Annu. Rev. Genomics Hum. Genet. 2005. 6:143–64
doi: 10.1146/annurev.genom.6.080604.162146

TRADE-OFFS IN DETECTING EVOLUTIONARILY CONSTRAINED SEQUENCE BY COMPARATIVE GENOMICS

Eric A. Stone,[1,2] Gregory M. Cooper,[3] and Arend Sidow[2,3]
Departments of Statistics,[1] Pathology,[2] and Genetics,[3] Stanford University, Stanford, California 94305; email: arend@stanford.edu

Key Words phylogenetic scope, functional elements, sequence conservation, genome evolution

■ **Abstract** As whole-genome sequencing efforts extend beyond more traditional model organisms to include a deep diversity of species, comparative genomic analyses will be further empowered to reveal insights into the human genome and its evolution. The discovery and annotation of functional genomic elements is a necessary step toward a detailed understanding of our biology, and sequence comparisons have proven to be an integral tool for that task. This review is structured to broadly reflect the statistical challenges in discriminating these functional elements from the bulk of the genome that has evolved neutrally. Specifically, we review the comparative genomics literature in terms of specificity, sensitivity, and phylogenetic scope, as well as the trade-offs that relate these factors in standard analyses. We consider the impact of an expanding diversity of orthologous sequences on our ability to resolve functional elements. This impact is assessed through both recent comparative analyses of deep alignments and mathematical modeling.

INTRODUCTION

Toward the goal of understanding human biology and disease, the use of experimentally amenable model organisms has been and remains the dominant paradigm in basic biomedical research. A chosen few of these organisms, for reasons practical and historical, have received intense concentration, and emerged as the standard bearers of their phyla. This select group, collectively the "Security Council of Model Genetic Organisms" (37), includes lambda phage (88), *Bacillus subtilis* (59), *Escherichia coli* (11), *Saccharomyces cerevisiae* (42), *Chlamydomonas reinhardtii* (92), *Arabidopsis thaliana* (5), *Caenorhabditis elegans* (18), *Drosophila melanogaster* (2), and *Mus musculus* (75), in addition to humans (50).

The model organisms of the Security Council were chosen for their suitability to experimental genetics. Recently, comparative sequence analysis of model organism genes and genomes has emerged as a powerful approach complementary to experimentation, facilitating the discovery of genomic elements that have common functions within the human and other lineages. Because mutations within functional regions usually confer a selective disadvantage, they are less likely to result in evolutionary change (56). This pressure to maintain function, defined as evolutionary constraint, restricts the space of sequences that evolution can explore, reducing evolutionary rates to less than neutral expectation and ultimately manifesting as sequence "conservation."

The relationship between biological function and evolutionary constraint provides a powerful mechanism for the discovery of functional elements based on orthologous sequence comparisons. The annotation of protein-coding genes in the human genome, for example, can be significantly improved through comparisons with the mouse genome sequence (83). The promise of this approach was supported by early observations that orthologous mouse and human coding exons typically exhibited 75% to 95% identity at the nucleotide level, contrasted with 60% or lower similarity levels throughout most aligned regions (75). One of the best examples of the effectiveness of human-mouse comparisons is the discovery of the *ApoAV* gene, which plays an important role in controlling triglyceride levels (84).

However, various studies over the past several years have made it readily apparent that most of the constrained fraction of the human genome does not code for proteins: While ∼5% to 6% of the genome is constrained (24, 75, 86), protein coding exons only occupy ∼1.2%, with untranslated regions of their associated transcripts occupying an additional ∼0.7% (51). The relative abundance of noncoding constrained elements can be found at all levels of stringency, ranging from intensely to weakly constrained (8, 24, 28). The abundance of these conserved noncoding elements, coupled with our relative lack of understanding of their functions, has elevated their annotation and characterization to a high priority (21, 34).

For those elements that function as noncoding RNAs, the existence of recognizable patterns in conjunction with comparative analyses has allowed substantial progress in both their identification and functional characterization throughout eukaryotic genomes (9, 29, 62, 65, 72, 79, 102). The vast majority of conserved noncoding sequences, however, fail to conform to any known regular pattern, and interspecies conservation remains the primary tool for their discovery and guides subsequent experimentation. In this review, we focus on the efficacy of this process in an era of diverse sequence data, specifically with regard to noncoding elements that do not appear to be transcribed. These include, but are not necessarily restricted to, elements that regulate chromatin structure and gene expression, such as promoters, enhancers, silencers, and insulator elements.

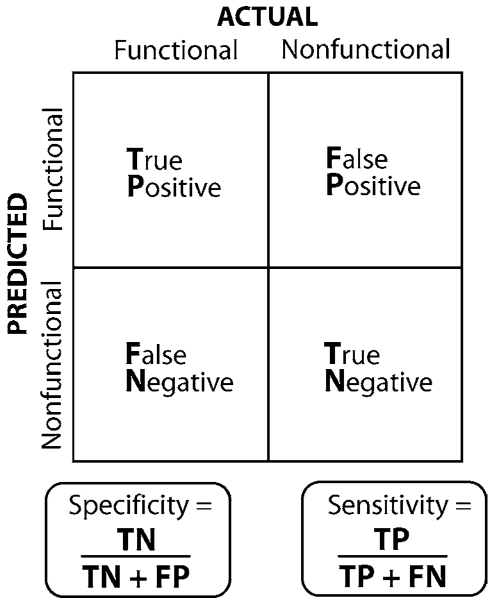
With the recent completion of genome sequences from a number of important organisms, the comparative genomics literature has grown considerably. In response to the rapid publication of methods and their applications, a variety of informative reviews have been written recently. In particular, there was a comprehensive survey of the field in this series last year (73), and Ureta-Vidal et al. (98) provided a detailed examination of important methods. Enard & Paabo (33) gave another review of comparative genomics that emphasized the impact of the field to our understanding of primate evolution. An interesting discussion of the role of phylogenetic scope, defined as the taxonomic range of species included in sequence comparisons, was provided by Boffelli et al. (13), contrasting the contributions of species placed on the proximal and distal edges of the vertebrate phylogeny. In addition, a discussion of the classes of human functional elements that can be found using various phylogenetic scopes can be found in Reference 23. For a short primer on the field, see Reference 47.

Though the extant review literature encompasses many aspects of comparative genomics, there does not appear to be any focused treatment of the quantitative impact of an expanding sequence repertoire. Supported by the results of recent studies that analyzed deep genomic sequence alignments, we embark on this task here. Comparative sequence analyses are necessarily based on multiple sequence alignments and well characterized species phylogenies. A number of methods exist for both computing genomic multiple sequence alignments (10, 15–17, 25, 90) and constructing trees (35, 36, 39, 53, 80, 100, 101), but these issues are beyond the scope of our review. We review the field of comparative genomics through the knowledge gained from sequence diversity and depth. We emphasize the impact of deep sequence on the universal features of comparative methodology: specificity, sensitivity (Figure 1), and phylogenetic scope. By reflecting on what we have learned, we speculate on the nature of future gains and provide a straightforward quantitative model that captures the relationships between sequence diversity, specificity, and sensitivity in the context of detecting the impact of purifying selection on genomic sequences.

SEQUENCE DIVERSITY PERMITS GREATER SPECIFICITY

The ability of genomic sequence comparisons to resolve cases of purifying selection is intimately tied to the neutral rate of evolution. Orthologous instances of functional elements that have evolved slowly due to selective constraints can be detected by their sequence similarity; however, ancestral sequences that have evolved slowly by chance are indistinguishable from their functional counterparts. The relative rate at which neutral sequences evolve, as compared to the rate of evolution for sequences under selective constraints, is the determining factor in the efficacy of comparative methods in distinguishing functional elements. Specificity, which quantifies the extent to which neutral sequence is misidentified as functional, is thus largely a function of evolutionary distance, expressed as nucleotide substitutions affecting neutrally evolving sites.

Many of the early successes of vertebrate comparative genomics were necessarily sequence comparisons between human and mouse (57, 68, 78). Human-mouse studies profit from an evolutionary distance between the species that is sufficient but not excessive (95); a large fraction of their genome sequences can be aligned, and functional elements can be identified by their conservation against the neutral background. Efficient identification of these functional elements, however, requires the adoption of a significance threshold strict enough to preclude many of the slower-evolving neutral regions that can be aligned (75, 91). Successful



PREDICTED \ ACTUAL	Functional	Nonfunctional
Functional	**T**rue **P**ositive	**F**alse **P**ositive
Nonfunctional	**F**alse **N**egative	**T**rue **N**egative

$$\text{Specificity} = \frac{\mathbf{TN}}{\mathbf{TN + FP}} \qquad \text{Sensitivity} = \frac{\mathbf{TP}}{\mathbf{TP + FN}}$$

Figure 1 Specificity, sensitivity, and prediction. Coincidence of prediction and truth define the four outcomes that may result from the statistical testing of genomic sequence for functionality. True positives (TP) represent correctly determined functional elements, whereas false negatives (FN) correspond to functional elements that are missed. Sensitivity measures the fraction of functional elements that a procedure is able to detect. False positives (FP) designate nonfunctional sequences mistakenly predicted to be functional, whereas true negatives (TN) are nonfunctional sequences that have been correctly determined as such. Specificity measures the fraction of total nonfunctional sequences that can be properly identified.

strategies based on a requirement of at least 70% identity [or 80%, e.g., (97)] over 100 consecutive bases have been used in numerous studies, including the identification of a coordinate regulator of Interleukins 4, 13, and 5 (63) and a comprehensive analysis of human chromosome 21 (27).

Generally, such threshold-based approaches fall into one of two categories, depending on the style of subsequent analysis. The first category consists of studies that nominate the most highly conserved elements for experimental assay, such as Reference 63, and a study that discovered a novel sequence regulating *ABCA1* (85). The second category consists of approaches that seek to broadly characterize conserved elements that meet threshold requirements. An example for this type of study includes the recent determination that a significant fraction of elements so chosen overlap predicted matrix-scaffold attachment regions (41). Approaches of the first category are highly specific, but by design lack any meaningful sensitivity; by contrast, the second category is largely composed of sensitive studies that are prone to admitting false positives (Figure 2*a*).

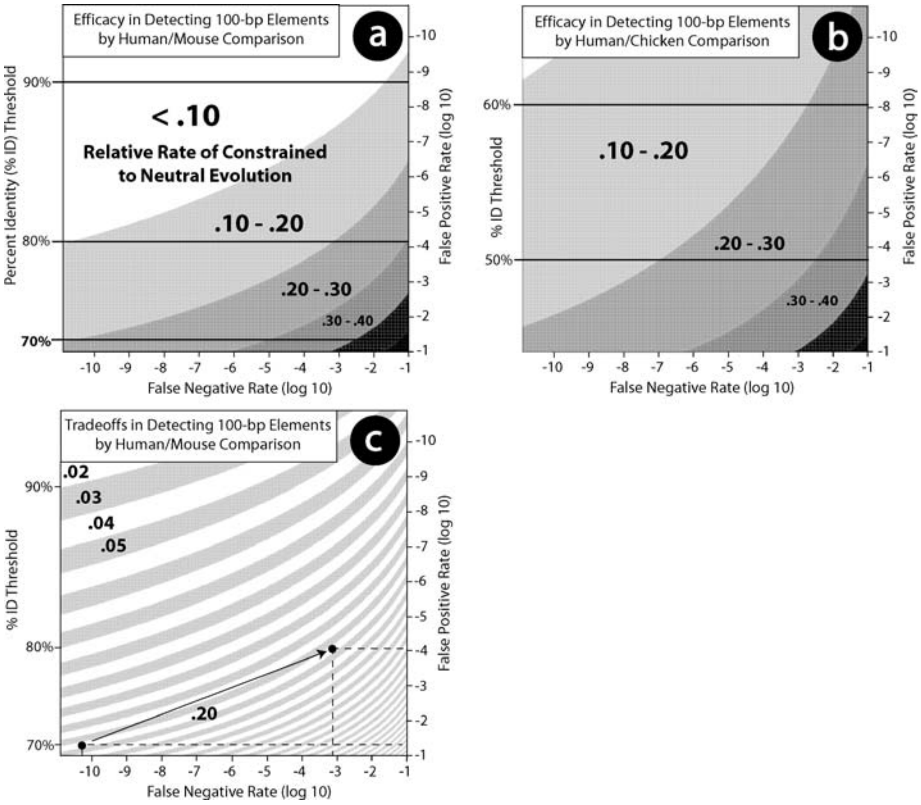


Figure 2 Efficacy of pairwise analyses. (*a*) The ability of 100-bp human-mouse sequence comparisons [at a neutral distance of 0.55 subs/site (24)] to resolve classes of functional elements is described in terms of false positive rates (*right vertical axis, log scale*) and false negative rates (*horizontal axis, log scale*). False positive rates are governed by the percent identity thresholds shown on the left vertical axis. False negative rates depend both on that threshold and the relative rate of neutral evolution at which conserved elements evolve; classes of relative rates are indicated by shaded regions of the plot, increasing in tint from light to dark. Calculations here and subsequent assume a Jukes-Cantor (J-C) one-parameter model of evolution (54). (*b*) The ability of 100-bp human-chicken sequence comparisons [at a neutral distance of 1.66 subs/site (49)] to resolve classes of functional elements is shown as in (*a*). (*c*) The plot of (*a*), with relative rates of functional elements now shown in steps of 0.01 as shaded contours. Both black circles correspond to the detection of elements evolving at 20% of the neutral rate. The lower left circle shows the performance of detection at a threshold of 70% identity; the upper right circle shows the same at a threshold of 80%.

By considering vertebrate genomes more distant than mouse for comparison to human, specificity can be remarkably improved (Figure 2*b*). For example, human-fugu comparisons were used to identify 25 highly conserved noncoding sequences flanking 4 developmental regulators (*SOX21*, *PAX6*, *HLXB9*, *SHH*), and 23 of these exhibited enhancer activity in one or more tissues of zebrafish embryos (99). An alternative approach to gaining specificity is to tighten the human-mouse threshold parameters at the sacrifice of some sensitivity (Figure 2*c*). In fact, a recent study examined the distributions of human-mouse and human-fugu conserved elements to derive threshold criteria for human-mouse that mimicked the efficiency of human-fugu comparisons (82). That such analogous criteria exist illustrates the duality between sensitivity and specificity, and underscores the analytical limitations inherent to pairwise comparisons. Better discrimination of functional elements is obtained by orthologous sequence from additional species.

In the simplest case of extension beyond pairwise comparisons, genomic sequence from a third species can be used to filter the identified conserved elements. By increasing the total evolutionary divergence considered, this strategy can improve specificity. If the third species is highly diverged, the initial list of candidate elements can be drastically reduced, although scope limitations will significantly reduce sensitivity. Chicken sequence, for instance, was used to isolate a unique element in the 10-kbp region upstream of the homeobox gene *Nkx2–5* from those initially obtained by human-mouse comparison (61). The availability of a complete draft of the chicken genome sequence should make this a prevalent technique to dramatically improve specificity of human-mouse comparisons, especially in light of the observation that only a small fraction of the human genome ($\sim$2.5%) can be aligned with the chicken genome (49).

The use of multiple species comparisons to dramatically improve specificity in the detection of functional elements is expanding rapidly, especially with the recent completion of several important genomes (4, 49, 52, 75, 86). A variety of examples highlight this literature, including the discovery of elements that regulate the expression of the stem cell leukemia (*SCL*) gene (45) and the identification of elements that enhance expression of the *Dlx* genes in the vertebrate forebrain (40). Another example is the comparative approach used to locate long-range enhancers of *DACH* (77). Using human-mouse comparisons, 1098 conserved noncoding sequences ($>$100 bp and with $>$70% identity) were discovered in a 2630-kbp targeted region. To limit the search, these 1098 sequences were filtered according to their presence in frog, zebrafish, and two pufferfish. By excluding those sequences not present in all species, the number of conserved sequences was reduced to 32; of 9 tested, 7 were enhancers. Examples of regulatory elements identified using comparisons confined to multiple mammalian sequences are also beginning to accumulate, including discoveries made using human, mouse, and rat (67), and also by inclusion of marsupial sequence, whose evolutionary distance from placental mammals can provide significant specificity gains (19).

Importantly, the specificity that multiple comparisons can achieve makes more refined investigations possible; the 100-bp windows common to human-mouse comparisons can be shortened considerably in deeper analyses. This was demonstrated in a multispecies analysis of the murine early enhancer of *Hoxc8* (94), which showed that eight diverse mammals were sufficient to resolve transcription factor binding sites (TFBSs) at significance using only small windows of sequence conservation. This point was further illustrated in comparisons of the human, mouse, and rat genome sequences, demonstrating reasonable specificity at a resolution of 50 bp (24), in agreement with a recent theoretical analysis (30).

With sufficient sequence diversity, it should be possible to dispense with length criteria entirely. Despite uncertainty in the number of sequenced genomes that may be required, resolving constraint at the nucleotide level is a goal that will ultimately be achieved (22, 30). We can already be confident that invariant nucleotides in deep multiple sequence comparisons are under some degree of purifying selection. Increased specificity through sequencing will allow us to identify less radical signatures of functional importance at the finest level of resolution possible.

SEQUENCE DIVERSITY LEADS TO INCREASED SENSITIVITY

The vast majority of functional elements, although under purifying selection, will stochastically accumulate substitutions at a nontrivial fraction of the neutral rate. Thus, although conservation is a reliable guide to locating important elements (46), the ability of functional sequence to tolerate such change complicates its discrimination from neutral DNA. By chance, some elements will have changed so much that they fall short of the conservation thresholds required for detection. Reflecting this, sensitivity quantifies the extent to which functional sequence is misidentified as neutral because of accumulated evolutionary change. As demonstrated by the examples of the previous section, sensitivity and specificity are interrelated in their opposition. But whereas above we were primarily concerned with neutral sequence that has evolved slowly by chance, the discussion here focuses on identifying functional sequence despite the accumulation of evolutionary change. The accurate resolution of both cases requires effective discrimination that relies on neutrally evolving sequence appearing less similar than its functional counterpart.

Achieving effective discrimination is a problem for comparisons both of distantly related and closely related species; however, the troubles in each case are distinct (Figure 3). Distant sequence comparisons are hampered by mutational saturation; given enough time, even slowly evolving functional sequence will accumulate enough changes to appear to have evolved unconstrained. Conversely, over short periods of time, even rapidly evolving neutral sequence will not accumulate enough changes to be distinguishable from an instance of purifying selection. Within the vertebrate scope, distant comparisons are subject to saturation in moderately constrained elements, but still allow useful levels of sensitivity for the most strongly constrained elements. For instance, human-fish sequence comparisons have been used to successfully identify both coding genes (1, 4) and regulatory

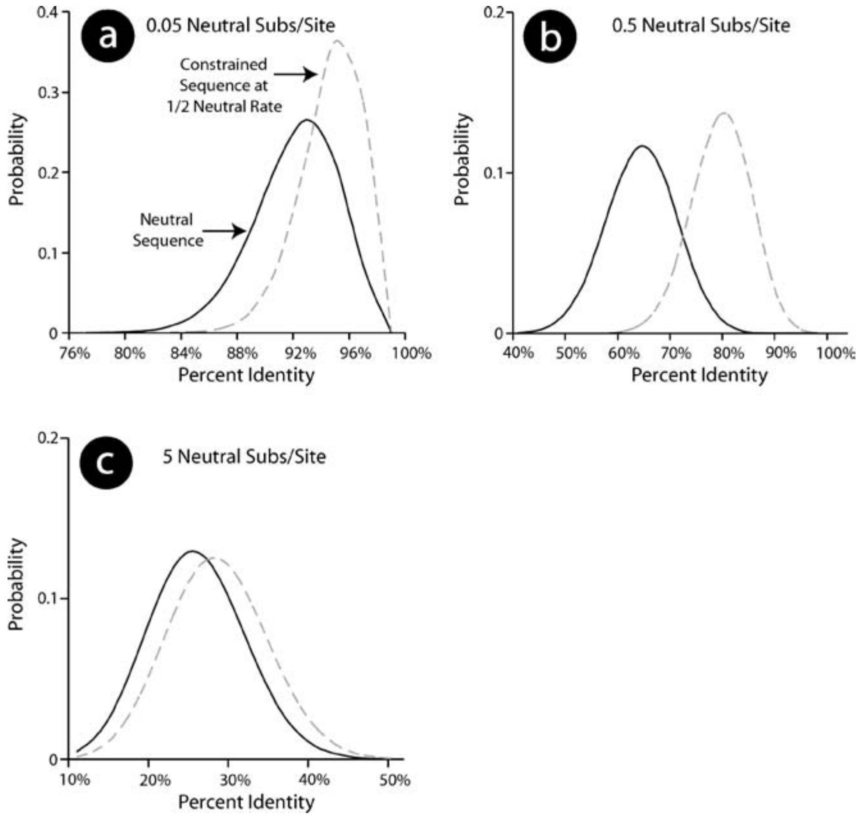


Figure 3 Discrimination at increasing neutral divergence. (*a*) The solid curve shows the probability distribution of pairwise identity between two 50-bp sequences diverged by 0.05 subs/site under the Jukes-Cantor (J-C) model. The dashed curve gives the distribution of conserved sequences that have diverged by 0.025 subs/site, equal to half the neutral distance. Both curves plot the probability mass function against percent sequence identity. (*b*) As in (*a*), assuming a neutral distance of 0.5 subs/site and a constrained distance of 0.25 subs/site. (*c*) As in (*a*), assuming a neutral distance of 5 subs/site and a constrained distance of 2.5 subs/site.

elements (7, 99). It must be borne in mind, however, that elements detectable by such comparisons are a small minority of all human functional elements.

Although more rapidly evolving functional elements may be rendered imperceptible by saturation, the inclusion of additional vertebrate genomes is straightforward. By contrast, the augmentation of comparisons of closely related species is hindered by a restrictive phylogenetic scope. Put another way, although one well-placed species can improve a comparison of distant sequences dramatically, the sequences of many closely related species are required to obtain the aggregate evolutionary divergence necessary to discriminate neutral and functional DNA. Based on this premise, phylogenetic shadowing (12, 81) was proposed as a method of orthologous sequence comparison to detect functional elements from extraordinary species diversity within a narrow evolutionary scope. As with other comparative approaches, shadowing relies on gathering sufficient neutral variation to contrast the signature of purifying selection. Thus, the method can be used to similar ends as distant comparisons, but it is uniquely suited to identifying lineage-specific functional elements. For example, by analyzing 18 primate sequences orthologous to a 1.6-kbp region of the *Apo*(*a*) locus, phylogenetic shadowing uncovered novel transcriptional regulation of a gene found only in Old World monkeys and hominids (12, 60).

The limitations on sensitivity to lineage-specific gains are paralleled by the obstacles presented by lineage-specific losses. In particular, the lineage-specific loss of functional elements (6, 58) undermines their discovery from a comparison of a limited number of species. The mouse and rat genomes, despite their frequent inclusion in comparative studies, may be suboptimal in this regard for discovering common mammalian functional elements. The relatively rapid rate of molecular evolution of rodent genomes, in terms of both small-scale (86, 96) and large-scale (14) changes, may have resulted in greater turnover of functional elements, especially in contrast to the more slowly evolving genomes characteristic of other mammalian lineages such as primate or elephant (14, 66, 76).

TFBSs are one class of functional elements for which there is direct evidence that evolutionary turnover may be a problem. Despite the fact that a number of studies have successfully leveraged sequence comparisons to detect TFBSs (55, 64, 71, 74, 94), experimental data from functional studies of 20 regulatory regions revealed that 32% to 40% of the human functional sites were not functional in rodents (26). It is probable in these cases that including rodent sequence in a comparative analysis will obscure the functional human element unless other genomes are also utilized. An additional concern is that although some lineages may be more prone to loss than others, the phenomenon and its effect on sensitivity are widespread. Illustrating this, a comparative analysis of horse, cow, pig, dog, cat, and mouse sequence orthologous to the *SIM2* gene interval on human chromosome 21 revealed that sequences conserved in various subsets of mammals were frequently functional (38).

The extent to which evolutionary variation obscures the detection of functional elements by comparative methods is difficult to estimate, owing to our limited understanding of noncoding functional sequence. Reliable analyses of sensitivity are hampered by the lack of any large-scale “gold standard,” an experimentally verified data set that comprehensively captures the functional elements in a given region for a given species (20, 32). In the absence of such validation, less representative experimental data can still be used to roughly estimate false negative rates. For example, human-mouse-rat comparisons recently demonstrated 100% sensitivity in detecting six experimentally identified functional binding sites using conservation criteria and a known binding motif (86). Regions that have been intensely studied, both by experimentation and comparative analyses, present the most immediate solution to improving the estimation of sensitivity; examples include the *Hox* gene clusters (3, 87, 89) and the *β-globin* (31, 68) and *SCL* loci (20, 43–45).

A useful, comprehensive, and currently more tractable alternative to experimental validation can be found in data sets derived by computation alone. This strategy is particularly suited to functional elements under strong purifying selection, as a reasonable validation set can be built from the conserved elements obtained from a deep multiple sequence comparison. For instance, multispecies conserved sequences (MCSs) were suggested as a set of candidate elements of functional importance (70). MCSs derived from 12 vertebrate sequences orthologous to a 1.9-Mbp stretch of the human *CFTR* region (34, 96) were used to estimate the ability of species subsets to resolve constrained elements (70, 96). For placing pairwise studies in perspective, the performance of human-mouse comparisons on the MCS validation set is noteworthy: At an 85% identity threshold for human-mouse alignment, only 41% of MCS bases were covered.

The sensitivity achievable by comparative analyses is strongly dependent on how functional sequence evolves. The genomes already sequenced and in the pipeline should collectively permit a wholesale identification of the noncoding functional elements that are highly conserved throughout vertebrate evolution. The extent to which purifying selection has acted on these widely conserved sequences suggests that they play fundamental roles in organismal function and development (8, 87). Many ubiquitously conserved noncoding elements are involved in regulating developmental genes, and recent work provides further evidence that this may be the case (99). Despite the high level of interest that these sequences generate, they represent only the lowest-hanging fruit. Comprehensively cataloging the functional elements of the human genome will require the identification of weaker signatures of selection, a task that can only be achieved with substantial sequence diversity.

MODELING THE TRADE-OFFS IN COMPARATIVE ANALYSES

We discuss sensitivity and specificity as complementary measures describing the efficacy of a comparative analysis in locating slowly evolving genomic regions. This efficacy is founded on a number of parameters that vary with each application, including the thresholds of X% identity over Y base pairs that have been used in pairwise comparisons. Such studies, although simple in their approach, are sufficient to illustrate many of the trade-offs inherent to comparative analyses. Figure 4 depicts the false positive and false negative error rates in using human-mouse comparisons to discover elements of fixed length Y evolving fivefold slower than neutral. For pairs of threshold parameters X and Y, the coincidence of false positive and false negative error rates is qualitatively described. We used false positive and false negative error rate boundaries of 10^{-6} per element to draw the

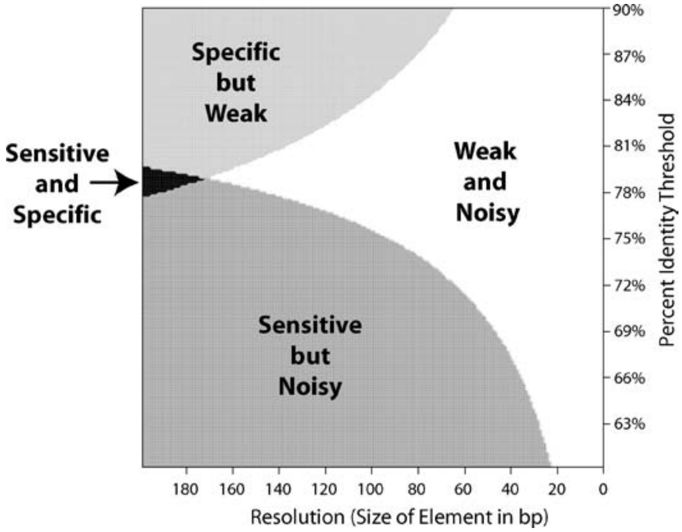


Figure 4 Balancing sensitivity and specificity in pairwise comparisons. The performance of human-mouse sequence comparisons is qualitatively described at varying detection thresholds. The horizontal axis shows the length of sequence considered; the vertical axis indicates the percent identity required for a sequence to be predicted as functional. A false negative rate of less than 10^{-6} is described as "sensitive," with the alternative considered "weak." A false positive rate of less than 10^{-6} is "specific" and the alternative is considered "noisy."

figure. As shown, it is easy to bound exactly one of the error rates by choosing restrictive values of X and Y; however, striking a useful balance between the two requires carefully chosen thresholds.

The practical difficulty with balancing false positive and false negative predictions is that calculations of false negative rates depend on the relative rate at which conserved elements are presumed to evolve. This factor, which we call gamma (above set to 0.2 for fivefold reduction), is another parameter that enters even the most basic analysis. The genome-wide distribution of these gammas is of fundamental interest and reflects the variable strength of selection on functional DNA. Related to this distribution is the density of functional elements within the human genome, that quantity, which, along with gamma, translates false positive and false negative rates into specificity and sensitivity (see Figure 1). Although there is evidence that the fraction of the human genome under purifying selection is around 5%, this value and the portion of it attributable to noncoding functional elements are only rough estimates of the truth. By illuminating the spectrum of conservation that selection processes induce across the genome, large-scale studies such as the ENCyclopedia of DNA Elements (ENCODE) project (34) will provide the first reliable estimates of these critical quantities.

Beyond Pairwise Comparisons

Despite the utility of pairwise analyses in illustrating the trade-offs in comparative genomics, issues such as phylogenetic scope and sequence diversity do not translate to this setting. To expand the discussion, we turn to a model that is suitable for capturing the trade-offs inherent to multiple comparisons (Figure 5). Our approach is similar to a recently published model that was accompanied by a thorough mathematical analysis (30). Here we concentrate on a quantitative treatment of the relationships between some of the parameters previously described. Because such treatment requires a set of strong simplifying assumptions, the precise conclusions should be regarded as qualitative.

As with the pairwise approach, the model assumes a perfect alignment of ungapped orthologous sequences. The expectation that small insertions and deletions will accumulate over large evolutionary distances suggests that their absence in a particular region of a deep multiple sequence alignment might alone be sufficient evidence to infer purifying selection. This signature of functional importance is invisible to our model, but its lack does not significantly detract from its utility. In particular, the potential impact of indels on the above analyses diminishes at finer scales of resolution, and also in tight phylogenetic scopes, and this is where we believe that our observations have their greatest importance. When using fine resolution to identify small, highly conserved sequences within larger conserved segments, the impact of indels will be smaller still.

Modeling Assumptions

In place of an estimate of the neutral rate, we now suppose that a correct phylogeny with branch lengths in neutral substitutions per site (subs/site) is available, and that sequence diversity has yielded a phylogeny dense enough that the sequences at each internal node of the tree can be accurately reconstructed. Thus, we take these to be given, ignoring their probabilistic nature, and treat the nucleotide bases on either end of every branch as observed. We simplify the analysis by disregarding any variation in branch lengths across the tree. This is accomplished by distributing the total branch length (in neutral subs/site) across each of the *2N-2* branches of the rooted, bifurcating phylogeny, where *N* is the number of species in the study. (Given our modeling framework, other idealized tree configurations would be possible, for example, distributing the total branch length evenly among only the terminal branches to produce a star phylogeny.) We suppose that the orthologous nucleotides at each position of a length-L sequence evolve independently along each branch of the phylogeny under a Jukes-Cantor (one parameter; all changes are equally likely) model of nucleotide evolution (54). Under neutral evolution, the probability that the nucleotides spanning a branch agree is $p = 1/4 + (3/4)\exp(-4D/3)$, where D is the total branch length T divided by *2N-2*. Functional nucleotides evolve more slowly by a factor of gamma (γ); thus, D is replaced by γD in the above. To incorporate phylogenetic scope, we use a parameter κ to represent the fraction of the phylogeny's total branch length that lies outside the scope

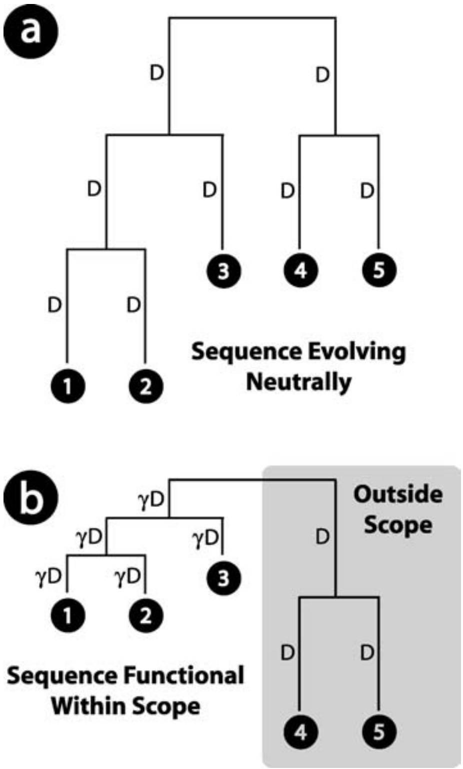


Figure 5 A model for multiple sequence comparisons. (*a*) Comparative sequences (here $N = 5$) related by a known phylogeny are shown. The total branch length ($8D$ as shown) is assumed to be equally distributed among the branches of the tree. It is presumed that the ancestral sequences are known, so each aligned nucleotide position contributes one pairwise comparison per branch. For neutrally evolving sequence, these pairwise comparisons are uniformly diverged by D subs/site. (*b*) Functional sequence evolves more slowly than neutral sequence by a factor of gamma (drawn to scale as $\gamma = 0.33$). Beyond the phylogenetic scope of the element (outside scope is shaded), sequence evolves neutrally. Each aligned nucleotide contributes five pairwise comparisons diverged by γD subs/site and three pairwise comparisons diverged by D subs/site. The fraction of branches outside the scope is denoted by kappa (here $\kappa = 3/8$).

of the functional element. Within the scope, purifying selection constrains the evolution of the element, but outside of the scope, the element evolves neutrally. Note that the same model applies to a loss of constraint in a set of lineages within the scope. Finally, we rely on asymptotic approximations in our calculations and figures. Taken together, this permits evaluation of the interplay between specificity, sensitivity, scope, resolution, and sequence diversity.

Sequence Diversity

It is well established that sequence diversity, as measured by the total branch length in neutral subs/site, correlates strongly with the performance of a comparative analysis (12, 13, 23, 30, 93). We considered this relationship in terms of specificity and sensitivity, respectively quantified by false positive and false negative rates (Figure 6*a*). To obtain the total branch length from the number of sequences included in the study, each species was assumed to contribute 0.2 subs/site of independent branch length. Small changes in this arbitrary value do not impact the observations; for example, changing to 0.1 subs/site per species and doubling the number of species generates similar results. Traversing from the lower right to the upper left of Figure 6*a* it is clear that increasing the total branch length uniformly reduces both error rates. The contours of the plot suggest that sensitivity can be obtained at far smaller branch lengths than is possible for specificity; however, this is influenced by our assumption, for this particular analysis, of functional sequence evolving five times slower than neutral (i.e., $\gamma = 0.2$).

Contaminating Neutral Sequence

To gauge the impact of including species beyond the scope of an element, we reconsidered this analysis supposing that the sequence evolved neutrally over 10% ($\kappa = 0.1$) of the total branch length. Contaminating the comparative analysis with 10% neutral sequence weakened the detection of functional elements considerably (Figure 6*b*), as the requirement of 7.7 subs/site for error rates of 10^{-6} in the original analysis becomes 10.5 subs/site after contamination. Even discounted for the 10% figure (i.e., 10.5 subs/site–1.05 subs/site), an additional 1.8 subs/site are necessary to counteract the effect of reaching slightly beyond scope. Although we have not considered the large-window detection of a small functional element embedded in neutral sequence, the same calculation applies and underscores the need for high sequence diversity to obtain sufficiently fine resolution.

Resolution Versus Branch Length

It was recently shown that branch length and resolution are roughly inversely proportional under certain modeling assumptions (30). The same phenomenon holds in our model (Figure 6*c*). We fixed the false positive rate at 10^{-4} to consider how branch length and element size interact to influence sensitivity. Reasonable false negative error rates can be achieved at 10-bp resolution with 6neutral subs/site; however, power falls off sharply as element size decreases (Figure 6*c*). Introducing 10% contamination is enough to compromise even 10-bp resolution (Figure 6*d*). The branch length necessary to restore the previous level of sensitivity is considerable: about 8.4 subs/site are required.

Realism

We are well aware that greater sophistication is necessary to obtain a truly realistic model of the quantitative trade-offs in comparative genomics. Yet there are

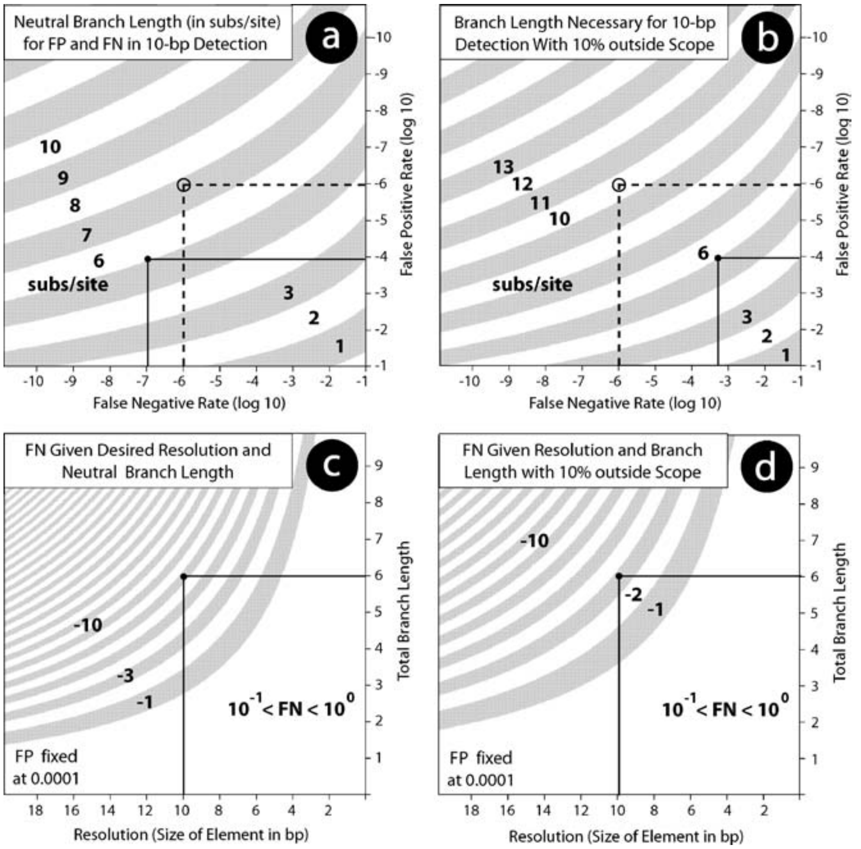


Figure 6 Trade-offs inherent to comparative analyses. (*a*) The ability to detect 10-bp functional elements is shown in terms of false positive (FP) rates (*vertical axis, log scale*) and false negative (FN) rates (*horizontal axis, log scale*). The total branch lengths required to achieve specified FP and FN levels are described by the plot area. Alternately shaded regions demarcate units of branch length, increasing from lower right to upper left. The solid horizontal line indicates a FP rate of 10^{-4}, which is used in subsequent panels; this intersects (at the *black dot*) the solid vertical line at the contour designating a total branch length of 6 subs/site. The dashed lines indicate matching FP and FN rates of 10^{-6} and meet at the hollow circle. (*b*) The plot of (*a*) is redrawn to account for comparisons in which 10% of the total branch length lies outside the phylogenetic scope of the element. (*c*) The ability to detect elements smaller than 20 bp is shown for increasing total branch lengths in terms of the FN rate, with the FP rate fixed at 10^{-4}. As the labels indicate, the plot area is alternately shaded in base-ten logarithmic units of FN. The horizontal line represents a total branch length of 6 subs/site, whereas the vertical line designates a resolution of 10 bp. The black dot is located at a position analogous to the same symbol used in (*a*). (*d*) The plot of (*c*) is modified for comparisons in which 10% of the total branch length lies outside the scope. Here the black dot is located analogously to the same symbol used in (*b*).

important issues that affect analysis even at this basic level. The discovery of functional elements by genomic comparison is fundamentally a scanning procedure not unlike LOD-score linkage analysis (48). Statistics that do not take this into account, including those that we describe, are instructive but may not accurately reflect practical significance levels. It should also be noted that comparative methods are typically large multiple testing procedures (69). For instance, scanning a 1-Mbp region for functional elements using nonoverlapping windows of 100 bp requires 1000 independent hypothesis tests. Despite this, simple approaches such as the one employed here are sufficient to reveal the interactions between important parameters that govern comparative analyses. Coupled with a greater understanding of noncoding functional sequence, managing the trade-offs in genomic sequence comparisons will lead to the best analyses possible.

CONCLUSION

The many recent successes of comparative genomics foreshadow a generation of sophisticated analyses that will benefit from an expanding diversity of sequence. Existing studies of deep, orthologous data sets allude to the insights that future work will provide. Discoveries of pan-vertebrate and pan-mammalian functional genomic elements will yield understanding of fundamental biological processes. Identifying less ubiquitous elements can shed light on the genomic differences that distinguish related species, such as those that set primates apart from their mammalian cousins, or those that set humans apart from other primates. Furthermore, competency at characterizing elements under weak purifying selection may lead to an understanding of their role in complex traits and human diseases.

The resolution of functional elements is impacted by sequence diversity in a number of ways. As we become confident in regional estimates of the neutral rate and are able to utilize more realistic models of neutral evolution, our ability to resolve neutral sequence variation will improve. The concomitant reduction of false positives in comparative analyses will enhance specificity, lending confidence that conserved elements are functional even in the absence of a positive experimental assay. By fostering the discovery of further functional sequences, diversity will permit analyses to identify complex patterns and motifs that define classes of noncoding functional elements; incorporating this knowledge into future studies will lead to as-yet-unachieved sensitivity.

Managing the trade-offs inherent to comparative genomics will remain an important part of future analyses, especially in the discovery of functional elements under weak purifying selection. The ability to comprehensively identify and annotate functional sequences is intimately tied to our understanding of the human genome and the processes that have shaped its evolution. Knowledge of these elements and their pathological variants will have a major impact on the comprehension and treatment of human disease.

ACKNOWLEDGMENTS

EAS acknowledges support from the Stanford Genome Training Program (NIH/NHGRI). GMC is a Howard Hughes Medical Institute Pre-Doctoral fellow. AS acknowledges financial support from the NIH/NHGRI.

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

LITERATURE CITED

1. Abrahams BS, Mak GM, Berry ML, Palmquist DL, Saionz JR, et al. 2002. Novel vertebrate genes and putative regulatory elements identified at kidney disease and NR2E1/fierce loci. *Genomics* 80:45–53
2. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, et al. 2000. The genome sequence of *Drosophila melanogaster*. *Science* 287:2185–95
3. Amores A, Force A, Yan YL, Joly L, Amemiya C, et al. 1998. Zebrafish hox clusters and vertebrate genome evolution. *Science* 282:1711–14
4. Aparicio S, Chapman J, Stupka E, Putnam N, Chia JM, et al. 2002. Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. *Science* 297:1301–10
5. Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408:796–815
6. Aravind L, Watanabe H, Lipman DJ, Koonin EV. 2000. Lineage-specific loss and divergence of functionally linked genes in eukaryotes. *Proc. Natl. Acad. Sci. USA* 97:11319–24
7. Bagheri-Fam S, Ferraz C, Demaille J, Scherer G, Pfeifer D. 2001. Comparative genomics of the SOX9 region in human and Fugu rubripes: conservation of short regulatory sequence elements within large intergenic regions. *Genomics* 78:73–82
8. Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, et al. 2004. Ultraconserved elements in the human genome. *Science* 304:1321–25
9. Berezikov E, Guryev V, van de Belt J, Wienholds E, Plasterk RHA, Cuppen E. 2005. Phylogenetic shadowing and computational identification of human microRNA genes. *Cell* 120:21–24
10. Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF, et al. 2004. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* 14:708–15
11. Blattner FR, Plunkett G III, Bloch CA, Perna NT, Burland V, et al. 1997. The complete genome sequence of *Escherichia coli* K-12. *Science* 277:1453–74
12. 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
13. Boffelli D, Nobrega MA, Rubin EM. 2004. Comparative genomics at the vertebrate extremes. *Nat. Rev. Genet.* 5:456–65
14. Bourque G, Pevzner PA, Tesler G. 2004. Reconstructing the genomic architecture of ancestral mammals: lessons from human, mouse, and rat genomes. *Genome Res.* 14:507–16
15. Bray N, Pachter L. 2004. MAVID: constrained ancestral alignment of multiple sequences. *Genome Res.* 14:693–99
16. 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

17. Brudno M, Poliakov A, Salamov A, Cooper GM, Sidow A, et al. 2004. Automated whole-genome multiple alignment of rat, mouse, and human. *Genome Res.* 14:685–92
18. *C. elegans* Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. The *C. elegans* Sequencing Consortium. *Science* 282:2012–18
19. 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
20. Chapman MA, Donaldson IJ, Gilbert J, Grafham D, Rogers J, et al. 2004. Analysis of multiple genomic sequence alignments: a web resource, online tools, and lessons learned from analysis of mammalian SCL loci. *Genome Res.* 14:313–18
21. Collins FS, Green ED, Guttmacher AE, Guyer MS. 2003. A vision for the future of genomics research. *Nature* 422:835–47
22. Cooper GM, Brudno M, Green ED, Batzoglou S, Sidow A. 2003. Quantitative estimates of sequence divergence for comparative analyses of mammalian genomes. *Genome Res.* 13:813–20
23. Cooper GM, Sidow A. 2003. Genomic regulatory regions: insights from comparative sequence analysis. *Curr. Opin. Genet. Dev.* 13:604–10
24. Cooper GM, Brudno M, Stone EA, Dubchak I, Batzoglou S, Sidow A. 2004. Characterization of evolutionary rates and constraints in three mammalian genomes. *Genome Res.* 14:539–48
25. 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
26. Dermitzakis ET, Clark AG. 2002. Evolution of transcription factor binding sites in mammalian gene regulatory regions: conservation and turnover. *Mol. Biol. Evol.* 19:1114–21
27. Dermitzakis ET, Reymond A, Lyle R, Scamuffa N, Ucla C, et al. 2002. Numerous potentially functional but non-genic conserved sequences on human chromosome 21. *Nature* 420:578–82
28. 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
29. Eddy SR. 2002. Computational genomics of noncoding RNA genes. *Cell* 109:137–40
30. Eddy SR. 2005. A model of the statistical power of comparative genome sequence analysis. *PLoS Biol.* 3:e10
31. Efstratiadis A, Posakony JW, Maniatis T, Lawn RM, O'Connell C, et al. 1980. The structure and evolution of the human beta-globin gene family. *Cell* 21:653–68
32. Elnitski L, Hardison RC, Li J, Yang S, Kolbe D, et al. 2003. Distinguishing regulatory DNA from neutral sites. *Genome Res.* 13:64–72
33. Enard W, Paabo S. 2004. Comparative primate genomics. *Annu. Rev. Genomics Hum. Genet.* 5:351–78
34. ENCODE Project Consortium. 2004. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* 306:636–40
35. Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* 17:368–76
36. Felsenstein J, Churchill GA. 1996. A Hidden Markov Model approach to variation among sites in rate of evolution. *Mol. Biol. Evol.* 13:93–104
37. Fink GR. 1998. Anatomy of a revolution. *Genetics* 149:473–77
38. Frazer KA, Tao H, Osoegawa K, de Jong PJ, Chen X, et al. 2004. Noncoding

sequences conserved in a limited number of mammals in the SIM2 interval are frequently functional. *Genome Res.* 14:367–72

39. Friedman N, Ninio M, Pe'er I, Pupko T. 2002. A structural EM algorithm for phylogenetic inference. *J. Comput. Biol.* 9: 331–53
40. Ghanem N, Jarinova O, Amores A, Long Q, Hatch G, et al. 2003. Regulatory roles of conserved intergenic domains in vertebrate Dlx bigene clusters. *Genome Res.* 13:533–43
41. Glazko GV, Koonin EV, Rogozin IB, Shabalina SA. 2003. A significant fraction of conserved noncoding DNA in human and mouse consists of predicted matrix attachment regions. *Trends Genet.* 19:119–24
42. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, et al. 1996. Life with 6000 genes. *Science* 274:546, 563–67
43. 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
44. 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
45. Göttgens B, Barton LM, Chapman MA, Sinclair AM, Knudsen B, et al. 2002. Transcriptional regulation of the stem cell leukemia gene (SCL)–comparative analysis of five vertebrate SCL loci. *Genome Res.* 12:749–59
46. Hardison RC. 2000. Conserved noncoding sequences are reliable guides to regulatory elements. *Trends Genet.* 16:369–72
47. Hardison RC. 2003. Comparative genomics. *PLoS Biol.* 1:E58
48. Hoh J, Ott J. 2000. Scan statistics to scan markers for susceptibility genes. *Proc. Natl. Acad. Sci. USA* 97:9615–17
49. International Chicken Genome Sequencing Consortium. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* 432:695–716
50. International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* 409:860–921
51. International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. *Nature* 431:931–45
52. Jaillon O, Aury JM, Brunet F, Petit JL, Stange-Thomann N, et al. 2004. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* 431:946–57
53. Jojic V, Jojic N, Meek C, Geiger D, Siepel A, et al. 2004. Efficient approximations for learning phylogenetic HMM models from data. *Bioinformatics* 20(Suppl.1): I161–I68
54. Jukes TH, Cantor CR. 1969. Evolution of protein molecules. In *Mammalian Protein Metabolism*, ed. HN Munro, pp. 21–132. New York: Academic
55. 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
56. Kimura M. 1983. *The Neutral Theory of Molecular Evolution*. Cambridge [Cambridgeshire]; New York: Cambridge Univ. Press. 367 pp.
57. 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
58. Krylov DM, Wolf YI, Rogozin IB, Koonin EV. 2003. Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated

in eukaryotic evolution. *Genome Res.* 13: 2229–35

59. Kunst F, Ogasawara N, Moszer I, Albertini AM, Alloni G, et al. 1997. The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature* 390:249–56
60. Lawn RM, Boonmark NW, Schwartz K, Lindahl GE, Wade DP, et al. 1995. The recurring evolution of lipoprotein(a). Insights from cloning of hedgehog apolipoprotein(a). *J. Biol. Chem.* 270: 24004–9
61. Lien CL, McAnally J, Richardson JA, Olson EN. 2002. Cardiac-specific activity of an Nkx2–5 enhancer requires an evolutionarily conserved Smad binding site. *Dev. Biol.* 244:257–66
62. Lim LP, Glasner ME, Yekta S, Burge CB, Bartel DP. 2003. Vertebrate microRNA genes. *Science* 299:1540
63. 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
64. 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
65. Lowe TM, Eddy SR. 1999. A computational screen for methylation guide snoRNAs in yeast. *Science* 283:1168–71
66. Madsen O, Scally M, Douady CJ, Kao DJ, DeBry RW, et al. 2001. Parallel adaptive radiations in two major clades of placental mammals. *Nature* 409:610–14
67. Major MB, Jones DA. 2004. Identification of a gadd45beta 3′ enhancer that mediates SMAD3- and SMAD4-dependent transcriptional induction by transforming growth factor beta. *J. Biol. Chem.* 279:5278–87
68. Makalowski W, Zhang J, Boguski MS. 1996. Comparative analysis of 1196 orthologous mouse and human full-length mRNA and protein sequences. *Genome Res.* 6:846–57
69. Manly KF, Nettleton D, Hwang JT. 2004. Genomics, prior probability, and statistical tests of multiple hypotheses. *Genome Res.* 14:997–1001
70. Margulies EH, Blanchette M, Haussler D, Green ED. 2003. Identification and characterization of multi-species conserved sequences. *Genome Res.* 13:2507–18
71. McCue LA, Thompson W, Carmack CS, Lawrence CE. 2002. Factors influencing the identification of transcription factor binding sites by cross-species comparison. *Genome Res.* 12:1523–32
72. McCutcheon JP, Eddy SR. 2003. Computational identification of non-coding RNAs in *Saccharomyces cerevisiae* by comparative genomics. *Nucleic Acids Res.* 31:4119–28
73. Miller W, Makova KD, Nekrutenko A, Hardison RC. 2004. Comparative genomics. *Annu. Rev. Genomics Hum. Genet.* 5:15–56
74. Moses AM, Chiang DY, Pollard DA, Iyer VN, Eisen MB. 2004. MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. *Genome Biol.* 5:R98
75. Mouse Genome Sequencing Consortium. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* 420:520–62
76. Murphy WJ, Eizirik E, Johnson WE, Zhang YP, Ryder OA, O'Brien SJ. 2001. Molecular phylogenetics and the origins of placental mammals. *Nature* 409:614–18
77. Nobrega MA, Ovcharenko I, Afzal V, Rubin EM. 2003. Scanning human gene deserts for long-range enhancers. *Science* 302:413
78. 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

79. Ohler U, Yekta S, Lim LP, Bartel DP, Burge CB. 2004. Patterns of flanking sequence conservation and a characteristic upstream motif for microRNA gene identification. *RNA* 10:1309–22
80. Olsen GJ, Matsuda H, Hagstrom R, Overbeek R. 1994. fastDNAmL: a tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. *Comput. Appl. Biosci.* 10:41–48
81. Ovcharenko I, Boffelli D, Loots GG. 2004. eShadow: a tool for comparing closely related sequences. *Genome Res.* 14:1191–98
82. Ovcharenko I, Stubbs L, Loots GG. 2004. Interpreting mammalian evolution using Fugu genome comparisons. *Genomics* 84: 890–95
83. 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
84. Pennacchio LA, Olivier M, Hubacek JA, Cohen JC, Cox DR, et al. 2001. An apolipoprotein influencing triglycerides in humans and mice revealed by comparative sequencing. *Science* 294:169–73
85. Qiu Y, Cavelier L, Chiu S, Yang X, Rubin E, Cheng JF. 2001. Human and mouse ABCA1 comparative sequencing and transgenesis studies revealing novel regulatory sequences. *Genomics* 73:66–76
86. Rat Genome Project Sequencing Consortium. 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* 428:493–521
87. Sabarinadh C, Subramanian S, Tripathi A, Mishra RK. 2004. Extreme conservation of noncoding DNA near HoxD complex of vertebrates. *BMC Genomics* 5: 75
88. Sanger F, Coulson AR, Hong GF, Hill DF, Petersen GB. 1982. Nucleotide sequence of bacteriophage lambda DNA. *J. Mol. Biol.* 162:729–73
89. Santini S, Boore JL, Meyer A. 2003. Evolutionary conservation of regulatory elements in vertebrate Hox gene clusters. *Genome Res.* 13:1111–22
90. 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. *Nucleic Acids Res.* 31:3518–24
91. Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, et al. 2003. Human-mouse alignments with BLASTZ. *Genome Res.* 13:103–7
92. Shrager J, Hauser C, Chang CW, Harris EH, Davies J, et al. 2003. Chlamydomonas reinhardtii genome project. A guide to the generation and use of the cDNA information. *Plant Physiol.* 131: 401–8
93. Sidow A. 2002. Sequence first. Ask questions later. *Cell* 111:13
94. Sumiyama K, Kim CB, Ruddle FH. 2001. An efficient *cis*-element discovery method using multiple sequence comparisons based on evolutionary relationships. *Genomics* 71:260–62
95. Tautz D. 2000. Evolution of transcriptional regulation. *Curr. Opin. Genet. Dev.* 10:575–79
96. 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
97. Toyoda A, Noguchi H, Taylor TD, Ito T, Pletcher MT, et al. 2002. Comparative genomic sequence analysis of the human chromosome 21 down syndrome critical region. *Genome Res.* 12:1323–32
98. Ureta-Vidal A, Ettwiller L, Birney E. 2003. Comparative genomics: genome-wide analysis in metazoan eukaryotes. *Nat. Rev. Genet.* 4:251–62
99. Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, et al. 2005. Highly

conserved non-coding sequences are associated with vertebrate development. *PLoS Biol.* 3:e7

100. Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *CABIOS* 13:555–56

101. Yang Z, Rannala B. 1997. Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo Method. *Mol. Biol. Evol.* 14:717–24

102. Yekta S, Shih IH, Bartel DP. 2004. MicroRNA-directed cleavage of HOXB8 mRNA. *Science* 304:594–96