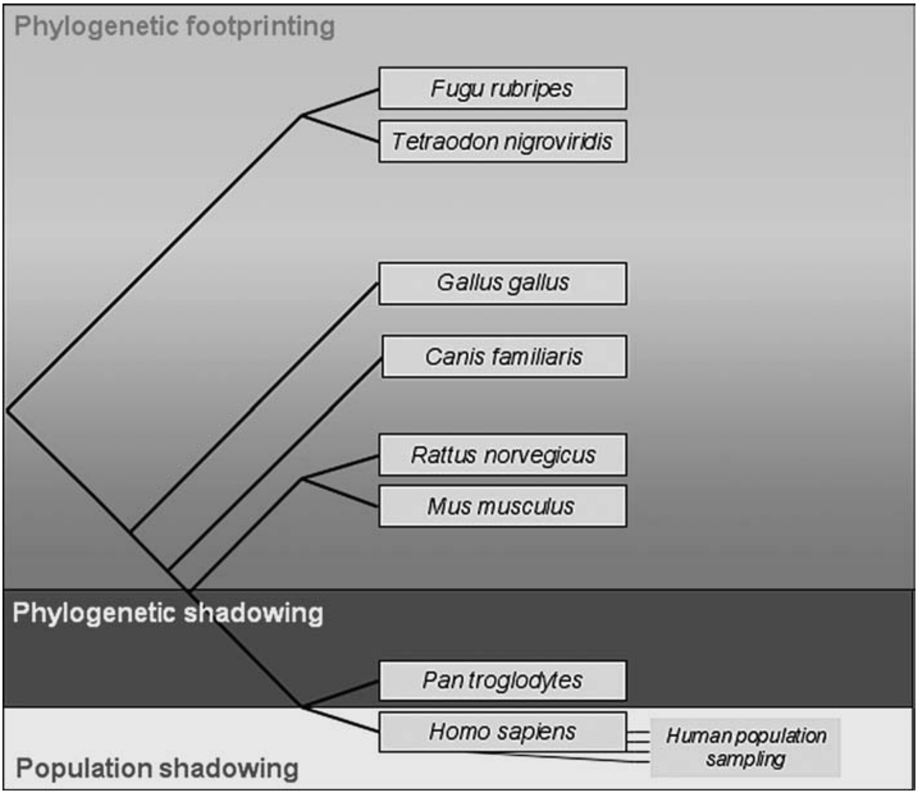


Figure 2 Use of whole-genome comparisons at various evolutionary distances to annotate the human sequence. Shaded areas representing different methods underlay a phylogenetic tree of selected vertebrates. Phylogenetic footprinting looks for the signature of negative selection, which shows regions that have undergone significantly less change than other, largely neutral DNA. Regulatory sequences (e.g., transcription factor binding sites), unlike protein-coding regions, are subject to rapid turnover (38, 116). Thus, predicting regulatory regions is more reliable when comparing genomic sequences at different divergence levels, such as the human (*Homo sapiens*) genome with the mouse (*Mus musculus*), rat (*Rattus norvegicus*), dog (*Canis familaris*), chicken (*Gallus gallus*), and fish (*Fugu rubripes* and *Tetraodon nigroviridis*) genomes. Some genes may have dramatically different expression levels in more divergent species (171). In this case comparing sequences from species belonging to the same class or family (phylogenetic shadowing) is more appropriate (18). Finally, to identify species-specific elements, one can analyze substitution polymorphisms in a sample of individual sequences from the same species, a technique termed population shadowing (120).

Several studies used a high level of sequence conservation to successfully predict and test enhancers. One of the most dramatic examples is a highly conserved noncoding region located between interleukin genes (114) of human and mouse. Removing this highly conserved sequence from a large DNA fragment containing several genes led to a loss of expression of the interleukin genes in about half the cells that normally express them, thus showing what is required for efficient expression. Other studies found particularly well-conserved sequences within known mammalian regulatory elements that led to the identification of additional protein-binding sites needed for full regulatory function (e.g., 48, 73). Studies of the *SCL* locus show that extending the phylogenetic distance of the comparison to human and chicken (63) allows detection of some but not all (64) functional enhancers. Analysis of a gene paralogous to *SCL*, *LYL1*, indicated that human-marsupial comparisons may be particularly effective for predicting regulatory elements (27). In some loci, such as *HOX* genes, CRMs are conserved between mammals and fish (6). Other studies, termed phylogenetic shadowing (Figure 2), focus on alignments of a larger number of more closely related species, such as several higher primates, to improve the accuracy of CRM detection (18, 72, 73).

However, the simple model of CRMs under strong purifying selection, and thus rarely changing, is not sufficient. Studies of the stripe 2 *eve* enhancer in several species of *Drosophila* show that many protein-binding sites have changed considerably while retaining the same expression pattern of *eve* (116). These data indicate a slow but continual turnover of protein binding sites under constant stabilizing selection, such that some changes compensate for alterations at other sites. In agreement with these results, a study of many CRMs shows that about 30% to 40% of the functional sites in human were no longer functional in mouse, which supports widespread turnover of transcription factor binding sites (38).

Studies were conducted to detect differences in binding sites in situations in which orthologous genes were differentially regulated. For instance, to support a longer gestation, higher primates have prolonged expression of the gamma-globin gene into the fetal period; other mammals, including primitive primates such as lemurs, express gamma-globin only in the embryo. Because of this difference, part of the regulatory region immediately upstream of the transcription start site is not conserved between primates and other mammals (85). Scientists identified a short segment contributing to the difference in expression by looking for regions that did not align (71).

Other factors add to the challenge of accurately predicting CRMs from sequence comparisons. One is the considerable variation in the rate of neutral evolution along chromosomes (77, 132, 186), which is discussed below. Thus, a noncoding conserved sequence with a particular alignment quality score in a slowly changing region is less significant than one with an identical score in a rapidly changing region. Methods are being developed to accommodate local rate variation in the evaluation of the likelihood that a sequence is under selection (108, 132).

A second complication is the substantial distance that can separate CRMs from their target genes. Within mammalian globin gene complexes, major regulatory elements can be 40–60 kb from the target promoters (67, 80), and the highly conserved noncoding sequence (mentioned above) can affect interleukin genes as much as 200 kb away (114). In some cases, promoters for genes not affected by the CRM are located between the CRM and its target. Thus, proximity is not always a reliable indicator of the target of a predicted CRM, and a considerable amount of DNA around any gene needs to be examined for a potential role in regulation. The long-distance effects, coupled with the variation in local rates of evolution, make the job of uniquely identifying CRMs even more difficult. For example, the major distal regulatory element of mammalian beta-globin gene complexes, the locus control region, stands out in interspecies comparisons as the only extensive noncoding region that aligns within about a 250-kb region (22, 75). In contrast, the major distal regulatory element of alpha-globin genes is one of many conserved noncoding sequences in a comparably sized region, and other noncoding regions are even more highly conserved than the active CRM (54).

A third complication concerns changes in CpG island density in mammals. CpG islands have long been associated with many regulatory elements, especially promoters for genes expressed in multiple tissues (16). Early reports estimated that the mouse genome lost roughly 20% of the CpG islands found in human (5, 124), and comparing the whole genomes showed that the proportion is closer to 43%. One example of this is found in the alpha-globin genes, *HBA1* and *HBA2*. Large CpG islands encompass the promoters and portions of some genes in humans, but the mouse homologs are not in or near CpG islands. The promoter sequences are poorly conserved between human and mouse, with the matches largely limited to one major transcription factor binding site.

Despite these serious problems in the bioinformatic prediction of CRMs based on sequence conservation, there are reasons for optimism. First, methods for predicting CRMs are in their infancy. The largely anecdotal information summarized here is derived from intensive study of a small number of loci. The whole-genome sequences and alignments open the door to large-scale studies that can monitor the local evolutionary rates and provide critical training sets for more sophisticated approaches to predicting CRMs. Second, most studies to date have been confined to pairwise comparisons. More genomes are being sequenced, and the additional information in multiple aligned sequences promises to provide a substantial increase in resolving power (175). Finally, it is clear that identifying clusters of defined transcription factor binding sites can be a powerful approach to predicting CRMs (14, 179a), even without information about interspecies conservation. It is reasonable to expect that progress in combining various independent methods of predicting CRMs will generate programs with greatly improved levels of sensitivity and specificity.

Current methods of using interspecies alignments for predicting CRMs evaluate the alignments either for their quality, with higher scores for more slowly changing regions, or for characteristic patterns. (Scoring for alignment quality is covered below.) Searching for characteristic patterns in alignments is in some respects similar to the conventional gene prediction algorithms discussed above. The aim is to find patterns in the alignments that are characteristic of known CRMs but are rarely seen in neutrally evolving DNA. One approach, described by Elnitski et al. (47), condenses the alignment to a small set of symbols that distinguish matches, mismatches, and gaps, and also distinguishes types of matches and types of mismatches. Alignments within a set of known CRMs serve as a training set to estimate the frequency of all, e.g., pentamer to hexamer transitions, which can be described as a fifth-order Markov model. Likewise, alignments within ancestral repeats provide a negative training set to estimate the frequency of these transitions in one model for neutral DNA. The ratios of the transition frequencies can be used to compute a log likelihood that any alignment has patterns of alignment symbols more characteristic of CRMs than neutral DNA. This score, called regulatory potential (47), can be generated because of the availability of the whole-genome alignments. Initial calibration and experimental tests of the effectiveness of this score are encouraging (76). Again, this is a fairly simple exploration of this approach and we expect improvements both from a greater amount of data (more genomes, larger training sets) and more statistical sophistication in the future.

About 5% of the Human Genome is Under Purifying Selection

Detailed comparative studies of several individual genes and gene clusters in mammals (often in humans and rodents) show that protein-coding exons are usually highly conserved, the untranslated regions are less well conserved, and often the *cis*-acting sequences regulating their expression are also conserved, albeit to varying extents. If these were the predominant functional DNA sequences in mammalian genomes, they would be the major targets for purifying selection. For example, the five active genes in the human *HBB* complex consist of about 2205 bp of coding exons, 920 bp of untranslated exons, and 3640 bp of known strong CRMs in an interval of about 68,000 bp. This corresponds to 3.3% coding exons, 1.3% untranslated exons, and 5.4% regulatory elements, for a total of 10% presumably subject to selection. However, this exon density is roughly twice that seen genome-wide. Analyses of both the human and mouse genomes indicate that coding exons probably occupy no more than 1.5% of these genomes (84, 132). Extrapolating from the ratio of all known functional sequences to coding sequences in the *HBB* complex (10% to 3.3%), one may predict that about 4.5% of the human genome would be under purifying selection.

This was tested by analyzing whole-genome human-mouse alignments (132). A good genome-wide monitor of the rate of neutral substitution is the set of aligning sites within ancestral repeats predating the human-mouse divergence, called AR sites. Advantages and disadvantages of AR sites as models for neutral DNA are covered below. Because the substitution rate at AR sites varies along chromosomal DNA (77, 132), quality scores for the alignments need to be adjusted for this rate variation. Thus, a conservation score was computed in 50-bp nonoverlapping windows for human DNA aligned with mouse (requiring at least 45 bp to be in alignments), adjusting for the matches expected based on the local neutral rate (empirically determined for AR sites within the surrounding region) and normalized by dividing by the standard deviation.

The genome-wide distribution of this conservation score overlaps considerably with the distribution of scores for ancestral repeats, but it is skewed substantially toward higher scores (132). This indicates that a second component is present at the higher scores, representing the sequences under selection. The distribution for all genomic alignments was decomposed into two components, an empirically determined distribution for neutral DNA (based on alignments in ancestral repeats) and an inferred distribution for sequences under purifying selection. The subset of sequences under selection contains at least 21% of the windows, which contain about 25% of the human genome. Therefore, at least 5% of the human genome is subject to purifying selection by this analysis. This is close to the value extrapolated from genome-wide coding exon density and the ratio of known functional sequences to coding sequences in the *HBB* complex.

The analysis summarized here does not allow each aligning segment to be unambiguously assigned as either neutral or under selection. However, one can compute a probability that a sequence with a particular conservation score belongs to the subset of windows under selection (132). One version of this analysis was implemented as L-scores (or Mouse cons track) at the UCSC Genome Browser (93). Other methods based on multiple sequence alignments are also being implemented (122, 164, 175), one of which is illustrated in Figure 3*A*.

The notion that at least 5% of the human genome is functional provides a rough initial guide for considering the scope of the problem of functional annotation. Because about 1.5% of the genome codes for protein and about 1% is in untranslated regions of genes (132), about 2.5% of the genome fulfills other functions. One major additional function is regulating gene expression, and some of the noncoding conserved sequences identified in this analysis fall into this category. Methods such as computing the regulatory potential and seeking conserved clusters of transcription factor binding sites should help in identifying such CRMs, as discussed above.

Some functional regions do not fall into the three categories listed so far. MicroRNAs are involved in regulating expression of genes in plants, worms, and other species, and many microRNAs have mammalian homologs (131). The number of examples of microRNAs involved in regulation in mammals is increasing steadily (139). Further studies are needed to estimate how much of a genome encodes microRNAs, and one may expect improved bioinformatic approaches for predicting their presence. It is likely that other classes of genomic sequences under selection have not yet been defined, such as sequences involved in chromosome replication and recombination.

One intriguing harbinger of future novel insights is a recent examination of highly conserved nongenic sequences on human Chromosome 21 (39). These are significantly more conserved than protein-coding exons and noncoding RNAs, and their pattern of substitution resembles that of protein-binding segments of DNA.

Other studies indicate that known regulatory regions are not as highly conserved as exons (47, 132), so these exceptionally well-conserved segments may comprise a novel class of critical sequences. Clearly, functional investigations of these sequences will be highly informative.

Despite the utility of this lower bound on the share of the human genome under selection, we expect that future studies will improve and refine the estimate. The accuracy of the decomposition of the conservation scores genome-wide depends completely on the reliability of the AR sites as models of neutral DNA. Although this is likely a good assessment for most sites, some of them are under selection. Also, AR sites align between human and mouse. It is possible that this set of ancestral repeats is biased toward those in more slowly changing regions; those in more rapidly changing regions might not align by current methods. Improved models of neutral DNA derived from comparisons of sequences from more closely related species may provide a better data set for the decomposition. Finally, the decomposition is performed on small windows, but even for windows that will likely be in the group under selection, not every nucleotide will be under selection. This tends to inflate the estimate of the share under selection, but it is somewhat balanced by the fact that short regions under selection may not be included in the windows under selection. These several limitations to the current analysis point to the need for additional work and should serve as a caution in interpreting the estimate of 5% of the genome under purifying selection.

Positively Selected Regions

Detecting DNA sequences evolving under positive selection is a challenge because they accumulate many changes. However, positively selected regions are among the most interesting regions in the genome because their evolution is likely adaptive and they frequently determine biological differences between organisms. Among the best-characterized examples of positively selected regions are genes encoding proteins involved in defense against pathogens, such as human histocompatibility determinants (83); reproduction, exemplified by the sperm protein lysin in abalone (105); speciation, including the homeodomain protein Odysseus in *Drosophila* (177); and adaptation to a new environment, e.g., lysozyme in langur monkeys (127). A comparative genomics approach can identify additional positively selected genes by contrasting their evolutionary rates with that of conserved genes. Genes detected by this approach are strong candidates for subsequent functional studies.

For instance, comparing the genomes of four species of yeast (*Saccharomyces cerevisiae*, *S. paradoxus*, *S. mikatae*, and *S. bayanus*) identified a rapidly evolving gene *YBR184W* (90). The *YBR184W* sequence has only 32% nucleotide identity and 13% amino acid identity across the four species. Based on absence of conservation, *YBR184W* was first considered a biologically meaningless open reading frame. However, we know that the gene is expressed. The K_A/K_S ratio for *YBR184W* is 0.689, which is quite high compared to the average of 0.11 for other yeast genes.

This suggests that positive selection played a role in the evolution of *YBR184W*. Its protein product is likely involved in gamete function because it is similar in sequence to a yeast spore-specific protein.

Positive selection has been shown to occur after gene duplication. Comparing mouse and human genomes located 25 paralogous clusters specific to the mouse, each of which is orthologous to a single gene in human (132). The median K_A/K_S values for these clusters are higher than the median value for mouse-human orthologs. This may reflect positive selection acting on new gene copies to evolve new functions in the mouse. In accordance with previous findings, the two major functional themes among mouse-specific clusters are reproduction and immunity.

Another example comes from a study of 250 young ($K_S < 0.3$) duplicate gene pairs in the human genome (191). These genes are duplicated in human but not in mouse. When the mouse ortholog is used as an outgroup, the K_A/K_S ratios differ significantly between the two duplicates for 25% of human gene pairs. This indicates that they experienced different functional constraints after gene duplication. Remarkably, 45% of gene pairs (113 out of 250) had $K_A/K_S > 1$, which suggests positive selection possibly connected with acquiring a new function.

Mechanisms and History of Mammalian Evolution

Whole-genome comparisons offer an unprecedented opportunity to follow the processes and rates of evolution in detail. About 25% of the human genome has been generated since the last common ancestor with rodents; these are lineage-specific repeats, such as *Alu* repeats, formed by transposition only in primates (84). About 40% of the human genome aligns with the mouse genome, but as discussed above only about 5% will likely be under purifying selection (132). That leaves about 35% of the human genome that does not align with mouse (using current technologies), and another roughly 35% that is conserved (still present in both human and mouse) but may have no detectable function. This large amount of aligning genomic DNA provides many insights into genome evolution, as discussed in this and the next two sections. The nonaligning DNA is more difficult to interpret, but we summarize some of the current thinking.

INDELS AND THEIR IMPORTANCE IN DETERMINING DIFFERENCES IN GENOME SIZE

Whole-genome comparisons contribute to a current debate about what determines genome size. An earlier analysis of three insects (grasshoppers, crickets, and fruit flies) suggests that genome sizes are inversely proportional to the rates of DNA loss by small indels (e.g., 145). However, recently this conclusion was criticized as premature (66), in part because of the small size of the data set. The availability of several complete genomes allows us to more precisely estimate differences in indel rates and to evaluate whether these or other factors contribute to variation in genome size. For instance, the mouse genome is ∼14% smaller than the human genome, and the rates of small deletions are about two times higher in mouse than in human. A detailed investigation of complete genome sequences reveals that this bias contributes only a small amount (1% to 2%) to the differences in genome size. Most likely, a higher rate of large deletions explains the smaller size of the mouse genome (132). A similar pattern was recently observed in pufferfish (134). Spiny pufferfish have genomes twice as large as smooth puffers. Surprisingly, the rate of small deletions in spiny puffers is higher than in smooth puffers. Large insertions and deletions possibly related to transposable element activity might play a more important role than small deletions in determining genome size in pufferfish (134). Thus, although it is evident that indel biases may be important in specifying genome size, additional whole-genome comparisons are necessary to delineate the relative contribution of small and large indels in this process.

SEGMENTAL DUPLICATIONS A high frequency of large segmental duplications was one of the unexpected findings of the original analysis of the human genome (84). About 3.5% to 5% of the finished human sequence consists of recent segmental duplications, defined as duplicated sequences that are >1 kb in size and 90% to 99.5% identity (9, 29). In contrast, only about 1.2% of the mouse genome is in recent segmental duplications (28), whereas about 3% of the rat genome is in such duplications (148). Segmental duplications are enriched at the breaks in conserved synteny between the human and mouse (7). This suggests that segmental duplications and chromosome rearrangements are connected processes in genome evolution. Sixteen orthologous genes are involved in independent recent segmental duplications in both human and mouse genomes (28). One can speculate that duplication provides an evolutionary advantage for these genes in human and mouse, as in the case of olfactory receptor genes (189).

EFFECTS OF NEIGHBORING BASES Analysis of genomic alignments assists in elucidating mutation patterns on a large scale. One example is the dependence of nucleotide substitution rates on the identity of neighboring bases. Earlier ("pregenomic") studies suggested that nucleotide substitutions are context-dependent (e.g., 101). In particular, the pattern of transitions is dominated by the effect of CpG dinucleotides due to rapid deamination of methylated cytosines. This was recently confirmed by a comparison of noncoding (and presumably neutral) ancestral repeat sequences at the 1.8 Mb of the CFTR locus among eight eutherian mammals (165). The most pronounced context effect discovered in this study was transitions at CpG sites. Such large-scale analyses identified additional, more subtle context-dependent effects on nucleotide substitutions. Currently, these effects do not have simple explanations and will be more evident after whole-genome alignments are examined. Context effects are also evident in coding sequences. Significant context effects occur even across codon boundaries (165). Considering the effects of neighboring bases significantly improves the goodness of fit of nucleotide substitution models. This is particularly important because such models are used in a variety of applications, e.g., phylogenetic reconstruction (165).

Nonuniformity of Neutral Evolutionary Rates Within Species

REGIONAL VARIATION IN THE RATES OF DNA CHANGE Since the early days of protein sequencing, it has been apparent that some proteins change very little over evolutionary time and others change enormously. The range in rates of amino acid substitution is about 1000-fold, which is usually interpreted as a reflection of variation both in the portion of amino acids that is under purifying selection and the severity of that selection (reviewed in 183). However, the rate of change in nucleotides that are under little or no purifying selection also varies substantially for different loci. In this section, we summarize some early experiments that revealed this variation and then cover some of the new insights into this phenomenon based on whole-genome sequences. A recent review provides additional insights (45).

The nonsynonymous substitution rate varies over a range comparable to that seen for amino acid substitutions (reviewed in 135). In keeping with the neutral theory of evolution (94, 95), substitutions occur at a much higher fraction of synonymous sites than of nonsynonymous sites. However, the proportion of synonymous sites that have changed (K_S) is not homogeneous. K_S varies roughly tenfold among different genes compared between the same two species (112). This is considerably less than the several hundred- to thousand-fold variation in the proportion of nonsynonymous sites that have changed (K_A), but it does not fit with a homogeneous neutral rate across the genome. Graur (65) pointed out that synonymous and nonsynonymous substitution rates are significantly correlated. This was re-examined and confirmed by several authors, including Makalowski & Boguski (118), who stated, "No satisfactory explanation has been found for this phenomenon."

At least part of the explanation lies in variation in the neutral evolutionary rate in different parts of the genome. Wolfe et al. (186) showed that the synonymous substitution rate (a model for neutral evolution) varied for different genes, and it correlated with the base composition of the genes and flanking DNA. They concluded that the variation in both K_S and base composition could be explained by systematic differences in the rate and pattern of mutation over regions of the genome. Other studies showed that conserved synonymous sites are significantly higher in GC content (the fraction of bases that are guanine or cytosine) than expected for random substitutions, consistent with a lower substitution rate in chromosomal regions high in GC content, as reviewed by Bernardi (15). Matassi et al. (123) showed that the synonymous substitution rate is significantly more similar for neighboring genes than for randomly chosen genes, strongly arguing for regional differences in evolutionary rates. The studies summarized here differ in the extent to which GC content explains the variation.

Variation in evolutionary rates is also observed in comparisons of noncoding sequences in long genomic DNA segments. Because much of the repetitive DNA in a species may be lineage-specific, it is useful to measure the fraction of nonrepetitive, noncoding DNA in a locus that aligns between two species. An early study contrasted the extent of conservation in two loci-encoding proteins with virtually identical functions. Comparing human and rabbit genomic sequences, Hardison et al. (74) showed that intergenic sequences were much less similar in the *HBA* gene cluster (encoding alpha-globins) than in the *HBB* gene cluster (encoding beta-globins). As additional long genomic sequences were compared in species from different mammalian orders (usually human and mouse), it became clear that some loci have extensive matches outside the coding region (46, 51, 99, 141), some have matches largely limited to the coding region (50, 103), and others have an intermediate level of noncoding sequence matches (4, 102, 121, 163). Koop (98) notes that these strikingly different patterns reflect a mosaic structure of the mammalian genome, with different regions changing at significantly different rates. Quantitative analysis shows that the fraction of noncoding, nonrepetitive genomic sequence that aligns between mammalian orders varies at least tenfold among different loci (40, 50).

COVARIATION IN RATES OF SUBSTITUTION, RECOMBINATION, INSERTION, AND DELETION Not only are the neutral substitution rates variable within genomes, but they covary with recombination rates and the level of intraspecies polymorphism (106, 133). The positive correlation between levels of polymorphism and recombination was observed in earlier studies in *Drosophila* (13) and mammals (e.g., 174). Comparing several long genomic DNA sequences between human and mouse showed significant positive correlation between the frequency of insertion (monitored as density of repeated DNA) and divergence as measured by the fraction of human sequence that does not align with mouse (30). The latter measure is determined at least partially by the amount of deletion in mouse (132). Studies of 4.7 Mb of noncoding sequences aligned from human, chimpanzee, and baboon showed that the local substitution rate covaries positively along the separate human and chimpanzee branches (166). This revealed that the mutation rate varies deterministically across primate chromosomes, indicating that factors such as GC content and compositional nonequilibrium affect the mutation rate. One striking example of these differences in genomic context is shown in Figure 3, in which the highly conserved, gene-rich, GC-rich, repeat-poor Class III region of the *MHC* is adjacent to the much less conserved, gene-poor, GC-poor, repeat-rich Class II region.

This evidence pointed to an association in the rates by which DNA changes by several processes. One simple explanation for this is that the mutation rate, or the tendency of the DNA to change, varies regionally. However, studies comparing *Drosophila* species had shown that the rate of between-species divergence did not correlate with recombination rates, whereas within-species diversity did correlate (13). This was a strong argument against using variation in the neutral mutation rate to explain the observed association of diversity and recombination. Instead, the authors proposed that the association could be explained by genetic hitchhiking associated with the fixation of advantageous mutants. Alleles linked to such advantageous mutations would be fixed rapidly in the population, leading to lower polymorphism. The lower the recombination rate of a region, the further this homogenizing effect would extend around the advantageous mutation. Other studies of the variable amount of divergence in mammalian loci argued that this reflects different levels of selective constraint exerted over long genomic regions (162).

The availability of whole-genome alignments between human and mouse provided the opportunity to examine these issues comprehensively using an unprecedented number of likely neutral sites. Most earlier studies used K_S to estimate the neutral substitution rate, but some studies showed evidence of selection even on synonymous sites (e.g., 52), and some synonymous sites are parts of splicing enhancers. Scientists examined two different models of neutral DNA to estimate the neutral substitution rate (132). Within coding regions, they measured substitutions at fourfold degenerate (4D) sites. They also developed a second, novel model for neutral DNA. Substitutions in repetitive elements in the human genome that aligned with their orthologs in mouse (AR sites) were also measured. These repeats that predate the human-mouse divergence are relics of currently inactive transposons, and it is likely that most of their nucleotides have no function. As Ellegren et al. (45) reviewed, neither the 4D nor AR sites are perfect models of neutral DNA. However, the features that compromise each model are distinct, and the two models are independent of each other. Thus, as neutral rates are estimated for each type of sequence within regions of the genome, it is highly unlikely that a region with some 4D sites that are under selection (e.g., a splice enhancer) would also have AR sites that are under selection (e.g., as enhancers of transcription; 86). Both models are present in sufficiently large numbers (about 2 million 4D sites and 165 million AR sites in the human genome) so that the issues of rate variation could be examined robustly at high resolution. The estimates of the average neutral rate computed from 4D and AR sites genome-wide fall within the range of previous estimates.

The neutral substitution rate estimated from 4D and AR sites varied significantly across the genome, and the two estimated rates varied together (77, 132). These neutral rates covaried with the rate of insertion of LTR repeats, with an inferred rate of deletion in the mouse genome, and with the recombination and polymorphism rates in the human genome. GC content of the human genome affects these rates in a complex manner. In regions with low GC content, these rates of DNA change are negatively correlated with GC content, whereas in regions with high GC content, the correlation is positive. In these studies, the dependency on GC content is a confounding variable, which may help explain the differing results from previous studies that necessarily examined a subset of the genome (15, 59, 123, 186). When the variation due to GC content is removed from the analysis, the residuals still show a strong correlation, which argues that the covariation in these rates is influenced by additional features besides GC content. Although the high divergence of some genes between human and mouse is associated with the change in GC content between the orthologs (26), the change in GC content does not explain much of the rate variation genome-wide (77). Additional studies show that the covariation extends to estimates of neutral substitution rates and insertions of most classes of transposable elements in mouse-rat comparisons and human-rodent comparisons (187). The only exceptions are the families of SINE repeats, which have a strong regional preference to insert into GC-rich, more slowly changing DNA.

Examining more closely related species also reveals strong correlations among recombination, between-species divergence, and within-species diversity (79). Regions with low levels of recombination and diversity in humans also have reduced divergence in chimpanzee-baboon comparisons. Thus, mutation is associated statistically with recombination in humans, in accord with other studies (106). Importantly, at least in humans, the positive correlation between diversity and recombination may have a purely neutral explanation.

Chromosomes in *C. elegans* also show regional variation in the rates of evolutionary change (25, 167). The arms of the chromosomes have higher rates of recombination, more insertions of repetitive elements, and more breaks in conserved synteny (with respect to *C. briggsae*) than do the central regions. Sequence comparisons with *C. briggsae* show that genes in the arms also show a higher K_A/K_S ratio and a higher K_A value compared to the central regions. The Ks value is quite high (much higher than in mammalian comparisons) and shows marked local variability, making it difficult to assess whether this measure of the neutral rate is higher in the arms than in the centers of chromosomes. The central, more slowly changing regions of the worm chromosomes are enriched in genes with single orthologs in the comparison species and in genes that when mutated have a lethal phenotype (25). The arms of *C. elegans* chromosomes are evolving more rapidly than the central portions, consistent with the regional differences in evolutionary rates seen in mammalian chromosome comparisons. The essential, more highly conserved genes reside in the more stable central region, whereas genes without clear orthologs, i.e., those subject to more active "evolutionary experimentation," are more prevalent in the chromosome arms (167).

It is not clear why these studies within primates (79), between primates and rodents (77, 132), and between worm species (167) show correlations between recombination and divergence, whereas the earlier comparisons of *Drosophila* species do not (13). The availability of more comparison genomes will allow scientists to examine these correlations in various species to see which apply broadly and which are found only in some taxa.

In summary, different regions of mammalian genomes show substantial covariation in their inherent tendency to change by a variety of processes. This is true for the rates of nucleotide substitution (both within and between species), recombination, and insertion of all classes of transposable elements except SINEs. The GC content only partially explains the variation.

The factors explaining the rate variation need further study. For example, recombination (106) and initiation of replication (58) appear to be mutagenic, and these processes may be more frequent in the more rapidly changing regions. Perhaps the regional differences in evolutionary rates correspond to a long-distance organization of the genome. For example, housekeeping genes tend to be clustered in the human genome, and these genes tend to be more highly expressed (107). Over evolutionary time, as recombination tests many alternative locations for genes, it may be favorable for the housekeeping genes to be retained in more slowly changing regions. This type of process has been invoked to explain the tantalizing observation that particular classes of genes tend to be in fast-changing versus slowly changing regions of the human genome (31). It is reminiscent of the preferential localization on worm chromosomes of essential genes to the slowly changing central portions and less essential genes to the more rapidly changing arms (167). This model differs from those that maintain that selection is acting on all the DNA in slowly changing regions (162). One interpretation of the regional preference for particular classes of genes is that some regions are inherently slow to change, and it is advantageous for some types of genes to be in those regions.

The variation in evolutionary rates has a practical impact on functional genomics. As discussed above, it is critical to know in what type of segment a gene resides to understand how to correlate some observable parameters (such as level of conservation) with function. Also, problems arise because orthologous sequences from neutral regions are sometimes beyond the threshold for reliable alignment among distant mammals. If the alignment method attempts to avoid aligning regions that do not match well, using ancient repeats biases the study toward slow-evolving regions. Similarly, a study of neutral rates between mouse and rat made in regions that do not align with human will be biased toward fast-evolving segments. Another concern is that insofar as alignments include nonorthologous genomic positions, evolutionary rates will tend toward overestimation.

MUTATION RATE DIFFERENCES BETWEEN SEX CHROMOSOMES AND AUTOSOMES One manifestation of the variation in mutation rate within a mammalian genome is the difference in rates between autosomes and sex chromosomes and between the two sex chromosomes due to male mutation bias. This observation can be employed to test whether mutations result from errors in DNA replication. Males undergo more germline cell divisions than do females. Thus, if mutations are replication-driven, one expects to observe the highest mutation rate at a male-specific chromosome (Y), an intermediate rate at autosomes, and the lowest rate at X, because it spends most of its time in females (reviewed in 113). Whole-genome comparisons allows us to investigate the phenomenon of male-mutation bias on a large scale, so that mutation rates can be estimated from many loci, compensating for the effects of local (within-chromosome) variation. A comparative analysis of the human and mouse genomes indicates a lower substitution rate on chromosome X as compared to autosomes (132). However, the number of germline cell divisions is different between human and mouse, which prevents a direct test of the hypothesis about the role of replication errors in mutagenesis. The availability of the rat genome sequence provided the opportunity to rigorously test this hypothesis because mouse and rat are similar in generation time and in the number of germline cell divisions. Recently, mouse-rat mutation rates were compared between chromosome X and autosomes (119, 148). There was an approximately twofold excess of nucleotide substitutions originating in males over that in females, which confirmed earlier studies. Unexpectedly, small indels in rodents appear to be male-biased as well; the male-to-female indel rate ratio is ~2.3. This contrasts with the recent evolutionary study in primates (based on a substantially smaller data set) that indicated no sex bias in small indels (172). Thus, both small indels and nucleotide substitutions originate twice as frequently in male than in female rodents. The ratio in the number of cell divisions between the male and female germlines in mouse and rat is also ~2. This suggests that both small indels and nucleotide substitutions occur primarily during DNA replication.

Nonuniformity of Evolution Along the Branches of Phylogeny

Whole-genome comparisons provide a conclusive test of the molecular clock hypothesis, which postulates that the rate of evolution is approximately constant over time in all evolutionary lineages (192). It is now apparent that a global molecular clock does not exist, but rather the rates of neutral evolution are not always uniform along branches of the phylogenetic tree.

Comparing genomic sequences confirmed earlier observations inconsistent with a global molecular clock. For instance, lower rates in primates than in rodents were inferred from DNA hybridization data (97) and later from analyses of gene sequences (e.g., 97, 111). In contrast, other studies argued that the rate did not differ between the two lineages (e.g., 43). A comparison of median divergences of AR sites between mouse and human indicates a twofold-higher nucleotide substitution rate in the mouse lineage than in the human lineage (132), in agreement with an earlier estimate (111). This result was corroborated by comparing the AR sites among human, mouse, and dog (96). Although the rates in the dog and human lineages are similar (0.189 and 0.167 substitutions per site, respectively), the mouse lineage exhibits a noticeable rate increase (0.375 substitutions per site). Analysis of AR sites in human, mouse, and rat shows a threefold-faster rate on the rodent lineage than on the primate lineage for the rate of substitutions at likely neutral sites (148).

Accumulating sequence data provide an opportunity to reexamine the hominoid-slowdown hypothesis. Based on immunological data and protein sequence data, Goodman (61) and Goodman et al. (62) proposed that the rate of molecular evolution has slowed in hominoids (humans and apes) after separating from the Old World monkeys. Sarich & Wilson (156) later challenged this conclusion. However, a large-scale comparison of noncoding sequences available for primates (188) provides convincing evidence supporting the original hominoid-slowdown hypothesis and shows that the rate in the Old World monkey lineage is ~33% faster than in the human lineage.

Deviations from a global molecular clock are emerging in other species besides mammals. For instance, a comparative analysis of four yeast genomes (*S. cerevisiae*, *S. paradoxus*, and *S. mikatae* with *S. bayanus* as an outgroup) shows that the substitution rate is similar in the *S. cerevisiae* and *S. mikatae* lineages, but is ~67% lower in the *S. paradoxus* lineage (90).

Additional Caveat About Whole-Genome Alignments

The art and science of genome alignments are still in an exploratory period. Available alignments of vertebrate genomes are computed by significantly different pipelines (21, 36, 160). The basic approach of each pipeline is similar: A rapid search for similar regions is followed by extensions of these initial hits. However, the details of implementation are complex and different among the pipelines. Each program has a number of adjustable parameters that affect output. For instance, gaps within an alignment are penalized by amounts that can be adjusted, and correct settings are poorly understood. Each program has a gap-penalty tuning knob that is set by guesswork, which influences the trade-off between percentage of mismatches on one hand, and the frequency and size distribution of gaps on the other. Hence, it is not easy to predict differences in behavior of the pipelines, and the results of the different whole-genome alignments have not been compared and evaluated objectively. This should not discourage investigators from using the alignments, but users must realize that results may depend significantly on the details of how the alignments were produced.

LEARNING MORE FROM EXISTING DATA

Individual investigators have many options for using the available wealth of genome sequence data to answer questions that interest them. Here are some suggestions for how to proceed.

Choice of Species

It is often difficult to predict which pair of species will permit functional regions of a desired type to stand out in an interspecies comparison, as illustrated by the alpha and beta cardiac myosin heavy chain genes (129). The balance between heart rate and energy consumption in small mammals favors a preponderance of the alpha form in their ventricles, whereas large animals do better with beta. Consequently, the expression pattern in humans is closer to that of pigs, for instance, than that of mice (149). This suggests that a human-pig comparison might be better than human-mouse for locating segments that regulate expression of these genes. However, the two genes lie head-to-head in a slowly evolving region, suggesting that two mammals may be too similar for optimal separation of functional from nonfunctional segments. When combined, the two criteria may suggest that the human sequence be compared with sequence from a large bird or reptile.

A combination of slow neutral evolutionary rate and fast evolution of the function of interest may mean that no single pair of aligned genomic sequences is adequate to inform investigators about conservation and functional inferences. A solution is to align sequences from many species at a range of phylogenetic distances (Figure 2). Genome sequences from more mammalian species such as rat (148) are or will be available in the near future. It is likely that multiple comparison genomes will be sequenced for other vertebrates (e.g., birds and fish) and insects.

A pilot study of genomic sequences from 12 species, each orthologous to a human 1.8-Mb region including *CFTR*, illustrates the power of the additional sequences (175). For example, measures of conservation based on multiple aligned sequences had much greater power to detect highly conserved noncoding sequences compared to the measures based on pairwise human-mouse comparisons. An example of the high resolution of an alignment score derived from the multiple aligned sequences is the phyloHMMcons track (164) illustrated in Figure 3*A*. No pair of sequences resolved the likely functional sequences as well as larger combinations of sequences, and one of the most effective combinations was the set of sequences from nonprimate mammals plus human. Critical insights into phylogenetic history can be gleaned from examining recent and ancient transposable elements, e.g., confirming that rodents and primates are sister groups, despite the large amount of sequence divergence between them. Also, aligning multiple species at appropriate evolutionary distances provides information about the direction of the sequence alterations; the types of nucleotide substitutions can be resolved and insertions can be distinguished from deletions. Much more detailed analyses of the rates of evolutionary processes are possible with the multiple alignments. Over the next few years, we expect this to be a very active area as more genome sequences are determined.

Choice of Tools

BROWSERS Genome browsers have become essential tools of molecular and genetic research in the life sciences. Although only a few years old, they have become so widely used that it is difficult to imagine research or teaching without them. They quickly and easily provide organized views of extensive biological information. Excellent browsers and databases are available for almost all model organisms; early examples include the *Saccharomyces* database and Flybase. We comment mainly on the browsers for mammalian genomes and some related servers (Table 1). The resources available change rapidly; the current description is for the end of 2003.

Browsers are Internet-based tools that provide integrative views of extensive annotation of genes or genomic regions. The user controls the types and level of detail of annotation, and the interval displayed can range from an entire chromosome to a few nucleotides. Three major browsers provide a wide variety of annotation; they are the UCSC Genome Browser (87, 93), the Ensembl browser (32, 82), and the MapViewer at NCBI (180, 181). Figure 3 gives images from two of these. Typically, they present a diagram illustrating genetic markers, known genes and genes predicted by various pipelines, exon-intron structure and direction of transcription, mRNAs and ESTs, CpG islands, and repetitive elements. Users can select among the several types of gene predictions or distinguish known full-length mRNAs from ESTs and spliced ESTs. Other tracks common to all the browsers include information on the depth of coverage of the genome assembly for a region, clones that are sequenced, and positions of gaps in the assembly. Data supporting the browsers are housed in large databases, and the browser serves as a portal to view information on a single genomic interval. The browsers support bulk data downloads as well.

Despite these substantial similarities among the browsers, the methods of displaying the data differ substantially (Figure 3), and some information is present on one browser but not on others. A preference based on method of display is largely a matter of personal taste, but there are some differences among the major browsers that can be objectively distinguished. The Ensembl project has one of the most sophisticated gene prediction pipelines, and it applies this same pipeline to all genomes analyzed (82). The Ensembl viewer and supporting annotation are similarly gene-oriented, whereas the UCSC Browser and NCBI MapViewer are more oriented to chromosome intervals. The Ensembl display in Figure 3*B* includes considerable information about proteins. A server at UCSC, the Gene Family Browser, finds groups of proteins or genes related by protein-level homology, similarity of gene expression profiles, or proximity along the genomic DNA.

Comparative genomics information differs significantly between Ensembl and the UCSC Genome Browser. The genome comparisons at Ensembl are generated by a pipeline that tends to bring out sequence matches in and around genes, whereas programs that are more sensitive in noncoding and intergenic regions generate the comparative genomics data at the UCSC Browser. This is illustrated by a comparison of panels A and B in Figure 3. The gene-oriented alignments in Ensembl (Figure 3*B*) are only seen in the Class III region, which is the left part of the display, whereas the UCSC Browser (Figure 3*A*) also shows some alignments in the Class II region, which is the right part of the display. Also, the UCSC Browser currently has a greater variety of comparative genomics information. Tracks available at the UCSC Browser include not only alignment information, but analyses of the alignments such as the phyloHMMcons track (Figure 3*A*), which plots the posterior probability that an aligned sequence is among the most slowly changing regions (164).

Other distinctions among the major browsers are choices about which data to include. For example, as of this writing, both the UCSC Browser and Ensembl provide access to the large data set of microarray expression data generated by GNF (171), whereas MapViewer does not. In contrast, MapViewer is unique among the three in having SAGE tag information. Another distinctive feature is the ability to view user-generated tracks, which the UCSC Browser and Ensembl support (albeit in different formats).

Other browsers are designed for one major purpose, e.g., to show the results of whole-genome alignments. For example, the Berkeley whole-genome alignment pipeline (36) finds anchors of conserved synteny between the two genomes compared (e.g., human and mouse) using the rapid local aligner BLAT (91), and then it computes global alignments within these blocks of conserved synteny using AVID (19). The pipeline was modified for multiple species, with the multiple alignment constructed using a combination global and local aligner MLAGAN

(21). The VISTA server displays the results of the alignments and their analysis (36). Results of MAVID alignments of human, mouse, and rat genomes are accessible via the K-BROWSER (Table 1). The CORG (Comparative Regulatory Genomics) server finds conserved noncoding blocks upstream of orthologous gene pairs among human, mouse, rat, and Fugu (41).

GENOME DATABASES The browsers provide detailed views of single genome intervals. For deeper data mining, e.g., examining many loci at once, users need to query on the supporting data. The browsers provide some query capacity, such as Table View at the UCSC Browser (88). NCBI's *Entrez* system is a text-based search and retrieval system that accesses genome data plus information in many databases (158). Also, users can download bulk data and develop their own databases. However, less computationally inclined users can access the two online databases of genome information.

GALA is a database of genome alignments and annotation (60). It provides access to information on genes (known and predicted), gene functions, gene ontology (8), expression patterns, genome landscape (such as repetitive elements), genome alignments and various analyses of these alignments (emphasizing estimates of likelihood of selection and regulatory potential), and conserved transcription factor binding sites predicted by TRANSFAC weight matrices (185). The data are imported from various sources, including the UCSC Genome Browser and LocusLink at NCBI (147, 180). All data are entered as chromosomal intervals, with new builds for GALA for each species (currently human, mouse, and rat) and each new assembly. The schema consists of many tables of relational data, all referencing chromosomal intervals.

The initial query page supports simple queries to GALA, and the results can be combined in many ways on a history page. GALA also supports proximity and clustering in addition to typical operations such as union and subtraction. It can make intersections in different ways depending on the desired outcome. Results can be viewed as specialized tracks on the UCSC Genome Browser in an interactive online alignment viewer (184) or as tables of results. Histograms of results can also be plotted online. An example of a complex query supported by GALA is finding, in the vicinity of a set of genes expressed in a particular tissue, all the predicted binding sites for one or more transcription factors of interest that are both conserved in mammals and are also in a region with a high likelihood of being a regulatory element. This result could be obtained in five steps: three simple queries on expression pattern, the selected conserved transcription factor binding sites, and regulatory potential score, and two operations to join the results in the desired manner (e.g., intersection or proximity). Results from this query serve as predicted or hypothesized regulatory elements that the user can test experimentally.

EnsMart (89) is a branch of the Ensembl project that allows users to retrieve lists of biological objects, such as genes or single nucleotide polymorphisms (SNPs). Users can mine the data about genes, gene function, gene expression, and disease to find new insights and formulate hypotheses. EnsMart integrates data from Ensembl and several other resources. It creates a generic data system from the various data sources, transforming the data from the primary sources into the EnsMart database. The organization of the EnsMart database differs from that of GALA. EnsMart follows a "warehouse star-schema" with central biological objects (e.g., genes or SNPs) connected to a set of satellite tables, such as disease, transcript, and Protein FAMily (PFAM) attributes. Users navigate EnsMart in three steps: (*a*) selecting the species and focus (e.g., genes), (*b*) filtering the selection (e.g., genes) with specified features, and (*c*) outputting the data set with user-selected attributes. The output is a table of the results, for which users can specify any of a variety of formats.

The differences between GALA and EnsMart mimic some of the differences between the UCSC and the Ensembl browsers. One major difference is the depth of use of comparative genomic data. GALA was developed primarily as a way to access the various whole-genome alignments and scores that measure properties of the alignments (e.g., quality that indicates likelihood of selection, similarity of patterns to those in known functional regions) and extensive annotation. The breadth of genome coverage by alignments and the depth of analysis of those alignments are greater in the data accessible through GALA. Much of the data, and all the comparative genomics data, are available as tracks on the UCSC Browser, and one effective way to view the query results is on the UCSC Browser.

In contrast, EnsMart was developed with the Ensembl data and data from external sources. The object-based organization of EnsMart fits with the gene-centric views of the Ensembl browser. Thus, EnsMart provides users with fast and effective access to deep data in and around genes.

These genome-wide databases change rapidly both in their internal implementation and in the data sets recorded, as with genome browsers. Both have been public for only a short time, but both are changing quickly. The issue of how much the two databases overlap in the services offered and how much is distinctive to each is hard to answer, and the accuracy of any answer is transient. This makes evaluating performance even more difficult than usual (see below). Researchers who avail themselves of these resources should find both to be portals into a wealth of useful information.

DOWNLOADS OF SEQUENCE DATA AND ANNOTATIONS Query interfaces to databases reflect the types of issues that some set of users brought to the designers. There will be other issues that are difficult or impossible to address within the designed query system. In those cases, researchers will want to download data in bulk to their own customized databases for further analysis. The UCSC Browser, Ensembl, EnsMart, NCBI, and GALA support downloads of all data unless the original source placed limits on data distribution.

ALIGNMENT SERVERS FOR SIMILAR SPECIES Online servers that align two or more sequences from related species have been available for several years. The BLAST suite of tools (2, 3), implemented at NCBI, is the most frequently used aligner in the life sciences. One tool, BLAST2Sequences, rapidly aligns two sequences, but the length of the sequences is limited and the sensitivity is not high. Thus, servers were specially designed to align two or more long genomic sequences at high sensitivity while detecting common rearrangements such as duplications. Servers computing local alignments are PipMaker (161), MultiPipMaker (159), and zPicture (142). Those computing global alignments are VISTA (125) and MAVID (20). The theoretical and practical similarities and differences in these servers were recently reviewed (57). In principle, local alignments should be better able to find similar sequences despite rearrangements, and they can easily use draft, fragmented sequences as input, whereas global aligners should be more sensitive within regions of conserved gene order and orientation. However, the servers are designed to minimize these theoretical limitations. As is often the case for resources that overlap in the services offered (see below), no objective third party has evaluated the performance of these alignment servers. In situations where new alignments are needed, users will likely find all alignment types useful and possibly complementary.

The entire human, mouse, and rat genomes were aligned using the software that powers the servers listed above. For many loci and many comparisons, users can access the desired alignments from the genome browsers and databases. However, if users want to find matches to any sequence that is not in the genome assemblies, they can use the alignment servers. For example, experimental work on a particular locus may be done in a model organism whose genome is not being sequenced. Researchers may have sequence data on that locus from this special organism and want to add it to the comparative data from the sequenced model organisms. It is common for a particular laboratory to have specific information about a gene, such as the 5′ end of an mRNA derived from sequencing 5′-RACE products. The alignment servers can easily locate these sequences. Also, the alignment portals to the UCSC Browser, which uses BLAT (91), and to Ensembl, which uses SSAHA (140), can be used for this task.

Some alignment servers provide additional information that makes them particularly attractive for a given problem. For instance, seeking candidate CRMs in highly conserved noncoding sequences is strongly complemented by finding conserved matches to a transcription factor binding site. This capacity is implemented in rVista (115). A new server, zPicture (142), computes alignments with BLASTZ (160) and provides many enhanced features, as Figure 4 illustrates. Sequence input is greatly facilitated at this server. Designated intervals identified within the UCSC Genome Browser, annotation of genes and exons, and masking of repeats can be used as inputs. Displaying the resulting alignments can use either the PipMaker format (horizontal lines for each ungapped alignment segment) or the smoothed format implemented in Vista. One innovation of zPicture is the dynamic display of results so that annotations can be changed and the results seen quickly. Finally, the results can be output to rVista so that conserved transcription factor binding sites and conserved regions are visible. Alignment servers with such enhanced features will become the tools of choice for future analyses.

ALIGNMENT SERVERS FOR DISTANT SPECIES AND PARALOGS The alignment servers discussed above are tuned for closely related species, such as those from different mammalian orders. To look at more distant species, or to look at paralogs generated by ancient duplications (such as those that predate the mammalian radiation and hence are present in both mouse and human), different alignment methods are needed. One major application of these is to look for common motifs in the upstream regions of coexpressed genes. Two examples of these approaches are MEME (11) and Gibbs sampling (157, 176). A survey of these tools is beyond the scope of this review.

DOWNLOADING ALIGNMENT SOFTWARE The above-mentioned network resources are sometimes inappropriate for aligning user-supplied sequences, and a user may prefer to download alignment software and run it locally. One potential justification is that the amount of sequence may exceed what is feasible for a remote server to align. Second, the software used by an alignment server typically has capabilities that cannot be accessed through the server. For instance, the alignment program may permit arbitrary scores for nucleotide substitutions, whereas the server uses default scores; for genome sequences at certain evolutionary distances or with unusual nucleotide content, one may need to change the scores to obtain better alignments. Third, the user may be unwilling to send the sequence over the Internet. Finally, a user may believe that the best program for their needs is not available through a public server.

Alignment servers and genome browsers are natural starting places to look for genome alignment software. Availability of software and the conditions required for its use are in a constant state of flux, so we do not attempt to enumerate what is available here. Likewise, we expect that before press time, many of the available tools will be significantly improved or even replaced, so we are not attempting an evaluation. We recommend that once you identify software that seems to suit your needs, inform the authors of your particular requirements and ask for their advice. They may be strongly motivated to help you take advantage of their program's full capability.

DOWNLOADING WHOLE-GENOME ALIGNMENTS Existing databases readily answer certain questions about genome-wide phenomena, as described above. However, one can imagine interesting questions that are difficult to answer completely or precisely, such as: Is the average percent identity between human and mouse genomic DNA higher in introns than in intergenic regions? In general, are the 1000 bp immediately upstream of the transcription start site better conserved for genes with tissue-specific expression than for ubiquitously expressed genes? How does the degree of conservation compare for various classes of nontranslated RNA genes? To answer such questions, it may be feasible to download whole-genome alignments that were computed elsewhere, assemble any other necessary information, and write special-purpose programs that interrogate these data in an appropriate way. Although the computation of whole-genome alignments may require more computational resources than are available to a typical investigator, downloading and analyzing such alignments requires only a good Internet connection and an inexpensive computer.

Currently, several Internet sites exist that provide access to alignments of entire genomes, and others will appear soon. Some, and probably most, of these will permit users to download the entire set of alignments, along with annotations such as gene and exon positions. These can be analyzed for novel purposes using programs that, at least in some cases, are relatively straightforward to write. For instance, a student with good programming skills should be able to write a program that reads alignments of the human and mouse genomes, reads positions of all human RefSeq genes, and determines average levels of substitutions in alignments of introns and of intergenic regions. The time to write the program might fall between one day and one week, depending on programming proficiency and luck, whereas running the program on the full set of alignments might take one or several hours on a typical workstation. Software to perform more complicated analyses of alignments and other data will take longer to write.

FUTURE OF COMPARATIVE GENOMICS

The next several years will bring many advances in comparative genomics. Genomes from a wide variety of species covering many taxa will be sequenced. This wide range of genomes will insure that the methods of comparative genomics will be applied to basic and complex issues in plant biology, developmental biology, pathobiology, behavior, and more. It would be presumptuous to speculate on the novel insights that will be gleaned, but previous experience with the several genomes sequenced to date makes us confident that many exciting new findings will be forthcoming.

We expect that the resources for comparative genomics will become even more user friendly, and that they will become part of the toolkit of virtually every experimental biologist. However, building the bioinformatic infrastructure to realize this exciting potential will require new developments. We summarize some of them here.

A small cadre of bioinformatics specialists has taken up the task of building better tools for producing whole-genome alignments. In parallel, a community of computationally oriented biologists has coalesced to use those alignments for various purposes. The two fields are coevolving: better alignment tools engender an expanded range of successful applications, and new uses for alignments suggest additions to the toolset. Although neither field has matured to the point that its trajectory can be reliably forecast in detail, we suggest several general steps to improve the quality of tools to compare genome sequences, as was previously attempted (42, 130).

Users of alignment tools and/or precomputed alignments may benefit from our discussion through a heightened awareness of areas in which current tools lack maturity. Take-home messages include that the field of genome sequence comparisons is still in its infancy, and that new tools and ideas should be greeted with a healthy skepticism. On the bright side, the continuing influx of experts from related fields will quickly improve the situation, particularly as the following goals are achieved.

Precise and Comprehensive Formulations of the Genome-Comparison Problem

The traditional concept of an alignment is intuitively natural and admits a straightforward and precise definition. In fact, many somewhat different definitions are straightforward, and to even approach the generality needed for whole-genome comparisons one must go beyond the traditional concept. For example, comparing genomes might be construed as the problem of computing an "appropriate" set of objects, each of which is constructed by the following steps: (*a*) select one or more genomes; (*b*) select one contiguous segment from each of the selected genomes; (*c*) replace zero or more of the segments by their reverse complement; (*d*) add zero or more dash characters to each selected segment so that the resulting padded segments have the same length; (*e*) place the padded sequences one over the other in an array, where each padded sequence forms a row, with one symbol (letter or dash) per column; and (*f*) discard all columns that consist entirely of dashes. "Appropriate" might be interpreted as requiring that, for the most part, two sequence positions are in the same column of an alignment if and only if they are descended (each by zero or more nucleotide substitutions) from the same position in the most recent common ancestor of the two species.

Developers and users of whole-genome alignment software have yet to agree on, or even to discuss in any detail, what it is that the software should attempt to compute. With maturation of the field of whole-genome sequence comparison, it becomes both more critical and more feasible to precisely formulate the genome comparison problem (or problems). Part of the task is to precisely specify the objects to be computed; early attempts to appropriately generalize the classical concept of an alignment (17, 78, 104) need improvement. Another part of the task is to formulate the biological criteria that the computed object should satisfy, i.e., to state how to tell if the computed object is accurate.

Alignment Software that Automatically and Accurately Handles a Wider Spectrum of Evolutionary Operations

The above definition of "alignment" models just a few evolutionary operations, namely insertions and deletions (both short by adding dashes and long by omitting species), inversions, and nucleotide substitutions. It is still more inclusive than the classes of similarities handled by many (probably most) current whole-genome alignment programs. In particular, it is common for a genome alignment program to be designed so that it will fail to record many inversions, particularly small ones. This is unfortunate, given the frequency of inversions within genomic DNA sequences (92) and the fact that functional regions can be inverted in one species relative to another. For instance, a 110-bp regulatory region lying over 4 kb downstream of the human *WNT1* gene (152) can be identified in whole-genome human-Fugu alignments computed by existing methods, provided the method accounts for the region's inversion between those species. Duplication events provide additional challenges for the would-be builder of whole-genome alignment tools. The currently available genome alignment pipelines fall short of accurately handling inversions and duplications, much less other kinds of rearrangements. Though the bioinformatics community has been painfully aware of the deficiency for several years (42), a completely satisfactory solution remains to be found.

Better Tools for Identifying Well-Conserved Regions within Long Alignments

The mammalian radiation was sufficiently recent such that mammalian genomes can frequently be aligned in neutrally evolving regions with some confidence. For example, about 40% of the human genome can be aligned to the mouse, a much larger fraction than what appears to be under evolutionary constraints (132). A major potential use for these alignments is to predict the locations of genomic segments that are functional (i.e., increase evolutionary fitness), hence the goal of finding the segments that are particularly well conserved within a pre-existing alignment. Several algorithms have been proposed for this problem (47, 81, 122, 151, 168). However, the problem is deeper than simply providing more methods. One need is for a system that lets the user readily select from among several methods, adjust parameters, pick which subset of the given species to consider, and see the computed regions displayed in the context of rich biological information. Another is for an objective evaluation of the existing tools, a need common to many classes of bioinformatics software.

Improved Methods to Evaluate Genome-Alignment Software

The sudden availability of huge amounts of genomic sequence data has fostered a "gold rush" mentality in a segment of the bioinformatics community, with several groups seeking the mother lode—an alignment program that will become the *de facto* standard for genome comparisons. Serious attempts to evaluate these tools are frequently left behind in the rush. Such an evaluation project is unlike the mundane task of setting up a hardware store to serve the gold miners because the job is more challenging (not to mention more useful to the biology community) than that of writing yet another alignment program. It is an order of magnitude easier to build two good genome-alignment programs than to tell which one is better.

One approach to evaluating genome alignment software is to measure the accuracy with which orthologous protein-coding regions are aligned. This has the advantages that accuracy is measured with respect to an important class of biologically functional regions and that those regions are relatively straightforward to locate by current experimental means. For DNA-based alignment software, it is a serious disadvantage that the methods evaluated are not primarily designed for matching coding regions, a job better done by programs that work with conceptual translations of the genomic sequence.

A more appropriate strategy is to measure efficacy at correctly matching functional noncoding regions, using naturally occurring DNA sequences and experimentally confirmed functional elements. The approach can apply to programs that either compute alignments or find well-conserved regions within existing alignments (168). Thus, programs are evaluated on a data set that is germane to their intended purpose. Unfortunately, it is currently difficult to obtain large data sets of confirmed elements to serve as the "gold standard," and it is not always clear how to quantify a program's success at finding the right answer (55). It is particularly problematic to experimentally verify that a genomic segment is completely nonfunctional (in all tissues, developmental stages, and environmental conditions), which makes it difficult to measure false positives among computationally predicted functional segments.

Besides predicting the location of functional noncoding segments, a second major application of genome-comparison programs is to investigate neutral evolution. In that context, success requires that a large majority of the aligned pairs of sequence positions are orthologous, meaning the two positions descend from the same position in the genome of the last common ancestral species, possibly via nucleotide substitutions. One way to quantify a program's ability to correctly align orthologous positions in neutrally evolving DNA is to simulate evolutionary processes (17, 169). The basic idea is to start with a hypothetical "ancestral" sequence and apply a realistic set of synthesized mutational operations and speciation events, resulting in an artificial set of "modern day" sequences; one then knows which pairs of sequence positions were obtained from the same ancestral position, i.e., one knows the correct output.

It is natural for bioinformatics specialists to prefer writing new programs to testing old ones. The burden may fall on editors and reviewers of biology and bioinformatics journals to require that published software papers provide compelling evidence of the new programs' superior performance for an important class of data. Also, publishing objective software evaluations by groups not heavily invested in any of the programs being evaluated, perhaps along the lines of Fortna & Gardiner's (56) recent survey of genome sequence analysis tools, should be strongly encouraged.

Improved Tools for Linking Alignments to Other Sequence-Based Information

Experimental data on functional, noncoding genomic regions, such as DNase-hypersensitive sites, protein-binding, assays and DNA-transfer experiments using cells or animals, needs to be better integrated with sequence data. A substantial body of data from traditional low- to medium-throughput experiments is in the published literature but not in databases. A major new initiative, the ENCODE project (35), started in the fall of 2003 to identify all functions of the DNA sequences in targets that cover 1% of the human genome. This initiative will generate large amounts of high-throughput functional data, and it is stimulating the development of additional high-throughput methods. The next phase is to apply the most effective methods to the entire human genome.

Data from the ENCODE project will be a testing ground to develop bioinformatic resources to display the experimental results and to integrate them with other sequence-based information. Figure 5 shows an early prototype with a coordinated view of hypersensitive sites and known regulatory regions with patterns of conservation at the *CFTR* locus. The DNA segments that have been investigated range from highly to moderately conserved. Some extremely well-conserved regions that also have conserved predicted transcription factor binding sites, such as the peak in the phyloHMMcons track at the far left, are tantalizing segments for experimental tests by a single investigator. The ENCODE data should include comprehensive tests of all segments, which will permit a critical evaluation of how well various measures, or combinations of them, predict function. As the ENCODE data are recorded and displayed, it will be desirable to harvest the vast amount of functional data already published and organize it into databases. It is reasonable to expect substantial advances in this area in the future.

Sequences and sequence alignments will continue to be integrated with functional genomic data in novel ways. One recent advance was the integration of microarray expression data from several different species guided by orthology relationships among the genes (170). This meta-analysis led to important new insights and should be considered an early effort in the large-scale integration efforts to come. Such advances will fuel efforts that lead to novel information and deeper understanding of biological processes.

The *Annual Review of Genomics and Human Genetics* is online at http://genom.annualreviews.org

LITERATURE CITED

1. Alexandersson M, Cawley S, Pachter L. 2003. SLAM: cross-species gene finding and alignment with a generalized pair hidden Markov model. *Genome. Res.* 13:496–502
2. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215:403–10
3. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* 25:3389–402
4. Ansari-Lari MA, Oeltjen JC, Schwartz S, Zhang Z, Muzny DM, et al. 1998. Comparative sequence analysis of a gene-rich cluster at human chromosome 12p13 and its syntenic region in mouse chromosome 6. *Genome. Res.* 8:29–40
5. Antequera FB, Bird A. 1993. Number of CpG islands and genes in human and mouse. *Proc. Natl. Acad. Sci. USA* 90:11995–99
6. Aparicio S, Morrison A, Gould A, Gilthorpe J, Chaudhuri C, et al. 1995. Detecting conserved regulatory elements with the model genome of the Japanese puffer fish, *Fugu rubripes*. *Proc. Natl. Acad. Sci. USA* 92:1684–88
7. Armengol L, Pujana MA, Cheung J, Scherer SW, Estivill X. 2003. Enrichment of segmental duplications in regions of breaks of synteny between the human and mouse genomes suggest their involvement in evolutionary rearrangements. *Hum. Mol. Genet.* 12:2201–8
8. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. 2000. Gene ontology: tool for the unification of biology. *Nat. Genet.* 25:25–29
9. Bailey JA, Gu Z, Clark RA, Reinert K,

Samonte RV, et al. 2002. Recent segmental duplications in the human genome. *Science* 297:1003–7

10. Bailey JA, Yavor AM, Viggiano L, Misceo D, Horvath JE, et al. 2002. Human-specific duplication and mosaic transcripts: the recent paralogous structure of chromosome 22. *Am. J. Hum. Genet.* 70:83–100
11. Bailey TL, Elkan C. 1995. The value of prior knowledge in discovering motifs with MEME. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 3:21–29
12. Batzoglou S, Pachter L, Mesirov JP, Berger B, Lander ES. 2000. Human and mouse gene structure: comparative analysis and application to exon prediction. *Genome. Res.* 10:950–58
13. Begun DJ, Aquadro CF. 1992. Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature* 356:519–20
14. Berman BP, Nibu Y, Pfeiffer BD, Tomancak P, Celniker SE, et al. 2002. Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. *Proc. Natl. Acad. Sci. USA* 99:757–62
15. Bernardi G. 1995. The human genome: organization and evolutionary history. *Annu. Rev. Genet.* 29:445–76
16. Bird AP. 1986. CpG-rich islands and the function of DNA methylation. *Nature* 321:209–13
17. Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AFA, et al. 2004. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome. Res.* 14:708–15
18. Boffelli D, McAuliffe J, Ovcharenko D, Lewis KD, Ovcharenko I, et al. 2003. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* 299:1391–94
19. Bray N, Dubchak I, Pachter L. 2003. AVID: A global alignment program. *Genome. Res.* 13:97–102
20. Bray N, Pachter L. 2003. MAVID multiple alignment server. *Nucl. Acids Res.* 31:3525–26
21. Brudno M, Do CB, Cooper GM, Kim MF, Davydov E, et al. 2003. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome. Res.* 13:721–31
22. Bulger M, Bender MA, von Doorninck JH, Wertman B, Farrell C, et al. 2000. Comparative structural and functional analysis of the olfactory receptor genes flanking the human and mouse β-globin gene clusters. *Proc. Natl. Acad. Sci. USA* 97:14560–65
23. Burge C. 1997. *Identification of genes in human genomic DNA*. PhD thesis. Stanford Univ., Stanford, Calif.
24. Burge C, Karlin S. 1997. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* 268:78–94
25. *C. elegans* Sequencing Consortium. 1998. Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science* 282:2012–18
26. Castresana J. 2002. Genes on human chromosome 19 show extreme divergence from the mouse orthologs and a high GC content. *Nucl. Acids Res.* 30:1751–56
27. Chapman MA, Charchar FJ, Kinston S, Bird CP, Grafham D, et al. 2003. Comparative and functional analyses of LYL1 loci establish marsupial sequences as a model for phylogenetic footprinting. *Genomics* 81:249–59
28. Cheung J, Wilson MD, Zhang J, Khaja R, MacDonald JR, et al. 2003. Recent segmental and gene duplications in the mouse genome. *Genome. Biol.* 4:R47
29. Cheung VG, Nowak N, Jang W, Kirsch IR, Zhao S, et al. 2001. Integration of cytogenetic landmarks into the draft sequence of the human genome. *Nature* 409:953–58
30. Chiaromonte F, Yang S, Elnitski L, Yap V, Miller W, Hardison RC. 2001. Association between divergence and

interspersed repeats in mammalian noncoding genomic DNA. *Proc. Natl. Acad. Sci. USA* 98:14503–8

31. Chuang J, Li H. 2004. Functional bias and spatial organization of genes in mutational hot and cold regions in the human genome. *Public Libr. Sci. Biol.* 2:253–63.
32. Clamp M, Andrews D, Barker D, Bevan P, Cameron G, et al. 2003. Ensembl 2002: accommodating comparative genomics. *Nucl. Acids Res.* 31:38–42
33. Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B, et al. 2003. Finding functional features in Saccharomyces genomes by phylogenetic footprinting. *Science* 301:71–76
34. Cliften PF, Hillier LW, Fulton L, Graves T, Miner T, et al. 2001. Surveying Saccharomyces genomes to identify functional elements by comparative DNA sequence analysis. *Genome. Res.* 11:1175–86
35. Collins FS, Green ED, Guttmacher AE, Guyer MS. 2003. A vision for the future of genomics research. *Nature* 422:835–47
36. Couronne O, Poliakov A, Bray N, Ishkhanov T, Ryaboy D, et al. 2003. Strategies and tools for whole-genome alignments. *Genome. Res.* 13:73–80
37. Crollius HR, Jaillon O, Bernot A, Dasilva C, Bouneau L, et al. 2000. Estimate of human gene number provided by genome-wide analysis using Tetraodon nigroviridis DNA sequence. *Nat. Genet.* 25:235–38
38. Dermitzakis E, Clark A. 2002. Evolution of transcription factor binding sites in mammalian gene regulatory regions: conservation and turnover. *Mol. Biol. Evol.* 19:1114–21
39. Dermitzakis ET, Reymond A, Scamuffa N, Ucla C, Kirkness E, et al. 2003. Evolutionary discrimination of mammalian conserved non-genic sequences (CNGs). *Science* 302:1033–35
40. DeSilva U, Elnitski L, Idol JR, Doyle JL, Gan W, et al. 2002. Generation and comparative analysis of approximately 3.3 Mb of mouse genomic sequence orthologous to the region of human chromosome 7q11.23 implicated in Williams syndrome. *Genome. Res.* 12:3–15
41. Dieterich C, Wang H, Rateitschak K, Luz H, Vingron M. 2003. CORG: a database for comparative regulatory genomics. *Nucl. Acids Res.* 31:55–57
42. Dubchak I, Pachter L. 2002. The computational challenges of applying comparative-based computational methods to whole genomes. *Brief Bioinform.* 3:18–22
43. Easteal S, Collet C. 1994. Consistent variation in amino-acid substitution rate, despite uniformity of mutation rate: protein evolution in mammals is not neutral. *Mol. Biol. Evol.* 11:643–47
44. Efstratiadis A, Posakony JW, Maniatis T, Lawn RM, O'Connell C, et al. 1980. The structure and evolution of the human β-globin gene family. *Cell* 21:653–68
45. Ellegren H, Smith NG, Webster MT. 2003. Mutation rate variation in the mammalian genome. *Curr. Opin. Genet. Dev.* 13:562–68
46. Ellsworth RE, Jamison DC, Touchman JW, Chissoe SL, Braden Maduro VV, et al. 2000. Comparative genomic sequence analysis of the human and mouse cystic fibrosis transmembrane conductance regulator genes. *Proc. Natl. Acad. Sci. USA* 97:1172–77
47. Elnitski L, Hardison RC, Li J, Yang S, Kolbe D, et al. 2003. Distinguishing regulatory DNA from neutral sites. *Genome. Res.* 13:64–72
48. Elnitski L, Miller W, Hardison R. 1997. Conserved E boxes function as part of the enhancer in hypersensitive site 2 of the β-globin locus control region: role of basic helix-loop-helix proteins. *J. Biol. Chem.* 272:369–78
49. Emorine L, Kuehl M, Weir L, Leder P, Max EE. 1983. A conserved sequence in the immunoglobulin Jκ-Cκ intron:

possible enhancer element. *Nature* 304: 447–49

50. Endrizzi M, Huang S, Scharf JM, Kelter AR, Wirth B, et al. 1999. Comparative sequence analysis of the mouse and human Lgn1/SMA interval. *Genomics* 60:137–51
51. Epp TA, Wang R, Sole MJ, Liew CC. 1995. Concerted evolution of mammalian cardiac myosin heavy chain genes. *J. Mol. Evol.* 41:284–92
52. Eyre-Walker A. 1999. Evidence of selection on silent site base composition in mammals: potential implications for the evolution of isochores and junk DNA. *Genetics* 152:675–83
53. Fickett JW, Tung CS. 1992. Assessment of protein coding measures. *Nucl. Acids Res.* 20:6441–50
54. Flint J, Tufarelli C, Peden J, Clark K, Daniels RJ, et al. 2001. Comparative genome analysis delimits a chromosomal domain and identifies key regulatory elements in the alpha globin cluster. *Hum. Mol. Genet.* 10:371–82
55. Florea L, Li M, Riemer C, Giardine B, Miller W, Hardison R. 2000. Validating computer programs for functional genomics in gene regulatory regions. *Curr. Genom.* 1:11–27
56. Fortna A, Gardiner K. 2001. Genomic sequence analysis tools: a user's guide. *Trends Genet.* 17:158–64
57. Frazer KA, Elnitski L, Church D, Dubchak I, Hardison RC. 2003. Cross-species sequence comparisons: a review of methods and available resources. *Genome. Res.* 13:1–12
58. Fullerton SM, Bond J, Schneider JA, Hamilton B, Harding RM, et al. 2000. Polymorphism and divergence in the beta-globin replication origin initiation region. *Mol. Biol. Evol.* 17:179–88
59. Fullerton SM, Carvalho AB, Clark AG. 2001. Local rates of recombination are positively correlated with GC content in the human genome. *Mol. Biol. Evol.* 18:1139–42
60. Giardine BM, Elnitski L, Riemer C, Makalowska I, Schwartz S, et al. 2003. GALA, a database for genomic sequence alignments and annotations. *Genome. Res.* 13:732–41
61. Goodman M. 1961. The role of immunochemical differences in the phyletic development of human behavior. *Hum. Biol.* 33:131–62
62. Goodman M, Barnabas J, Matsuda G, Moore GW. 1971. Molecular evolution in the descent of man. *Nature* 233:604–13
63. Gottgens B, Barton LM, Gilbert JG, Bench AJ, Sanchez MJ, et al. 2000. Analysis of vertebrate SCL loci identifies conserved enhancers. *Nat. Biotechnol.* 18:181–86
64. Gottgens B, Gilbert JG, Barton LM, Grafham D, Rogers J, et al. 2001. Long-range comparison of human and mouse SCL loci: localized regions of sensitivity to restriction endonucleases correspond precisely with peaks of conserved noncoding sequences. *Genome. Res.* 11:87–97
65. Graur D. 1985. Amino acid composition and the evolutionary rates of protein-coding genes. *J. Mol. Evol.* 22:53–62
66. Gregory TR. 2003. Is small indel bias a determinant of genome size? *Trends Genet.* 19:485–88
67. Grosveld F, van Assendelft GB, Greaves D, Kollias G. 1987. Position-independent, high-level expression of the human β-globin gene in transgenic mice. *Cell* 51:975–85
68. Guigo R. 1998. Assembling genes from predicted exons in linear time with dynamic programming. *J. Comput. Biol.* 5: 681–702
69. Guigo R, Agarwal P, Abril JF, Burset M, Fickett JW. 2000. An assessment of gene prediction accuracy in large DNA sequences. *Genome. Res.* 10:1631–42
70. Gumucio D, Shelton D, Zhu W, Millinoff D, Gray T, et al. 1996. Evolutionary strategies for the elucidation of *cis* and

trans factors that regulate the developmental switching programs of the beta-like globin genes. *Mol. Phylogent. Evol.* 5:18–32

71. Gumucio DL, Blanchard-McQuate KL, Heilstedt-Williamson H, Tagle D, Gray TA, et al. 1991. γ-Globin gene regulation: evolutionary approaches. In *The Regulation of Hemoglobin Switching*, ed. G Stamatoyannopoulos, AW Nienhuis, pp. 277–89. Baltimore, MD: Johns Hopkins Univ. Press
72. Gumucio DL, Heilstedt-Williamson H, Gray TA, Tarle SA, Shelton DA, et al. 1992. Phylogenetic footprinting reveals a nuclear protein which binds to silencer sequences in the human γ and ε globin genes. *Mol. Cell. Biol.* 12:4919–29
73. Gumucio DL, Shelton DA, Bailey WJ, Slightom JL, Goodman M. 1993. Phylogenetic footprinting reveals unexpected complexity in *trans* factor binding upstream from the ε-globin gene. *Proc. Natl. Acad. Sci. USA* 90:6018–22
74. Hardison R, Krane D, Vandenbergh D, Cheng J-F, Mansberger J, et al. 1991. Sequence and comparative analysis of the rabbit α-like globin gene cluster reveals a rapid mode of evolution in a G+C rich region of mammalian genomes. *J. Mol. Biol.* 222:233–49
75. Hardison R, Oeltjen J, Miller W. 1997. Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome. *Genome. Res.* 7:959–66
76. Hardison RC, Chiaromonte F, Kolbe D, Wang H, Petrykowska H, et al. 2004. Global predictions and tests of erythroid regulatory regions. In *The Genome of Homo sapiens*. Cold Spring Harbor, NY: Cold Spring Harbor
77. Hardison RC, Roskin KM, Yang S, Diekhans M, Kent WJ, et al. 2003. Covariation in frequencies of substitution, deletion, transposition and recombination during eutherian evolution. *Genome. Res.* 13:13–26
78. Hein J. 1989. A new method that simultaneously aligns and reconstructs ancestral sequences for any number of homologous sequences, when the phylogeny is given. *Mol. Biol. Evol.* 6:649–68
79. Hellmann I, Ebersberger I, Ptak SE, Paabo S, Przeworski M. 2003. A neutral explanation for the correlation of diversity with recombination rates in humans. *Am. J. Hum. Genet.* 72:1527–35
80. Higgs D, Wood W, Jarman A, Sharpe J, Lida J, et al. 1990. A major positive regulatory region located far upstream of the human α-globin gene locus. *Genes Dev.* 4:1588–601
81. Huang X, Pevzner P, Miller W. 1994. Parametric recomputing in alignment graphs. In *Combinatorial Pattern Matching*, pp. 87–101. Springer Lecture Notes in Computer Science, 807. Springer-Verlag
82. Hubbard T, Barker D, Birney E, Cameron G, Chen Y, et al. 2002. The Ensembl genome database project. *Nucl. Acids Res.* 30:38–41
83. Hughes AL, Nei M. 1988. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* 335:167–70
84. International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* 409:860–921
85. Jane SM, Ney PA, Vanin EF, Gumucio DL, Nienhuis AW. 1992. Identification of a stage selector element in the human γ-globin gene promoter that fosters preferential interaction with the 5′ HS2 enhancer when in competition with the β-promoter. *EMBO J.* 11:2961–69
86. Jordan IK, Rogozin IB, Glazko GV, Koonin EV. 2003. Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends Genet.* 19:68–72
87. Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, et al. 2003. The

UCSC Genome Browser Database. *Nucl. Acids Res.* 31:51–54

88. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, et al. 2004. The UCSC Table Browser data retrieval tool. *Nucl. Acids Res.* 32:D493–96
89. Kasprzyk A, Keefe D, Smedley D, London D, Spooner W, et al. 2003. EnsMart—a generic system for fast and flexible access to biological data. *Genome Res.* 14:160–69
90. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 423:241–54
91. Kent WJ. 2002. BLAT–the BLAST-like alignment tool. *Genome. Res.* 12:656–64
92. Kent WJ, 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. USA* 100:11484–89
93. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, et al. 2002. The human genome browser at UCSC. *Genome. Res.* 12:996–1006
94. Kimura M. 1968. Evolutionary rate at the molecular level. *Nature* 217:624–26
95. Kimura M. 1977. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature* 267:275–76
96. Kirkness EF, Bafna V, Halpern AL, Levy S, Remington K, et al. 2003. The dog genome: survey sequencing and comparative analysis. *Science* 301:1898–903
97. Kohne DE. 1970. Evolution of higher-organism DNA. *Q. Rev. Biophys.* 3:327–75
98. Koop BF. 1995. Human and rodent sequence comparisons: a mosaic model of genomic evolution. *Trends Genet.* 11: 367–71
99. Koop BF, Hood L. 1994. Striking sequence similarity over almost 100 kilobases of human and mouse T-cell receptor DNA. *Nat. Genet.* 7:48–53
100. Korf I, Flicek P, Duan D, Brent MR. 2001. Integrating genomic homology into gene structure prediction. *Bioinformatics* 17 Suppl 1:S140–48
101. Krawczak M, Ball EV, Cooper DN. 1998. Neighboring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes. *Am. J. Hum. Genet.* 63:474–88
102. Lamerdin JE, Montgomery MA, Stilwagen SA, Scheidecker LK, Tebbs RS, et al. 1995. Genomic sequence comparison of the human and mouse XRCC1 DNA repair gene regions. *Genomics* 25:547–54
103. Lamerdin JE, Stilwagen SA, Ramirez MH, Stubbs L, Carrano AV. 1996. Sequence analysis of the ERCC2 gene regions in human, mouse, and hamster reveals three linked genes. *Genomics* 34:399–409
104. Lee C, Grasso C, Sharlow MF. 2002. Multiple sequence alignment using partial order graphs. *Bioinformatics* 18: 452–64
105. Lee YH, Ota T, Vacquier VD. 1995. Positive selection is a general phenomenon in the evolution of abalone sperm lysin. *Mol. Biol. Evol.* 12:231–38
106. Lercher MJ, Hurst LD. 2002. Human SNP variability and mutation rate are higher in regions of high recombination. *Trends Genet.* 18:337–40
107. Lercher MJ, Urrutia AO, Hurst LD. 2002. Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nat. Genet.* 31:180–83
108. Li J, Miller W. 2003. Significance of interspecies matches when evolutionary rate varies. *J. Comput. Biol.* 10:537–54
109. Li W-H. 1997. *Molecular Evolution*. Sunderland, MA: Sinauer
110. Li W-H, Gouy M, Sharp P, O'hUigin C, Yang Y-W. 1990. Molecular phylogeny of rodentia, lagomorpha, primates,

artiodactyla and carnivora and molecular clocks. *Proc. Natl. Acad. Sci. USA* 87:6703–7

111. Li WH, Wu CI. 1987. Rates of nucleotide substitution are evidently higher in rodents than in man. *Mol. Biol. Evol.* 4:74–82
112. Li WH, Wu CI, Luo CC. 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol. Biol. Evol.* 2:150–74
113. Li WH, Yi S, Makova K. 2002. Male-driven evolution. *Curr. Opin. Genet. Dev.* 12:650–56
114. Loots GG, Locksley RM, Blankespoor CM, Wang ZE, Miller W, et al. 2000. Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* 288:136–40
115. Loots GG, Ovcharenko I, Pachter L, Dubchak I, Rubin EM. 2002. rVista for comparative sequence-based discovery of functional transcription factor binding sites. *Genome. Res.* 12:832–39
116. Ludwig MZ, Patel NH, Kreitman M. 1998. Functional analysis of eve stripe 2 enhancer evolution in *Drosophila*: rules governing conservation and change. *Development* 125:949–58
117. Lukashin AV, Borodovsky M. 1998. GeneMark.hmm: new solutions for gene finding. *Nucl. Acids Res.* 26:1107–15
118. Makalowski W, Boguski MS. 1998. Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences. *Proc. Natl. Acad. Sci. USA* 95:9407–12
119. Makova K, Yang S, Chiaromonte F. 2004. Small indels are male-biased too: a whole-genome analysis in rodents. *Genome. Res.* 14:567–73
120. Makova KD, Ramsay M, Jenkins T, Li WH. 2001. Human DNA sequence variation in a 6.6-kb region containing the melanocortin 1 receptor promoter. *Genetics* 158:1253–68
121. Margot JB, Demers GW, Hardison RC. 1989. Complete nucleotide sequence of the rabbit β-like globin gene cluster: analysis of intergenic sequences and comparison with the human β-like globin gene cluster. *J. Mol. Biol.* 205:15–40
122. Margulies EH, Blanchette M, NISC Comparative Sequencing Program, Haussler D, Green ED. 2003. Identification and characterization of multi-species conserved sequences. *Genome. Res.* 13:2507–18
123. Matassi G, Sharp PM, Gautier C. 1999. Chromosomal location effects on gene sequence evolution in mammals. *Curr. Biol.* 9:786–91
124. Matsuo K, Clay O, Takahashi T, Silke J, Schaffner W. 1993. Evidence for erosion of mouse CpG islands during mammalian evolution. *Somat. Cell Mol. Genet.* 19:543–55
125. Mayor C, Brudno M, Schwartz JR, Poliakov A, Rubin EM, et al. 2000. VISTA: visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics* 16:1046–47
126. McClelland M, Florea L, Sanderson K, Clifton SW, Parkhill J, et al. 2000. Comparison of the Escherichia coli K-12 genome with sampled genomes of a Klebsiella pneumoniae and three salmonella enterica serovars, Typhimurium, Typhi and Paratyphi. *Nucl. Acids Res.* 28:4974–86
127. Messier W, Stewart CB. 1997. Episodic adaptive evolution of primate lysozymes. *Nature* 385:151–54
128. Meyer IM, Durbin R. 2002. Comparative ab initio prediction of gene structures using pair HMMs. *Bioinformatics* 18:1309–18
129. Miller W. 2000. So many genomes, so little time. *Nat. Biotechnol.* 18:148–49
130. Miller W. 2001. Comparison of genomic DNA sequences: solved and

unsolved problems. *Bioinformatics* 17: 391–97

131. Moss EG, Tang L. 2003. Conservation of the heterochronic regulator Lin-28, its developmental expression and microRNA complementary sites. *Dev. Biol.* 258:432–42
132. Mouse Genome Sequencing Consortium. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* 420:520–62
133. Nachman MW. 2001. Single nucleotide polymorphisms and recombination rate in humans. *Trends Genet.* 17:481–85
134. Neafsey DE, Palumbi SR. 2003. Genome size evolution in pufferfish: a comparative analysis of diodontid and tetraodontid pufferfish genomes. *Genome. Res.* 13:821–30
135. Nei M. 1987. *Molecular Evolutionary Genetics*. New York: Columbia Univ. Press
136. Nekrutenko A, Chung WY, Li WH. 2003. ETOPE: Evolutionary test of predicted exons. *Nucl. Acids Res.* 31:3564–67
137. Nekrutenko A, Chung WY, Li WH. 2003. An evolutionary approach reveals a high protein-coding capacity of the human genome. *Trends Genet.* 19:306–10
138. Nekrutenko A, Makova KD, Li WH. 2002. The K(A)/K(S) ratio test for assessing the protein-coding potential of genomic regions: an empirical and simulation study. *Genome. Res.* 12:198–202
139. Nelson P, Kiriakidou M, Sharma A, Maniataki E, Mourelatos Z. 2003. The microRNA world: small is mighty. *Trends Biochem. Sci.* 28:534–40
140. Ning Z, Cox AJ, Mullikin JC. 2001. SSAHA: a fast search method for large DNA databases. *Genome. Res.* 11:1725–29
141. Oeltjen JC, Malley TM, Muzny DM, Miller W, Gibbs RA, Belmont JW. 1997. Large-scale comparative sequence analysis of the human and murine Bruton's tyrosine kinase loci reveals conserved regulatory domains. *Genome. Res.* 7:315–29
142. Ovcharenko I, Loots GG, Hardison RC, Miller W, Stubbs L. 2004. *zPicture*: dynamic alignment and visualization tool for analyzing conservation profiles. *Genome. Res.* 14:472–77
143. Pachter L, Alexandersson M, Cawley S. 2002. Applications of generalized pair hidden Markov models to alignment and gene finding problems. *J. Comput. Biol.* 9:389–99
144. Parra G, Agarwal P, Abril JF, Wiehe T, Fickett JW, Guigo R. 2003. Comparative gene prediction in human and mouse. *Genome. Res.* 13:108–17
145. Petrov DA, Sangster TA, Johnston JS, Hartl DL, Shaw KL. 2000. Evidence for DNA loss as a determinant of genome size. *Science* 287:1060–62
146. Pribnow D. 1975. Nucleotide sequence of an RNA polymerase binding site at an early T7 promoter. *Proc. Natl. Acad. Sci. USA* 72:784–88
147. Pruitt KD, Maglott DR. 2001. RefSeq and LocusLink: NCBI gene-centered resources. *Nucl. Acids Res.* 29:137–40
148. Rat Genome Sequencing Consortium. 2004. Evolution of the Mammalian Genome: sequence of the Genome of the Brown Norway Rat. *Nature.* 428:493–521
149. Reiser PJ, Kline WO. 1998. Electrophoretic separation and quantitation of cardiac myosin heavy chain isoforms in eight mammalian species. *Am. J. Physiol.* 274:H1048–53
150. Rogic S, Mackworth AK, Ouellette FB. 2001. Evaluation of gene-finding programs on mammalian sequences. *Genome. Res.* 11:817–32
151. Roskin K, Diekhans M, Haussler D. 2004. Score functions for assessing conservation in locally aligning regions of DNA from two species. *J. Comput. Biol.* In press
152. Rowitch DH, Echelard Y, Danielian PS, Gellner K, Brenner S, McMahon AP.

1998. Identification of an evolutionarily conserved 110 base-pair cis-acting regulatory sequence that governs Wnt-1 expression in the murine neural plate. *Development* 125:2735–46

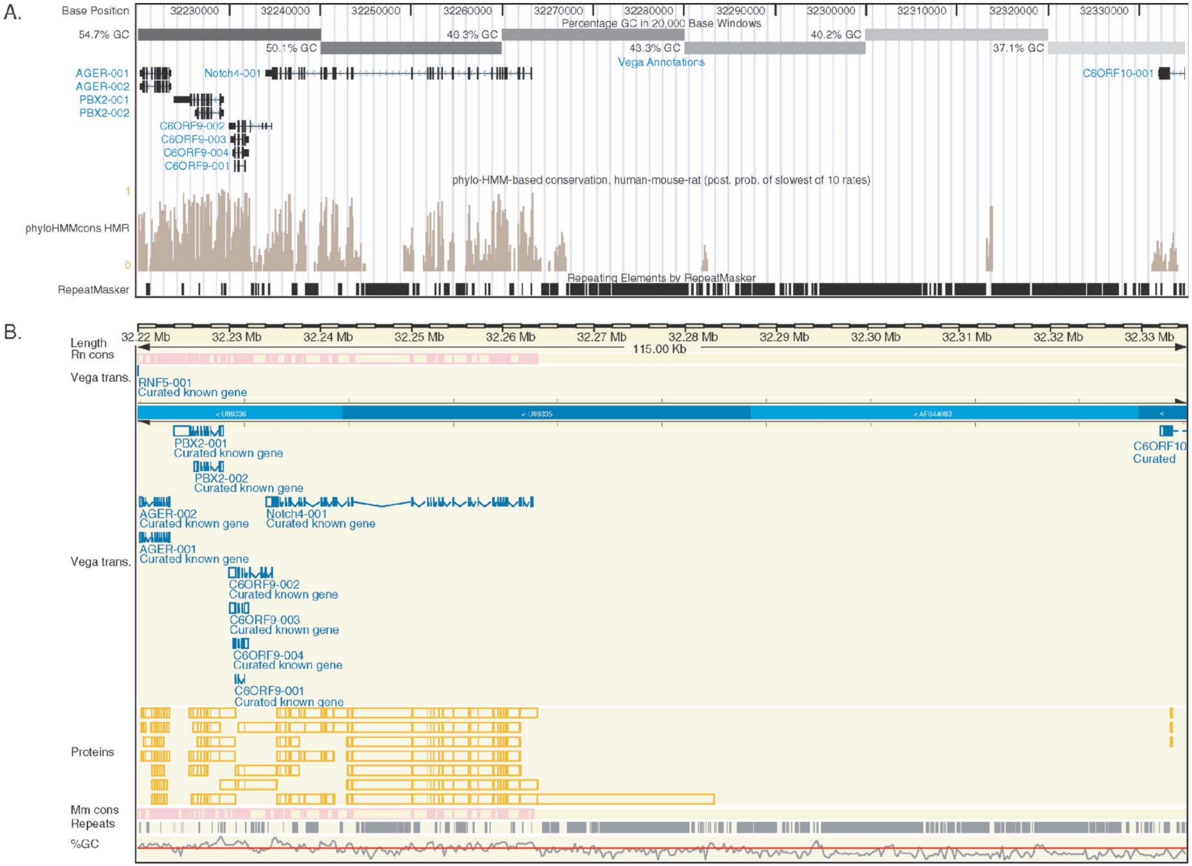
153. Rubin GM, Yandell MD, Wortman JR, Gabor Miklos GL, Nelson CR, et al. 2000. Comparative genomics of the eukaryotes. *Science* 287:2204–15
154. Saitoh N, Bell AC, Recillas-Targa F, West AG, Simpson M, et al. 2000. Structural and functional conservation at the boundaries of the chicken beta-globin domain. *EMBO J.* 19:2315–22
155. Salamov AA, Solovyev VV. 2000. Ab initio gene finding in Drosophila genomic DNA. *Genome. Res.* 10:516–22
156. Sarich VM, Wilson AC. 1967. Immunological time scale for hominid evolution. *Science* 158:1200–3
157. Schug J, Overton GC. 1997. Modeling transcription factor binding sites with Gibbs sampling and minimum description length encoding. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 5:268–71
158. Schuler G, Epstein J, Ohkawa H, Kans J. 1996. Entrez: molecular biology database and retrieval system. *Methods Enzymol.* 266:141–62
159. Schwartz S, Elnitski L, Li M, Weirauch M, Riemer C, et al. 2003. MultiPipMaker and supporting tools: alignments and analysis of multiple genomic DNA sequences. *Nucl. Acids Res.* 31:3518–24
160. Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, et al. 2003. Human-mouse alignments with *Blastz*. *Genome. Res.* 13:103–5
161. Schwartz S, Zhang Z, Frazer KA, Smit A, Riemer C, et al. 2000. PipMaker—a web server for aligning two genomic DNA sequences. *Genome Res.* 10:577–86
162. Shabalina SA, Ogurtsov AY, Kondrashov VA, Kondrashov AS. 2001. Selective constraint in intergenic regions of human and mouse genomes. *Trends Genet.* 17:373–76
163. Shehee R, Loeb DD, Adey NB, Burton FH, Casavant NC, et al. 1989. Nucleotide sequence of the BALB/c mouse β-globin complex. *J. Mol. Biol.* 205:41–62
164. Siepel A, Hausser D. 2004. Phylogenetic hidden Markov models. In *Statistical Methods in Molecular Evolution*, ed. R Nielsen. In press
165. Siepel A, Haussler D. 2004. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Mol. Biol. Evol.* 21:468–88
166. Smith NG, Webster MT, Ellegren H. 2002. Deterministic mutation rate variation in the human genome. *Genome. Res.* 12:1350–56
167. Stein LD, Bao Z, Blasiar D, Blumenthal T, Brent MR, et al. 2003. The genome sequence of *Caenorhabditis briggsae*: a platform for comparative genomics. *Public Libr. Sci. Biol.* 1:166–92
168. Stojanovic N, Florea L, Riemer C, Gumucio D, Slightom J, et al. 1999. Comparison of five methods for finding conserved sequences in multiple alignments of gene regulatory regions. *Nucl. Acids Res.* 27:3899–910
169. Stoye J, Evers D, Meyer F. 1997. Generating benchmarks for multiple sequence alignments and phylogenetic reconstructions. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 5:303–6
170. Stuart JM, Segal E, Koller D, Kim SK. 2003. A gene-coexpression network for global discovery of conserved genetic modules. *Science* 302:249–55
171. Su A, Cooke M, Ching K, Hakak Y, Walker J, et al. 2002. Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl. Acad. Sci. USA* 99:4465–70
172. Sundstrom H, Webster MT, Ellegren H. 2003. Is the rate of insertion and deletion mutation male biased?: Molecular evolutionary analysis of avian and primate sex chromosome sequences. *Genetics* 164:259–68
173. Tagle DA, Koop BF, Goodman M,

Slightom J, Hess DL, Jones RT. 1988. Embryonic ε and γ globin genes of a prosimian primate (*Galago crassicaudatus*): Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *J. Mol. Biol.* 203:7469–80

174. Templeton AR, Clark AG, Weiss KM, Nickerson DA, Boerwinkle E, Sing CF. 2000. Recombinational and mutational hotspots within the human lipoprotein lipase gene. *Am. J. Hum. Genet.* 66:69–83
175. Thomas JW, Touchman JW, Blakesley RW, Bouffard GG, Beckstrom-Sternberg SM, et al. 2003. Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* 424:788–93
176. Thompson W, Rouchka EC, Lawrence CE. 2003. Gibbs Recursive Sampler: finding transcription factor binding sites. *Nucl. Acids Res* 31:3580–85
177. Ting CT, Tsaur SC, Wu ML, Wu CI. 1998. A rapidly evolving homeobox at the site of a hybrid sterility gene. *Science* 282:1501–4
178. Uberbacher EC, Mural RJ. 1991. Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. *Proc. Natl. Acad. Sci. USA* 88:11261–65
179. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. 2001. The sequence of the human genome. *Science* 291:1304–51

179a. Wasserman WW, Sandelin A. 2004. Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.* 5:276–87

180. Wheeler DL, Church DM, Federhen S, Lash AE, Madden TL, et al. 2003. Database resources of the National Center for Biotechnology. *Nucl. Acids Res.* 31:28–33
181. Wheeler DL, Church DM, Lash AE, Leipe DD, Madden TL, et al. 2002. Database resources of the National Center for Biotechnology Information: 2002 update. *Nucl. Acids Res.* 30:13–16
182. Wiehe T, Gebauer-Jung S, Mitchell-Olds T, Guigo R. 2001. SGP-1: prediction and validation of homologous genes based on sequence alignments. *Genome. Res.* 11:1574–83
183. Wilson AC, Carlson SS, White TJ. 1977. Biochemical evolution. *Annu. Rev. Biochem.* 46:573–639
184. Wilson MD, Riemer C, Martindale DW, Schnupf P, Boright AP, et al. 2001. Comparative analysis of the gene-dense ACHE/TFR2 region on human chromosome 7q22 with the orthologous region on mouse chromosome 5. *Nucl. Acids Res.* 29:1352–65
185. Wingender E, Chen X, Fricke E, Geffers R, Hehl R, et al. 2001. The TRANSFAC system on gene expression regulation. *Nucl. Acids Res.* 29:281–83
186. Wolfe KH, Sharp PM, Li WH. 1989. Mutation rates differ among regions of the mammalian genome. *Nature* 337:283–85
187. Yang S, Smit AF, Schwartz S, Chiaromonte F, Roskin KM, et al. 2004. Patterns of insertions and their covariation with substitutions in the rat, mouse and human genomes. *Genome. Res.* 14:517–27
188. Yi S, Ellsworth DL, Li WH. 2002. Slow molecular clocks in Old World monkeys, apes, and humans. *Mol. Biol. Evol.* 19:2191–98
189. Young JM, Trask BJ. 2002. The sense of smell: genomics of vertebrate odorant receptors. *Hum. Mol. Genet.* 11:1153–60
190. Zhang MQ. 2002. Computational prediction of eukaryotic protein-coding genes. *Nat. Rev. Genet.* 3:698–709
191. Zhang P, Gu Z, Li WH. 2003. Different evolutionary patterns between young duplicate genes in the human genome. *Genome. Biol.* 4:R56
192. Zuckerkandl E, Pauling L. 1965. Molecules as documents of evolutionary history. *J. Theor. Biol.* 8:357–66



See legend on next page

Figure 3 Images from the UCSC Browser and Ensembl. A region of 115 kb containing the junction between the Class III and Class II regions of the human major histocompatibility complex is pictured. The Class III region shows much higher gene density, GC content (the fraction of bases that are guanine or cytosine), and interspecies conservation than the Class II region, and a lower density of interspersed repeats. Both the UCSC Browser (*panel A*) and Ensembl (*panel B*) illustrate these features, using different types of icons and displays for identical (e.g., Vega gene annotation) or highly similar (repetitive elements) data. Some features are distinctive to each browser, such as the phylHMMcons (164) track at UCSC, which gives an estimate of the likelihood that an aligned sequence is a more slowly changing region, and the proteins track at Ensembl.

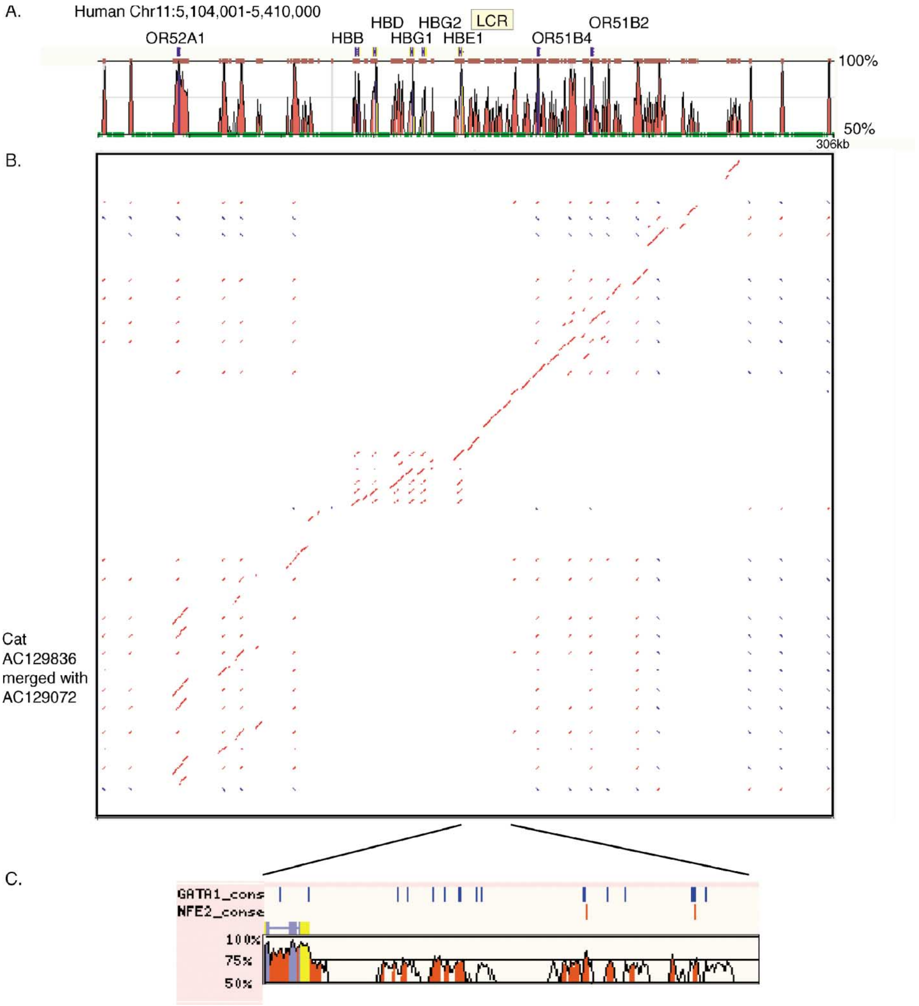


Figure 4 Use of the zPicture server to compare sequences from human and cat. A region of about 300 kb containing the beta-like globin gene cluster embedded in a cluster of olfactory receptor (OR) genes is pictured. Panel A shows human-cat conservation and gene annotations that were automatically extracted from the UCSC Browser. Panel B presents a dot-plot representation of the alignments (at the same scale as panel A), revealing a triplication of multiple OR genes in cat. Not all of the OR genes are annotated in the human; the matrices of short matches in the upper and lower left and right parts of the plot are matches between OR genes. Red lines are matches in the same orientation; blue is a match with the reverse complement. The matrix of matches in the center of the plot shows results from alignments among globin genes. The locus control region (LCR) is strongly conserved. Panel C shows conservation and predicted GATA1 and NFE2 binding sites that are conserved between human and cat in the LCR. These correspond to experimentally verified binding sites in localized segments of the LCR. Panel C was generated by exporting the alignments to rVista, a function supported by zPicture.

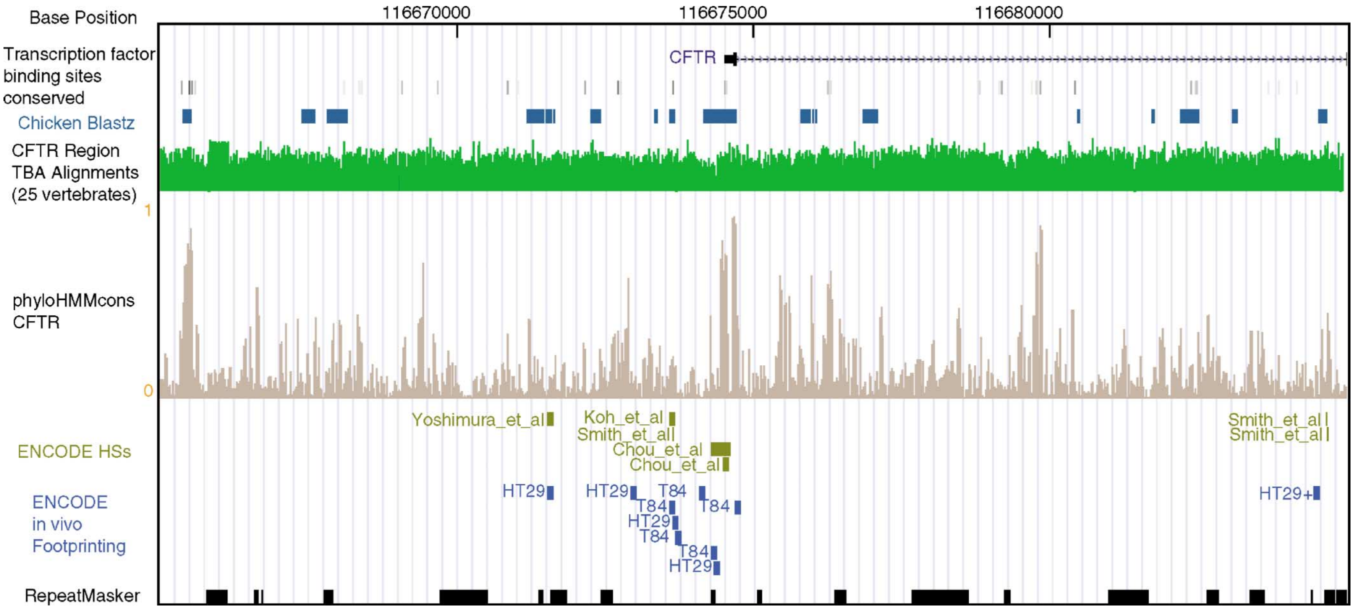


Figure 5 Prototype version of the UCSC Browser for showing functional data in register with sequence conservation. Experimentally confirmed DNAse I hypersensitive sites and protein-binding segments in the 20 kb surrounding the first exon of the CFTR gene are pictured, along with (*a*) computationally predicted transcription-factor binding sites conserved between human, mouse, and rat, (*b*) segments that align in a whole-genome comparison of human to chicken, (*c*) regions conserved in a multiple alignment of sequences from 25 vertebrates (using the phyloHMMcons program), and (*d*) positions of interspersed repeats.