Perspective

# Qualifying the relationship between sequence conservation and molecular function

Gregory M. Cooper[1,3,4] and Christopher D. Brown[2,3]

[1]*Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA;* [2]*Institute for Genomics and Systems Biology, University of Chicago, Chicago, Illinois 60637, USA*

Quantification of evolutionary constraints via sequence conservation can be leveraged to annotate genomic functional sequences. Recent efforts addressing the converse of this relationship have identified many sites in metazoan genomes with molecular function but without detectable conservation between related species. Here, we discuss explanations and implications for these results considering both practical and theoretical issues. In particular, phylogenetic scope influences the relationship between sequence conservation and function. Comparisons of distantly related species can detect constraint with high specificity due to the loss of conserved neutral sequence, but sensitivity is sacrificed as a result of functional changes related to lineage-specific biology. The strength of natural selection operating on functional sequence is also important. Mutations to functional sequences that result in small fitness effects are subject to weaker constraints. Therefore, particularly when comparing highly divergent species, functional sequences that are degenerate or biologically redundant will be prone to turnover, wherein functional sequences are replaced by effectively equivalent, but nonorthologous counterparts. Finally, considering the size and complexity of metazoan genomes and the fact that many nonconserved sequences are associated with sequence-degenerate, low-level molecular functions, we find it likely that there exist many biochemically functional sequences that are not under constraint. This hypothesis does not lead to the conclusion that huge amounts of vertebrate genomes are functionally important, but rather that such "functionality" represents molecular noise that has weak or no effect on organismal phenotypes.

## Introduction

The identification of functional elements within large complex genomes has been aided by comparative genomics, in particular, via the quantification of evolutionary constraints (Pennacchio et al. 2001, 2006; Göttgens et al. 2002; Kellis et al. 2003). Recently, however, high-throughput functional genomics techniques have allowed for an initial assessment of the converse relationship, namely, the quantification of selective constraint on large, unbiased collections of functional elements. These studies include cell-based assays on a genomic scale (Kim et al. 2005; Borneman et al. 2007; The ENCODE Project Consortium 2007; Heintzman et al. 2007; Xi et al. 2007) and in vivo assays for developmentally important functions in individual loci in animal model organisms (Fisher et al. 2006; Brown et al. 2007). Interestingly, they have demonstrated that there are large numbers of functional sequences that are not detectably conserved across both distant (McGaughey et al. 2008, this issue) and close (Moses et al. 2006; Margulies et al. 2007) evolutionary timescales. This lack of conservation has several explanations in principle, each of which has distinct implications for functional annotation of complex genomes and a better understanding of genomic evolution. Here, we address these possibilities in light of variation in phylogenetic scope and the quantitative relationship between sequence function and evolutionary rate.

### The basic premise

The application of comparative sequence analysis to annotate genomic functional sequences is dependent upon the basic principles laid out by Kimura in the neutral theory of molecular evolution (Kimura 1983). Most evolutionary change between species is the result of mutations with minimal or no functional impact that are fixed via random genetic drift. In contrast, mutations in functional elements (e.g., exons, *cis*-regulatory elements) are likely to impair function, be deleterious to the organism, and subsequently be eliminated by purifying selection. The detection of sequences affected by purifying selection, which are said to be under evolutionary constraint, can therefore be used to annotate functional sites in genomes. Detection and quantification of constraint is usually accomplished through statistical evaluations of interspecific genomic sequence conservation. We note that it is important to distinguish "conservation," which is an observation of similarity, from "constraint," which is a hypothesis about the effects of purifying selection. Conservation, when observed to be in excess of the levels predicted by a neutral model, can be used to infer constraint. However, the presence of conservation does not necessarily imply constraint nor does its absence imply a lack of constraint. This distinction is critical to the interpretation of results from comparative genomic analyses. Indeed, conservation statistics should never be utilized in the absence of the context provided by the levels of neutral sequence conservation/divergence.

### Phylogenetic scope

One of the most important parameters of a comparative genomics study is phylogenetic scope, defined as the minimal evolutionary span that captures all of the included species. For example, analyses comparing sequence from human, mouse, and dog have a placental mammalian scope. Because constraint analyses require an assumption of orthology (or at the very least homology), the phylogenetic scope of the analysis enforces a

**[3]These authors contributed equally to this work.**
**[4]Corresponding author.**
**E-mail coopergm@u.washington.edu; fax (206) 221-5795.**
Article is online at http://www.genome.org/cgi/doi/10.1101/gr.7205808.

limit on sensitivity to only those functional sequences present in the species' last common ancestor. Phylogenetic scope is also correlated with levels of genomic sequence divergence, defined in this context as the average number of nucleotide changes affecting neutral sites. Since the inference of constraint requires a statistically significant difference between the conservation seen for neutral sites and that seen for constrained sites, the level of neutral sequence divergence is a direct contributor to the specificity of a sequence comparison.

Phylogenetic scope thus has direct, predictable consequences on specificity and sensitivity for a comparative analysis. It is difficult to measure the effects of selection on any given nucleotide of the human genome when comparing only closely related ape genome sequences (Eddy 2005; Stone et al. 2005), for example, as the vast majority of neutral nucleotides remain conserved between these species. On the other hand, comparisons between human and more distant vertebrates like fishes, or even among distantly related fishes like zebrafish and *Fugu*, are so divergent that neutral sites have been completely saturated with nucleotide changes (both substitutions and deletions), and any sequence that is reliably aligned between these species is almost certainly under constraint. However, such comparisons are known to miss a large number of highly constrained lineage-specific functional elements (Cooper et al. 2005). Thus, it should not be regarded as surprising that many functional elements are not conserved when comparing extremely distant species (e.g., as seen in McGaughey et al. 2008).

### Sequence function and evolutionary rate

The sensitivity of constraint-based methods to identify functional sequence is also dependent on the quantitative relationship between sequence function and evolutionary rate, which is mediated by the strength and efficacy of natural selection. In general, nucleotides with important molecular functions will evolve more slowly than the rate predicted by a neutral model. However, this is not a discrete phenomenon. The selection coefficient, a quantitative measure of the effects of selective pressure, varies continuously in relation to both the sensitivity of the molecular function to nucleotide change (degeneracy) and the importance of the molecular function to survival and reproductive success (dispensability). Quantitative variation in selection coefficients in turn produces quantitative variation in the rate of sequence change. That this is a generalizable property of both protein-coding and noncoding sequences is supported by several lines of evidence.

With respect to coding DNA, it is well established that proteins evolve at vastly different rates. Protein expression level, functional category, structural characteristics, and participation in intermolecular interactions have all been suggested to contribute to this evolutionary rate variation (Li 1997; Pal et al. 2001; Wall et al. 2005; Drummond et al. 2006; Kim et al. 2006). In addition, within a given protein, the rates of evolution of individual amino acids vary greatly, largely as a result of the structure-function requirements for a given amino acid at a particular position within that protein. For example, active sites of enzymes, DNA-binding domains of transcription factors, and residues important for structural maintenance evolve slowly, as substitutions in these residues are particularly deleterious (Suckow et al. 1996; Simon et al. 2002).

With respect to other classes of functional sequence, recent estimates suggest that 70% of the nucleotides evolving under purifying selection in mammalian genomes are not within exons of protein-coding genes ("noncoding") and, except for the extreme constraint seen on some critical proteins (e.g., histones), the range of selection coefficients affecting these positions appears similar to that for protein-coding DNA (Mouse Genome Sequencing Consortium 2002; Rat Genome Sequencing Project Consortium 2004; King et al. 2007). Furthermore, analysis of the regulatory function and biochemical specificity of individual transcription-factor binding sites also supports the presence of a quantitative spectrum of selective strength in noncoding functional sequences. Across transcription-factor binding sites, sites that contribute more to the total regulatory activity of a *cis*-regulatory element accumulate fewer substitutions than those that contribute less (Brown et al. 2007). In addition, nucleotide-by-nucleotide binding specificity within a transcription-factor binding site is inversely proportional to the evolutionary rate of the position (Mirny and Gelfand 2002; Moses et al. 2003).

## Interpreting nonconserved genomic functionality

Results from constraint-based comparative genomic analyses should be interpreted in light of the principles described above in addition to practical considerations. The discovery of many nonconserved functional sequences in metazoan genomes can thus be explained by several nonexclusive possibilities, including technical challenges, divergent biology related to phylogenetic scope, loss of conservation resulting from weak constraints, and unconstrained molecular functionality. We address each of these explanations in turn.

### Technical challenges

Some constrained functional elements are likely to be misclassified as nonconserved ("false negatives") by comparative sequence analyses due to technical challenges. Consider a small functional element (<10 bp) present within a long stretch of neutral sequence. Even if the element itself is highly constrained and persistent across a wide phylogenetic scope, without similar sequence nearby to provide a reliable alignment "anchor" (Batzoglou 2005), such an element would likely not manifest as a conserved sequence. While such obstacles are more problematic when comparing highly divergent species, they are not restricted to comparisons in wide scopes. Genomic sequence alignment, a prerequisite to any constraint-based analysis, remains a challenging problem even for relatively closely related species (Pollard et al. 2006; Margulies et al. 2007).

Experimental limitations also contribute to false negatives. For example, many functional genomic datasets are plagued by poor resolution: Transcription factor "binding sites" identified by "ChIP-chip" experiments, for example, often span hundreds of nucleotides, while the extent of a functional sequence is likely to be substantially smaller. This problem has been shown to obscure the relationship between constraint and function (Brown et al. 2007; Margulies et al. 2007) since the conservation signal indicative of constraint on the functional nucleotides is diluted by the noise resulting from the inclusion of many nonfunctional and neutrally evolving sites. In addition, nearly all sequence comparisons of functional sites derive functional annotation from only one species. Simultaneous annotation of function independently in multiple species can significantly clarify the relationship between sequence conservation and molecular function, contrasting conservation that may simply be obscured due to technical

challenges (Brown et al. 2007) from legitimate primary sequence turnover of functional binding sites (Borneman et al. 2007; Odom et al. 2007).

### Divergent biology

Pathway modularity and functional exaptation notwithstanding, functional elements that relate to environmental, developmental, physiological, or other biological factors that are not common to the entire phylogenetic scope of an analysis are likely to be systematically missed. Indeed, it has been shown that many regulatory elements in the human genome are restricted to particular clades and are likely to play important, but clade-specific roles (King et al. 2007); sequences involved in the articulation of digits in the developing mammalian limb bud are unlikely to be systematically captured in a human–fish comparison, for example. Even for those elements present in the common ancestral genome, changes in genomic or biological context that alter the strength of selection are likely to be major contributors to a loss of sensitivity in the detection of constraint. Lineage-specific loss of function, for example, can have a major effect on sensitivity even when only a minor subset of the analyzed lineages is affected (Stone et al. 2005). Additionally, even for functionality that is persistent across the entire phylogenetic scope, changes in genomic context can lead to decreased sensitivity. Duplication events, a prominent feature in the evolution of genomes (Ohno 1970; Wolfe and Shields 1997; Dehal and Boore 2005), in principle, allow for relaxed constraint on one or both copies of a duplicated functional element (Lynch and Conery 2000; Kondrashov et al. 2002). As such, inclusion of only one member of a lineage-specific duplicated sequence, as is routinely done by the popular genomic sequence alignment tools (Margulies et al. 2007), will provide an incomplete picture of the constraint–function relationship.

### Weak constraints

The strength of selection operating on any particular genomic sequence is related to both the sequence degeneracy and organismal importance of its molecular function. As such, it is anticipated that functional sequences that have a small influence on organismal fitness or are sequence-degenerate will be under weaker evolutionary constraints and thus more likely to change or "turnover" as the amount of neutral divergence increases. For example, enhancer sequences that contribute only a small portion of the total regulatory information for a given gene have been shown to evolve more swiftly than enhancers with larger effect, even when they regulate genes with critical developmental function (Brown et al. 2007). Additionally, consider transcriptional promoters of human protein-coding genes (Trinklein et al. 2003; Kim et al. 2005): While these regions are important for transcriptional regulation and strongly enriched for constrained sequences, many individual promoters lack strong sequence conservation, even among placental mammals (The ENCODE Project Consortium 2007), and may be influenced by a significant level of individual binding site changes (Odom et al. 2007). This is likely a consequence of flexibility in sequence that can give rise to promoter function relating to either the sequence degeneracy or redundancy of individual functional elements. An additional possibility is the need for secondary structural or other characteristics that are only indirectly related to primary sequence (e.g., Greenbaum et al. 2007). Altogether, these observations suggest that promoter sequences are generally under constraint and as a class evolve more slowly than neutral DNA, but possess enough sequence degeneracy such that they are affected by a significant level of nucleotide divergence.

### Unconstrained molecular functionality

We speculate that there may be many sequences capable of molecular function in complex genomes, but lacking any significant effect on organismal fitness. Such sequences would evolve neutrally and therefore contribute to the discovery of nonconserved functional sequences. For example, recent studies in human cells describe extensive but low-level transcriptional activity spread across the genome, the vast majority of which yields no detectable signals of evolutionary constraint in mammalian genomic sequence (The ENCODE Project Consortium 2007; Kapranov et al. 2007; Margulies et al. 2007). While it certainly is possible that some of these functional sequences are under constraint but appear to be false negatives for reasons described above, two observations support the idea that many are truly not under constraint. First, some classes of experimentally annotated functional sequences fail to show enrichment for constrained nucleotides (Margulies et al. 2007). If these elements were truly, but weakly constrained, some enrichment would be expected, as is seen for promoters of protein-coding genes. Second, bulk distribution analyses comparing rates of evolution in ancient mobile element insertion fragments ("ancestral repeats" or "ARs") to those in unique sequence find that there are unlikely to be a large number of truly constrained bases in the human genome that are not currently annotated (The ENCODE Project Consortium 2007). While it is clear that some ARs include functionally constrained DNA (Cooper et al. 2005; Bejerano et al. 2006; Xie et al. 2006), most are unlikely to possess specific and important molecular functions. Considering then that they can often be recognized as orthologous, alignable DNA amongst related mammals, ARs are likely to constitute a good empirical model for neutral evolution. This hypothesis is supported by the global regional correlations between rates of evolution at these sites and synonymous sites in protein-coding genes, and also a strong concordancy of results between AR-based and independently constructed null models (Mouse Genome Sequencing Consortium 2002; Hardison et al. 2003; Margulies et al. 2007).

If these functional sequences are under little to no constraint, it becomes critical to characterize their origins. One possibility is a result of the interplay between functional degeneracy and genome size and complexity. Indeed, given the impossibility of perfect molecular fidelity, we speculate that such "molecular noise" must be a common phenomenon, particularly for those functions that would arise frequently in large genomes at random and have a very minimal impact on the overall molecular activity of the cell. For example, given the variety of primary sequences that can give rise to their function, there are likely to be many transcriptional promoters occurring at random in the human genome; in fact, mobile elements like *Alu*s are capable of some promoter function, and randomly selected fragments of the human genome often show at least minimal promoter activity (Smit 1996; Khambata-Ford et al. 2003). Furthermore, such events may even show reproducible spatiotemporal specificity due to differential local chromatin regulation (Thurman et al. 2007). Thus, it is plausible, if not likely, to expect low levels of reproducible transcriptional activity and weak protein–DNA binding widely distributed across large complex genomes with no particular purpose.

## Conclusions

Large amounts of sequence data have provided a wealth of insights into the evolution of genes and genomes from the perspective of mutation and divergence. Improvements in functional genomics technologies and the development of appropriate model systems promise to provide similar insights from the perspective of molecular function. Recent efforts adopting an unbiased approach to discover functional sequences in complex genomes are already providing a glimpse of such insights. While of tremendous interest, we argue that the discovery of nonconserved functional sequences is largely in line with expectations.

First, we note that these results highlight gaps in our current data and analytical tools and the need for careful study design. Improved computational techniques related to sequence alignment and genomic sequence data from additional species will significantly boost the sensitivity to detect constrained and, therefore, functional sequences (Boffelli et al. 2003). Comparative studies of model organisms that are currently restricted to extreme phylogenetic scopes would benefit tremendously from additional genome sequences from more closely related species, such as the recent effort to surround the *Drosophila melanogaster* genome sequence with data from many other *Drosophilids* (*Drosophila* 12 Genomes Consortium 2007). Additionally, higher-resolution functional annotations and the development of experimental platforms for model organism "sister" species are also likely to clarify this relationship (Brown et al. 2007; Margulies et al. 2007). Second, these results also point to the influence of functional sequence turnover (Ludwig et al. 2000; Moses et al. 2006; Odom et al. 2007). We note that this phenomenon may apply to even developmentally important functionality, particularly for comparisons of distantly related species to discover elements that are individually minor contributors to the overall functional output (McGaughey et al. 2008).

Finally, we speculate that there are many functional sequences that are unlikely to have a major phenotypic effect and are therefore of minimal or no relevance to organismal fitness. It is important to keep in mind that we are not suggesting that such "molecular noise" is irrelevant to biology. Quite to the contrary, beyond the fact that characterizing these functions is necessary for a more complete understanding of biology, it seems possible that such "background" functionality serves some more general role. Synonymous sites in protein-coding DNA are often considered to be neutral (Kimura 1983), for example, but serve the abstract, yet critical function of generating a richer genetic code. Additionally, sequences with subtle molecular functionality may constitute a set of elements adaptable for the generation of novel genes or regulatory elements; mobile element activity may play a role in recruiting new genes to particular regulatory networks (Wang et al. 2007), for example, and there exists at least one example of a "promoter-like" sequence that is turned into a novel functional element (albeit pathogenic) via a single-nucleotide change in humans (De Gobbi et al. 2006). In any case, the accumulation of neutral "functional" changes is likely to be a common and important biological phenomenon. This idea has already received support from analyzing transcriptional "drift" in the evolution of humans and chimpanzees (Khaitovich et al. 2004). Much as the neutral theory of molecular evolution emphasized the role of chance in the evolution of genomic sequences, such a model seems appropriate as the default interpretation for the evolution of genomic function.

## Acknowledgments

We thank Mark Rieder, Arend Sidow, Nadia Singh, and three anonymous reviewers for helpful comments on the manuscript. G.M.C. is supported by a Merck, Jane Coffin Childs Memorial Fund Fellowship.

## References

Batzoglou, S. 2005. The many faces of sequence alignment. *Brief. Bioinform.* **6:** 6–22.

Bejerano, G., Lowe, C.B., Ahituv, N., King, B., Siepel, A., Salama, S.R., Rubin, E.M., Kent, W.J., and Haussler, D. 2006. A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature* **441:** 87–90.

Boffelli, D., McAuliffe, J., Ovcharenko, D., Lewis, K.D., Ovcharenko, I., Pachter, L., and Rubin, E.M. 2003. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* **299:** 1391–1394.

Borneman, A.R., Gianoulis, T.A., Zhang, Z.D., Yu, H., Rozowsky, J., Seringhaus, M.R., Wang, L.Y., Gerstein, M., and Snyder, M. 2007. Divergence of transcription factor binding sites across related yeast species. *Science* **317:** 815–819.

Brown, C.D., Johnson, D.S., and Sidow, A. 2007. Functional architecture and evolution of transcriptional elements that drive gene coexpression. *Science* **317:** 1557–1560.

Cooper, G.M., Stone, E.A., Asimenos, G., Green, E.D., Batzoglou, S., and Sidow, A. 2005. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* **15:** 901–913.

De Gobbi, M., Viprakasit, V., Hughes, J.R., Fisher, C., Buckle, V.J., Ayyub, H., Gibbons, R.J., Vernimmen, D., Yoshinaga, Y., de Jong, P., et al. 2006. A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter. *Science* **312:** 1215–1217.

Dehal, P. and Boore, J.L. 2005. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol.* **3:** e314. doi: 10.1370/journal.pbio.0030314.

*Drosophila* 12 Genomes Consortium. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* **450:** 203–218.

Drummond, D.A., Raval, A., and Wilke, C.O. 2006. A single determinant dominates the rate of yeast protein evolution. *Mol. Biol. Evol.* **23:** 327–337.

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

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

Fisher, S., Grice, E.A., Vinton, R.M., Bessling, S.L., and McCallion, A.S. 2006. Conservation of RET regulatory function from human to zebrafish without sequence similarity. *Science* **312:** 276–279.

Göttgens, B., Barton, L.M., Chapman, M.A., Sinclair, A.M., Knudsen, B., Grafham, D., Gilbert, J.G., Rogers, J., Bentley, D.R., and Green, A.R. 2002. Transcriptional regulation of the stem cell leukemia gene (SCL)—Comparative analysis of five vertebrate SCL loci. *Genome Res.* **12:** 749–759.

Greenbaum, J.A., Parker, S.C., and Tullius, T.D. 2007. Detection of DNA structural motifs in functional genomic elements. *Genome Res.* **17:** 940–946.

Hardison, R.C., Roskin, K.M., Yang, S., Diekhans, M., Kent, W.J., Weber, R., Elnitski, L., Li, J., O'Connor, M., Kolbe, D., et al. 2003. Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. *Genome Res.* **13:** 13–26.

Heintzman, N.D., Stuart, R.K., Hon, G., Fu, Y., Ching, C.W., Hawkins, R.D., Barrera, L.O., Van Calcar, S., Qu, C., Ching, K.A., et al. 2007. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* **39:** 311–318.

Kapranov, P., Cheng, J., Dike, S., Nix, D.A., Duttagupta, R., Willingham, A.T., Stadler, P.F., Hertel, J., Hackermuller, J., Hofacker, I.L., et al. 2007. RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* **316:** 1484–1488.

Kellis, M., Patterson, N., Endrizzi, M., Birren, B., and Lander, E.S. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423:** 241–254.

Khaitovich, P., Weiss, G., Lachmann, M., Hellmann, I., Enard, W., Muetzel, B., Wirkner, U., Ansorge, W., and Paabo, S. 2004. A neutral model of transcriptome evolution. *PLoS Biol.* **2:** e132. doi: 10.1371/journal.pbio.0020132.

Khambata-Ford, S., Liu, Y., Gleason, C., Dickson, M., Altman, R.B., Batzoglou, S., and Myers, R.M. 2003. Identification of promoter regions in the human genome by using a retroviral plasmid library-based functional reporter gene assay. *Genome Res.* **13:** 1765–1774.

Kim, T.H., Barrera, L.O., Zheng, M., Qu, C., Singer, M.A., Richmond, T.A., Wu, Y., Green, R.D., and Ren, B. 2005. A high-resolution map of active promoters in the human genome. *Nature* **436:** 876–880.

Kim, P.M., Lu, L.J., Xia, Y., and Gerstein, M.B. 2006. Relating three-dimensional structures to protein networks provides evolutionary insights. *Science* **314:** 1938–1941.

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

King, D.C., Taylor, J., Zhang, Y., Cheng, Y., Lawson, H.A., Martin, J., ENCODE groups for Transcriptional Regulation and Multispecies Analysis, Chiaromonte, F., Miller, W., and Hardison, R.C. 2007. Finding *cis*-regulatory elements using comparative genomics: Some lessons from ENCODE data. *Genome Res.* **17:** 775–786.

Kondrashov, F.A., Rogozin, I.B., Wolf, Y.I., and Koonin, E.V. 2002. Selection in the evolution of gene duplications. *Genome Biol.* **3:** RESEARCH0008. doi: 10.1186/gb-2002-3-2-research0008.

Li, W.-H. 1997. *Molecular evolution*. Sinauer Associates, Sunderland, MA.

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

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

Margulies, E.H., Cooper, G.M., Asimenos, G., Thomas, D.J., Dewey, C.N., Siepel, A., Birney, E., Keefe, D., Schwartz, A.S., Hou, M., et al. 2007. Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. *Genome Res.* **17:** 760–774.

McGaughey, D.M., Vinton, R.M., Huynh, J., Al-Saif, A., Beer, M.A., and McCallion, A.S. 2008. Metrics of sequence constraint overlook regulatory sequences in an exhaustive analysis at *phox2b*. *Genome Res.* **18:** (this issue). doi: 10.1101/gr.6929408.

Mirny, L.A. and Gelfand, M.S. 2002. Structural analysis of conserved base pairs in protein–DNA complexes. *Nucleic Acids Res.* **30:** 1704–1711.

Mouse Genome Sequencing Consortium. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.

Moses, A.M., Chiang, D.Y., Kellis, M., Lander, E.S., and Eisen, M.B. 2003. Position specific variation in the rate of evolution in transcription factor binding sites. *BMC Evol. Biol.* **3:** 19. doi: 10.1186/1471-2148-3-19.

Moses, A.M., Pollard, D.A., Nix, D.A., Iyer, V.N., Li, X.Y., Biggin, M.D., and Eisen, M.B. 2006. Large-scale turnover of functional transcription factor binding sites in *Drosophila*. *PLoS Comput. Biol.* **2:** e130. doi: 10.1371/journal.pcbi.0020130.

Odom, D.T., Dowell, R.D., Jacobsen, E.S., Gordon, W., Danford, T.W., MacIsaac, K.D., Rolfe, P.A., Conboy, C.M., Gifford, D.K., and Fraeknel, E. 2007. Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nat. Genet.* **39:** 730–732.

Ohno, S. 1970. *Evolution by gene duplication*. Springer-Verlag, Berlin/New York.

Pal, C., Papp, B., and Hurst, L.D. 2001. Highly expressed genes in yeast evolve slowly. *Genetics* **158:** 927–931.

Pennacchio, L.A., Olivier, M., Hubacek, J.A., Cohen, J.C., Cox, D.R., Fruchart, J.C., Krauss, R.M., and Rubin, E.M. 2001. An apolipoprotein influencing triglycerides in humans and mice revealed by comparative sequencing. *Science* **294:** 169–173.

Pennacchio, L.A., Ahituv, N., Moses, A.M., Prabhakar, S., Nobrega, M.A., Shoukry, M., Minovitsky, S., Dubchak, I., Holt, A., Lewis, K.D., et al. 2006. In vivo enhancer analysis of human conserved non-coding sequences. *Nature* **444:** 499–502.

Pollard, D.A., Moses, A.M., Iyer, V.N., and Eisen, M.B. 2006. Detecting the limits of regulatory element conservation and divergence estimation using pairwise and multiple alignments. *BMC Bioinformatics* **7:** 376.

Rat Genome Sequencing Project Consortium. 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428:** 493–521.

Simon, A.L., Stone, E.A., and Sidow, A. 2002. Inference of functional regions in proteins by quantification of evolutionary constraints. *Proc. Natl. Acad. Sci.* **99:** 2912–2917.

Smit, A.F. 1996. The origin of interspersed repeats in the human genome. *Curr. Opin. Genet. Dev.* **6:** 743–748.

Stone, E.A., Cooper, G.M., and Sidow, A. 2005. Trade-offs in detecting evolutionarily constrained sequence by comparative genomics. *Annu. Rev. Genomics Hum. Genet.* **6:** 143–164.

Suckow, J., Markiewicz, P., Kleina, L.G., Miller, J., Kisters-Woike, B., and Muller-Hill, B. 1996. Genetic studies of the Lac repressor. XV: 4000 single amino acid substitutions and analysis of the resulting phenotypes on the basis of the protein structure. *J. Mol. Biol.* **261:** 509–523.

Thurman, R.E., Day, N., Noble, W.S., and Stamatoyannopoulos, J.A. 2007. Identification of higher-order functional domains in the human ENCODE regions. *Genome Res.* **17:** 917–927.

Trinklein, N.D., Aldred, S.J., Saldanha, A.J., and Myers, R.M. 2003. Identification and functional analysis of human transcriptional promoters. *Genome Res.* **13:** 308–312.

Wall, D.P., Hirsh, A.E., Fraser, H.B., Kumm, J., Giaever, G., Eisen, M.B., and Feldman, M.W. 2005. Functional genomic analysis of the rates of protein evolution. *Proc. Natl. Acad. Sci.* **102:** 5483–5488.

Wang, T., Zeng, J., Lowe, C.B., Sellers, R.G., Salama, S.R., Yang, M., Burgess, S.M., Brachmann, R.K., and Haussler, D. 2007. Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. *Proc. Natl. Acad. Sci.* **104:** 18613–18618.

Wolfe, K.H. and Shields, D.C. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387:** 708–713.

Xi, H., Shulha, H.P., Lin, J.M., Vales, T.R., Fu, Y., Bodine, D.M., McKay, R.D., Chenoweth, J.G., Tesar, P.J., Furey, T.S., et al. 2007. Identification and characterization of cell type-specific and ubiquitous chromatin regulatory structures in the human genome. *PLoS Genet.* **3:** e136. doi: 10.1371/journal.pgen.0030136.

Xie, X., Kamal, M., and Lander, E.S. 2006. A family of conserved noncoding elements derived from an ancient transposable element. *Proc. Natl. Acad. Sci.* **103:** 11659–11664.