# HHS Public Access

# A high-resolution map of human evolutionary constraint using 29 mammals

*A full list of authors and affiliations appears at the end of the article.*

## Abstract

Comparison of related genomes has emerged as a powerful lens for genome interpretation. Here, we report the sequencing and comparative analysis of 29 eutherian genomes. We confirm that at least 5.5% of the human genome has undergone purifying selection, and report constrained elements covering ~4.2% of the genome. We use evolutionary signatures and comparison with experimental datasets to suggest candidate functions for ~60% of constrained bases. These elements reveal a small number of new coding exons, candidate stop codon readthrough events, and over 10,000 regions of overlapping synonymous constraint within protein-coding exons. We find 220 candidate RNA structural families, and nearly a million elements overlapping potential promoter, enhancer and insulator regions. We report specific amino acid residues that have undergone positive selection, 280,000 non-coding elements exapted from mobile elements, and ~1,000 primate- and human-accelerated elements. Overlap with disease-associated variants suggests our findings will be relevant for studies of human biology and health.

## Introduction

A key goal in understanding the human genome is to discover and interpret all functional elements encoded within its sequence. While only ~1.5% of the human genome encodes protein sequence[1], comparative analysis with the mouse[2], rat[3] and dog[4] genomes showed

**Additional contributors:** Broad Institute Sequencing Platform and Whole Genome Assembly Team: Jen Baldwin[1], Toby Bloom[1], Chee Whye Chin[1], Dave Heiman[1], Robert Nicol[1], Chad Nusbaum[1], Sarah Young[1], Jane Wilkinson[1].

Baylor College of Medicine Human Genome Sequencing Center: Andrew Cree[14], Huyn H. Dihn[14], Gerald Fowler[14], Shalili Jhangiani[14], Vandita Joshi[14], Sandra Lee[14], Lora R. Lewis[14], Lynne V. Nazareth[14], Geoffrey Okwuonu[14], Jireh Santibanez[14].

Genome Institute at Washington University: Kim Delehaunty[15], David Dooling[15], Catrina Fronik[15], Lucinda Fulton[15], Bob Fulton[15], Tina Graves[15], Patrick Minx[15], Erica Sodergren[13,15].

that at least 5% is under purifying selection and thus likely functional, of which ~3.5% consists of non-coding elements with likely regulatory roles. Detecting and interpreting these elements is particularly relevant to medicine, as loci identified in genome-wide association studies (GWAS) frequently lie in non-coding sequence[5].

Whereas initial comparative mammalian studies could estimate the overall proportion of the genome under evolutionary constraint, they had little power to detect most of the constrained elements – especially the smaller ones. Thus, they focused only on the top 5% of constrained sequence, corresponding to less than ~0.2% of the genome[4,6]. In 2005, we began an effort to generate sequence from a large collection of mammalian genomes with the specific goal of identifying and interpreting functional elements in the human genome based on their evolutionary signatures[7-8]. Here, we report our results to systematically characterize mammalian constraint, using 29 eutherian (placental) genomes. We identify 4.2% of the genome as constrained and ascribe potential function to ~60% of these bases using diverse lines of evidence for protein-coding, RNA, regulatory and chromatin roles, and we present evidence of exaptation and accelerated evolution. All datasets described here are publicly available in a comprehensive set at Broad Institute and UCSC (see below for links).

## Sequencing, assembly and alignment

We generated genome sequence assemblies for 29 mammalian species selected to achieve maximum divergence across the four major mammalian clades (Figure 1a, Text S1 and Table S1). For nine species, we used genome assemblies based on ~7-fold coverage shotgun sequence, and for 20 species we generated ~2-fold coverage (2X), to maximize the number of species sequenced with available resources on capillary machines. Twenty genomes are first reported here, and nine were previously described (See supplement). 0

The power to detect constrained elements depends largely on the total branch length of the phylogenetic tree connecting the species[9]. The 29 mammals correspond to a total effective branch length of ~4.5 substitutions per site, compared to ~0.68 for human-mouse-rat-dog (HMRD), and thus should offer greater power to detect evolutionary constraint: the probability that a genomic sequence not under purifying selection will remain fixed across all 29 species is $P_1 < 0.02$ for single bases and $P_{12} < 10^{-25}$ for 12-mers, compared to $P_1 \sim 0.50$ and $P_{12} \sim 10^{-3}$ for HMRD.

For 2x mammals, our assisted assembly approach[10] resulted in a typical contig size $N50_C$ of 2.8 kb and a typical scaffold size $N50_S$ of 51.8 kb (Text S2 and Table S1) and high sequence accuracy (96% of bases had a <1% error rate = Q20)[11]. Compared to high-quality sequence across the 30 Mb of the ENCODE pilot project[12], we estimated average error rates of 1-3 miscalled bases per kilobase[11], which is ~50-fold lower than the typical nucleotide sequence difference between the species, enabling high-confidence detection of evolutionary constraint (Text S3).

We based our analysis on whole-genome alignments by MultiZ (Text S4). The average number of aligned species was 20.9 at protein-coding positions in the human genome and 23.9 at the top 5% HMRD-conserved non-coding positions, with an average branch length

of 4.3 substitutions per base in these regions (Figure S1, S2). By contrast, whole-genome average alignment depth is only 17.1 species with 2.9 substitutions per site, likely due to large deletions in non-functional regions[4]. The depth at ancestral repeats is 11.4 (Figure S1a) consistent with repeats being largely non-functional[2,4].

### Detection of constrained sequence

Our analysis did not substantially change the estimate of the *proportion* of genome under selection. By comparing genome-wide conservation to that of ancestral repeats, we estimated the overall fraction of the genome under evolutionary constraint to be 5.36% at 50-bp windows (5.44% at 12-bp windows), using the SiPhy-$\omega$ statistic[13], a measure of overall substitution rate (Figure S3), consistent with previous similar estimates[2,4,14]. However, alternative methods[15-16] and different ways of correcting for the varying alignment depths give higher estimates (see Text S5 for details).

The additional species had a dramatic effect on our ability to identify the *specific elements under constraint*. With 29 mammals, we identify 3.6 million elements spanning 4.2% of the genome, at a finer resolution of 12 bp (Figure 1b, Text S6, Figure S4, Table S2, S3), compared to <0.1% of the genome for HMRD 12-bp elements and 2.0% for HMRD 50-bp elements[4]. Elements previously detected using five vertebrates[17] also detect a larger fraction of the genome (~4.1%), but only cover 45% of the mammalian elements detected here, suggesting a large fraction of our elements are mammalian-specific. The mean element size (36bp) is considerably shorter than both previously-detected HMRD elements (123bp) and five-vertebrate elements (104bp)[17]. For example, it is now possible to detect individual binding sites for the neuron-restrictive silencer factor (NRSF) in the promoter of the *NPAS4* gene, which are beyond detection power in previous datasets (Figure 2, Figure S5). We found a similar regional distribution of 12-bp elements (including the 2.6 million newly-detected constrained elements) to previously-detected HMRD elements (r = 0.94, Figure S6),. Similar results were obtained with the PhastCons[17] statistic (see Text S6).

Using a new method, SiPhy-$\pi$, sensitive not just to the substitution rate but also to biases in the substitution pattern (e.g. Figure S7), we detected an additional 1.3% of the human genome in constrained elements (see Table S2, S3). Most of the newly-detected constrained nucleotides extend elements found by rate-based methods, but 22% consist of new elements (average length 17 bp), and are enriched in noncoding regions.

**Constraint within the human population—**We observed that the evolutionary constraint acting on the 29 mammals is correlated with constraint within the human population, as assessed from human polymorphism data (Text S7) and consistent with previous studies[18]. Mammalian constrained elements show a depletion in single-nucleotide polymorphisms (SNPs)[19], and more constrained elements show even greater depletion. For example, in the top 1% most-strongly-conserved non-coding regions, SNPs occur at a 1.9-fold lower rate than the genome average, and the derived alleles have a lower frequency, consistent with purifying selection at many of these sites in the human genome.

Moreover, at positions with biased substitution patterns across mammals, the observed human SNPs show a similar bias to the one observed across mammals (Figure S7). Thus, not

only are constrained regions less likely to exhibit polymorphism in humans, but when such polymorphisms are observed, the derived alleles in humans tend to match the alleles present in non-human mammals, indicating a preference for the same alleles across both mammalian and human evolution.

## Functional annotation of constraint

We first studied the overlap of the 3.6 million evolutionarily constrained elements ($\omega < 0.8$ with $P < 10^{-15}$) with known gene annotations (Figure 1b). Roughly 30% of constrained elements were associated with protein-coding transcripts: ~25.3% overlap mature mRNAs (including 19.6% in coding exons, 1.2% in 5′-UTRs, and 4.4% in 3′-UTRs), and an additional 4.4% reside within 2 kb of transcriptional start sites (1.2% of which is within 200 bases).

The majority of constrained elements however reside in intronic and intergenic regions (29.7% and 38.6%, respectively). To study their biological roles and provide potential starting points to understand these large and mostly uncharted territories, we next studied their overlap with *evolutionary signatures*[7-8,20-21] characteristic of specific types of features and a growing collection of public large-scale experimental data.

**Protein-coding genes and exons—**Despite intense efforts to annotate protein-coding genes over the past decade[20,22-24], we detected 3,788 candidate new exons (a 2% increase) using evolutionary signatures characteristic of protein-coding exons[25]. Of these, 54% reside outside protein-coding genes, 19% within introns, and 13% in UTRs of known coding genes (Text S8, Table S4, S5). Our methods recovered 92% of known coding exons that were >10 codons and that fall in syntenic regions, the remainder showing non-consensus splice sites, unusual features, or poor conservation.

The majority of new exon candidates (>58%) are supported by evidence of transcription measured in 16 human tissues[26] (Figure S8a) or similarity to known Pfam protein domains. 31% of intronic and 13% of intergenic predictions extend known transcripts, and 5% and 11% respectively reside in new transcript models. The newly detected exons are more tissue-specific than known exons (mean of 3 tissues, vs. 12) and are expressed at 5-fold lower levels. Directed experiments and manual curation will be required to complete the annotation of the few hundred protein-coding genes that likely remain unannotated[27].

We found apparent stop codon readthrough[28] of four genes based on continued protein-coding constraint after an initial conserved stop codon[29] and until a subsequent stop codon (Text S9, Figure S8b). Readthrough in *SACM1L* could be triggered by an 80-base conserved RNA stem loop predicted by RNAz[30], lying four bases downstream of the readthrough stop codon.

We also detected coding regions with a very low synonymous substitution rate, indicating additional sequence constraints beyond the amino acid level (Text S9). We found >10,000 such synonymous constraint elements (SCEs) in more than one-quarter of all human genes[31]. Initial analysis suggests potential roles in splicing regulation (34% span an exon-exon junction), A-to-I editing, microRNA (miRNA) targeting, and developmental

regulation. *Hox* genes contain several top candidates (Figure 3a), including two previously-validated developmental enhancers[32-33].

**RNA structures and families of structural elements—**We next used evolutionary signatures characteristic of conserved RNA secondary structures[34] to reveal 37,381 candidate structural elements (Text S10, Figure S9a), covering ~1% of constrained regions. For example, the *XIST* lincRNA, known to bind chromatin and enable X-inactivation[35], contains a newly-predicted structure in its 5′ end (Figure S9bc), distinct from other known structures[36], that seems to be the source of chromatin-associated short RNAs[37].

Sequence- and structure-based clustering of predictions outside protein-coding exons revealed 1,192 novel families of structural RNAs (Text S10). We focused on a high-scoring subset consisting of 220 families with 725 instances, which also showed the highest thermodynamic stability[30] (Figure S9a, S10), DNase hypersensitivity, expression pattern correlation across tissues and intergenic expression enrichment (Figure S9a). We also expanded both known and novel families by including additional members detected by homology to existing members.

Noteworthy examples include: a glycyl-tRNA family, including a new member in *POP1,* involved in tRNA maturation, and likely involved in feedback regulation of *POP1*; three intronic families of long hairpins in ion-channel genes known to undergo A-to-I RNA editing and possibly involved in regulation of the editing event; an additional member of a family of 5′UTR hairpins overlapping the start codon of collagen genes and potential new miRNA genes that extend existing families[37].

Two of the largest novel families consist of short AU-rich hairpins of 6-7 bp that share the same strong consensus motif in their stem. These occur in the 3′UTRs of genes in several inflammatory response pathways, whose post-transcriptional regulation often involves structural AU-rich elements (AREs). Indeed, two homologous hairpins in *TNF* and *CSF3* correspond to known mRNA-destabilization elements, suggesting roles in mRNA stability for the two families[37].

Lastly, a family of six conserved hairpin structures (Figure S9d) was found in the 3′UTR of the *MAT2A* gene[37], which is involved in the synthesis of S-adenosyl-methionine (SAM), the primary methyl donor in human cells. All six hairpins consist of a 12-18 bp-stem and a 14-bp loop region with a deeply-conserved sequence motif (Figure S9), and may be involved in sensing SAM concentrations, which are known to affect *MAT2A* mRNA stability[38].

**Conservation patterns in promoters—**As different types of conservation in promoters may imply distinct biological functions[39], we classified the patterns of conservation within core promoters into three categories: those with uniformly 'high' constraint (7,635 genes, 13,996 transcripts), uniformly 'low' constraint (2,879 genes, 4,135 transcripts), and 'intermittent' constraint, consisting of alternating peaks and troughs of conservation (14,271 genes and 29,814 transcripts) (Figure S11a). 'High' and 'intermittent' constraint promoters are both associated with CpG islands (~66%), while 'low' constraint promoters have

significantly lower overlap (~41%), and all three classes show similar overlap with functional TATA boxes (2-3%, see Text S11).

These groups show distinct Gene Ontology enrichments (Figure S11b), with high-constraint promoters involved in development ($P_{bonf}<10^{-30}$), intermittent-constraint in basic cellular functions ($P_{bonf.}<5\times10^{-4}$), and low-constraint promoters in immunity, reproduction and perception, functions expected to be under positive selection and lineage-specific adaptation[2].

High constraint may reflect cooperative binding of many densely-binding factors, as previously suggested for developmental genes[6]. Intermittent constraint promoters, whose peak-spacing distribution was suggestive of the periodicity of the DNA helix turns, may reflect loosely-interacting factors (Figure S11cd). Low constraint may reflect rapid motif turnover, under neutral drift or positive selection.

**Identifying specific instances of regulatory motifs**—Data from just four species (HMRD) was sufficient to create a catalog of known and novel motifs with many conserved instances across the genome[21]. The power to discover such motifs was high, because one can aggregate data across hundreds of motif instances. Not surprisingly, the additional genomes therefore had little effect on the ability to discover new motifs (known motifs showed 99% correlation in genome-wide motif conservation scores, Figure S12 and S13).

In contrast, the 29 mammalian genomes dramatically improved our ability to detect individual motif instances, making it possible to predict specific target sites for 688 regulatory motifs corresponding to 345 transcription factors (Figure S14). We chose to identify motif instances at a false discovery rate (FDR) of 60%, representing a reasonable compromise between specificity and sensitivity given the available discovery power (Text S12), and matching the experimental specificity of Chromatin Immunoprecipitation (ChIP) experiments for identifying biologically-significant targets[40]. Higher levels of stringency could be obtained by sequencing additional species.

We identified 2.7 million conserved instances (Table S6), enabling the construction of a regulatory network linking 375 motifs to predicted targets, with a median of 21 predicted regulators per target gene (25$^{th}$ percentile: 10; 75$^{th}$: 39). The number of target sites (average: 4277; 25$^{th}$ percentile: 1407; 75$^{th}$: 10,782) are comparable to those found in ChIP experiments, and have the advantage that they are detected at nucleotide resolution, enabling us to use them to interpret disease-associated variants for potential regulatory functions. However, some motifs never reached high confidence values, and others did so at very few instances.

The motif-based targets show strong agreement with experimentally-defined binding sites from ChIP experiments (Table S7). For long and distinct motifs, such as CTCF and NRSF, the fraction of instances overlapping experimentally observed binding matches the fraction predicted by the confidence score (e.g. at 80% confidence 70% of NRSF motif instances overlapped bound sites, and at ~50% confidence 40% overlapped), despite potential confounding aspects such as condition-specific binding, overlapping motifs between factors,

or non-specific binding. Moreover, increasing confidence levels showed increasing overlap with experimental binding (Figure S14-16). For example, *YY1* enrichment for bound sites increased from 42-fold to 168-fold by focusing on conserved instances. Lastly, combining motif conservation and experimental binding led to increased enrichment for candidate tissue-specific enhancers, suggesting the two provide complementary information. Within bound regions, the evolutionary signal reveals specific motif instances with high precision (e.g., Figure 2, Figure 4, Figure S17).

**Chromatin signatures—**To suggest potential functions for the ~68% of 'unexplained' constrained elements outside coding regions, UTRs, or proximal promoters, we used chromatin state maps from CD4 T-cells[41] (Figure S18) and nine diverse cell types[42] (Text S13, Figure S19). In T-cells, constrained elements were most enriched for promoter-associated states (up to 5-fold), an insulator state and a specific repressed state (2.2-fold), and numerous enhancer states (1.5-2-fold), together covering 7.1% of the unexplained elements at 2.1-fold enrichment. In the nine cell types, enriched promoter, enhancer and insulator states, cover 36% of unexplained elements at ~1.75-fold enrichment, with locations active in multiple cell types showing even stronger enrichment (Figure S20).

Overall, chromatin states suggest possible functions (at 1.74-fold enrichment) for 37.5% (N=987,985) of unexplained conserved elements (27% of all conserved elements), suggesting meaningful association for at least 16% of unexplained constrained bases. While current experiments only provide nucleosome-scale (~200-bp) resolution, we expect higher-resolution experimental assays that more precisely pinpoint regulatory regions to show further increases in enrichment. The increase observed with additional cell types suggests that new cell types will help elucidate additional elements. Of course, further experimental tests will be required to validate the predicted functional roles.

**Accounting for constrained elements—**Overall, ~30% of constrained elements overlap protein-coding genes, ~27% specific enriched chromatin states, ~1.5% novel RNA structures, and ~3% conserved regulatory motif instances (Text S14). Together, ~60% of constrained elements overlap one of these features, with enrichments ranging from 1.75-fold for chromatin states (compared to unannotated regions) up to 17-fold for protein-coding exons (compared to the whole genome).

**Implications for interpreting disease-associated variants—**In the non-protein-coding genome, SNPs associated with human diseases in genome-wide association studies are 1.37-fold enriched for constrained regions, relative to HapMap SNPs (Text S15, Table S8). This is striking, since only a small proportion of the associated SNPs are likely to be causative while the rest are merely in linkage disequilibrium (LD) with causative variants.

Accordingly, constrained elements should be valuable in focusing the search for causative variants amongst multiple variants in LD. For example, in an intergenic region between *HOXB1* and *HOXB2* associated with tooth development phenotypes[43], the reported SNP (rs6504340) is not conserved, but a linked SNP (rs8073963) sits in a constrained element 7.1 kb away. Moreover, rs8073963 disrupts a deeply-conserved Foxo2 motif instance within a predicted enhancer (Figure 4), making it a candidate mutation for further follow-up. Similar

examples of candidate causal variants are found for diverse phenotypes such as height or multiple sclerosis, and similar analyses could be applied to case-control resequencing data.

## Evolution of constrained elements

We next sought to identify signatures of positive selection that may accompany functional adaptations of different species to diverse environments and new ecosystems. ***Codon-specific selection.*** We used the ratio $d_N/d_S$ of non-synonymous to synonymous codon substitutions as evidence of positive selection (>1) or negative selection (<1). While $d_N/d_S$ is typically calculated for whole genes, the additional mammals sequenced enabled analysis at the codon level – simulations predicted a 250-fold gain in sensitivity compared to HMRD, identifying 53% of positive sites at 5% FDR (Text S16).

Applying this test to 6.05 million codons in 12,871 gene trees, we found evidence of strong purifying selection ($d_N/d_S$<0.5) for 84.2% of codons and positive selection ($d_N/d_S$>1.5) for 2.4% of codons (with 94.1% of sites <1 and 5.9% >1; Table S9). At 5% FDR, we found 15,383 positively-selected sites in 4431 proteins. The genes fall into three classes based on the distribution of selective constraint: 84.8% of genes show uniformly high purifying selection, 8.9% show distributed positive selection across their length, and 6.3% show localized positive selection concentrated in small clusters (Figure 3b, Figure S21, Table S10-11).

Genes with distributed positive selection were enriched in such functional categories as immune response ($p_{Bonf}$ <$10^{-16}$) and taste perception ($p_{Bonf}$ <$10^{-10}$), which are known to evolve rapidly, but also in some unexpected functions such as meiotic chromosome segregation ($p_{Bonf}$<$10^{-23}$) and DNA-dependent regulation of transcription ($p_{Bonf}$<$10^{-19}$, Table S12). Localized positive selection was enriched in core biochemical processes, including microtubule-based movement ($p_{Bonf}$<$10^{-10}$), DNA topological change ($p_{Bonf}$<$10^{-4}$) and telomere maintenance ($p_{Bonf}$<$7×10^{-3}$), suggesting adaptation at important functional sites.

Focusing on 451 unique Pfam protein-domain annotations, we found abundant purifying selection, with 225 domains showing purifying selection for >75% of their sites, and 447 domains showing negative selection for >50% of their sites (Table S13). Domains with substantial fractions of positively-selected sites include CRAL/TRIO involved in retinal binding (2.6%), proteinase-inhibitor-cystatin involved in bone remodeling (2.2%), and secretion-related *Emp24/GOLD/p24* family (1.6%).

**Exaptation of mobile elements—**Mobile elements provide an elegant mechanism for distributing a common sequence across the genome, which can then be retained in locations where it confers advantageous regulatory functions to the host - a process termed *exaptation*. Our data revealed >280,000 mobile element exaptations common to mammalian genomes covering ~7Mb (Text S17), dramatically expanding from ~10,000 previously-recognized cases[44]. Of the ~1.1 million constrained elements that arose during the 90 million years between the divergence from marsupials and the eutherian radiation, we can trace >19% to mobile element exaptations. Often only a small fraction (median ~11%) of each mobile element is constrained, in some cases matching known regulatory motifs. Recent exaptations

are generally found near ancestral regulatory elements, except in gene deserts which are abundant in ancestral elements but show few recent exaptations ($p<10^{-300}$, Figure S22).

**Accelerated evolution in the primate lineage**—Lineage-specific rapid evolution in ancestrally-constrained elements previously revealed human positive selection associated with brain and limb development[45]. Applying this signature to the human and primate lineages, we identified 563 human-accelerated regions (HARs) and 577 primate-accelerated regions (PARs) at FDR<10% (Text S18, Table S14, S15), significantly expanding the 202 previously-known HARs[46]. Fifty-four HARs (9.4%) and 49 PARs (8.5%) overlap enhancer-associated chromatin marks and experimentally validated enhancers (Text S18). Substitution patterns in HARs suggest that GC-biased gene conversion (BGC) is not responsible for the accelerated evolution in the vast majority of these regions (~15% show evidence of BGC).

Genes harboring or neighboring HARs and PARs are enriched for extra-cellular signaling, receptor activity, immunity, axon guidance, cartilage development, and embryonic pattern specification (Figure S23). For example, the *FGF13* locus associated with an X-linked form of mental retardation contains four HARs near the 5′-ends of alternatively-spliced isoforms of *FGF13* expressed in the nervous system, epithelial tissues and tumors, suggesting human-specific changes in isoform regulation (Figure S24).

## Discussion

Comparative analysis of 29 mammalian genomes reveals a high-resolution map of >3.5 million constrained elements that encompass ~4% of the human genome and suggest potential functional classes for ~60% of the constrained bases; the remaining 40% show no overlap and remain uncharacterized. We report previously-undetected exons and overlapping functional elements within protein-coding sequence, new classes of RNA structures, promoter conservation profiles, and predicted targets of transcriptional regulators. We also provide evidence of evolutionary innovation, including codon-specific positive selection, mobile element exaptation and accelerated evolution in the primate and human lineages.

By focusing our comparison on only eutherian mammals, we discover functional elements relevant to this clade, including recent eutherian innovations. This is especially important for discovering regulatory elements, which can be subject to rapid turnover[47]. Indeed, a previous comparison suggest that only 80% of 50-bp non-coding elements are shared with opossum, while the current 12-bp analysis shows ~64% of non-coding elements shared with opossum[48], and only 6% with stickleback fish. Many eutherian elements are thus likely missing from previous maps of vertebrate constraint[17].

Sequencing of additional species should enable discovery of lineage-specific elements within mammalian clades, and provide increased resolution for shared mammalian constraint. We estimate that 100-200 eutherian mammals (15-25 neutral substitutions per site) will enable single-nucleotide resolution. The majority of this branch length is present within the Laurasiatherian and Euarchontoglire branches, which also contain multiple model organisms. These are ideal next targets for sequencing as part of the Genome 10K effort[49],

aiming to sequence 10,000 species from all walks of life. Within the primate clade, a branch length of ~1.5 could be achieved, enabling primate-specific selection studies albeit at lower resolution. Lastly, human-specific selection should be detectable by combining data across genomic regions and by comparing thousands of humans[50].

The constrained elements reported here can be used to prioritize disease-associated variants for subsequent study, providing a powerful lens for elucidating functional elements in the human genome complementary to ongoing large-scale experimental endeavors such as ENCODE and Roadmap Epigenomics. Experimental studies require prior knowledge of the biochemical activity sought and reveal regions active in specific cell types and conditions. Comparative approaches provide an unbiased catalog of shared functional regions independent of biochemical activity or condition, and thus can capture experimentally-intractable or rare activity patterns. With increasing branch length, they can provide information on ancestral and recent selective pressures across mammalian clades and within the human population. Ultimately, the combination of disease genetics, comparative and population genomics and biochemical studies have important implications for understanding human biology, health and disease.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Authors

Kerstin Lindblad-Toh[1,2,$], Manuel Garber[1,*], Or Zuk[1,*], Michael F. Lin[1,3,*], Brian J. Parker[4,*], Stefan Washietl[3,*], Pouya Kheradpour[1,3,*], Jason Ernst[1,3,*], Gregory Jordan[5,*], Evan Mauceli[1,*], Lucas D. Ward[1,3,*], Craig B. Lowe[6,7,8,*], Alisha K. Holloway[9,*], Michele Clamp[1,10,*], Sante Gnerre[1,*], Jessica Alfoldi[1], Kathryn Beal[5], Jean Chang[1], Hiram Clawson[6], James Cuff[11], Federica Di Palma[1], Stephen Fitzgerald[5], Paul Flicek[5], Mitchell Guttman[1], Melissa J Hubisz[12], David B. Jaffe[1], Irwin Jungreis[3], W James Kent[9], Dennis Kostka[9], Marcia Lara[1], Andre L Martins[12], Tim Massingham[5], Ida Moltke[4], Brian J. Raney[6], Matthew D. Rasmussen[3], Jim Robinson[1], Alexander Stark[13], Albert J. Vilella[5], Jiayu Wen[4], Xiaohui Xie[1], Michael C. Zody[1], Broad Institute Sequencing Platform and Whole Genome Assembly Team[†], Kim C. Worley[14], Christie L. Kovar[14], Donna M. Muzny[14], Richard A. Gibbs[14], Baylor College of Medicine Human Genome Sequencing Center[†], Wesley C. Warren[15], Elaine R Mardis[15], George M. Weinstock[14,15], Richard K. Wilson[15], Genome Institute at Washington University[†], Ewan Birney[5], Elliott H. Margulies[16], Javier Herrero[5], Eric D. Green[17], David Haussler[6,8], Adam Siepel[12], Nick Goldman[5], Katherine S. Pollard[9,18], Jakob S. Pedersen[4,19], Eric S. Lander[1,$], and Manolis Kellis[1,3,$]

## Affiliations

[1]Broad Institute of Harvard and Massachusetts Institute of Technology (MIT), 7 Cambridge Center, Cambridge, Massachusetts 02142, USA

[2]Science for Life Laboratory, Department of Medical Biochemistry and Microbiology, Uppsala University, Box 582, SE-751 23 Uppsala, Sweden

[3]MIT Computer Science and Artificial Intelligence Laboratory, 32 Vassar St. Cambridge MA 02139, USA

[4]The Bioinformatics Centre, Department of Biology, University of Copenhagen, DK-2200 Copenhagen, Denmark

[5]EMBL-EBI, Wellcome Trust Genome Campus, CB10 1SD Hinxton, UK

[6]Center for Biomolecular Science and Engineering, University of California, Santa Cruz, CA 95064

[7]Department of Developmental Biology, Stanford University, Stanford, CA 94305

[8]Howard Hughes Medical Institute

[9]Gladstone Institutes, University of California, 1650 Owens Street, San Francisco, CA 94158

[10]BioTeam Inc, 7 Derosier Drive, Middleton, MA

[11]Research Computing, Division of Science, Faculty of Arts and Sciences, Harvard University, Cambridge MA 02138

[12]Dept. of Biological Statistics & Computational Biology, Cornell University, Ithaca, NY 14853

[13]Research Institute of Molecular Pathology (IMP), A-1030 Vienna, Austria

[14]Human Genome Sequencing Center, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77031, USA

[15]Genome Institute at Washington University, Washington University School of Medicine, 1 Childrens Place, Saint Louis, MO 63110, USA

[16]Genome Informatics Section, Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda MD 20892 USA

[17]NISC Comparative Sequencing Program, Genome Technology Branch and NIH Intramural Sequencing Center, National Human Genome Research Institute, National Institutes of Health, Bethesda MD 20892 USA

[18]Institute for Human Genetics, and Division of Biostatistics, University of California, 1650 Owens Street, San Francisco, CA 94158

[19]Department of Molecular Medicine (MOMA), Aarhus University Hospital, Skejby, DK-8200 Aarhus N, Denmark

## Acknowledgements

## Literature cited

1. Lander ES, et al. Initial sequencing and analysis of the human genome. Nature. 2001; 409:860–921. doi:10.1038/35057062. [PubMed: 11237011]

2. Waterston RH, et al. Initial sequencing and comparative analysis of the mouse genome. Nature. 2002; 420:520–562. doi:10.1038/nature01262 nature01262 [pii]. [PubMed: 12466850]

3. Gibbs RA, et al. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature. 2004; 428:493–521. doi:10.1038/nature02426 nature02426 [pii]. [PubMed: 15057822]

4. Lindblad-Toh K, et al. Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature. 2005; 438:803–819. doi:nature04338 [pii] 10.1038/nature04338. [PubMed: 16341006]

5. Altshuler D, Daly MJ, Lander ES. Genetic mapping in human disease. Science. 2008; 322:881–888. doi:322/5903/881 [pii] 10.1126/science.1156409. [PubMed: 18988837]

6. Bejerano G, et al. Ultraconserved elements in the human genome. Science. 2004; 304:1321–1325. doi:10.1126/science.1098119 1098119 [pii]. [PubMed: 15131266]

7. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES. Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature. 2003; 423:241–254. doi:10.1038/nature01644 nature01644 [pii]. [PubMed: 12748633]

8. Stark A, et al. Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. Nature. 2007; 450:219–232. doi:nature06340 [pii] 10.1038/nature06340. [PubMed: 17994088]

9. Cooper GM, Brudno M, Green ED, Batzoglou S, Sidow A. Quantitative estimates of sequence divergence for comparative analyses of mammalian genomes. Genome Res. 2003; 13:813–820. doi: 10.1101/gr.1064503 13/5/813 [pii]. [PubMed: 12727901]

10. Gnerre S, Lander ES, Lindblad-Toh K, Jaffe DB. Assisted assembly: how to improve a de novo genome assembly by using related species. Genome Biol. 2009; 10:R88. doi:gb-2009-10-8-r88 [pii] 10.1186/gb-2009-10-8-r88. [PubMed: 19712469]

11. Hubisz MJ, Lin MF, Kellis M, Siepel A. Error and Error Mitigation in Low-Coverage Genome Assemblies. PLoS ONE. 2011

12. Thomas JW, et al. Comparative analyses of multi-species sequences from targeted genomic regions. Nature. 2003; 424:788–793. doi:10.1038/nature01858 nature01858 [pii]. [PubMed: 12917688]

13. Garber M, et al. Identifying novel constrained elements by exploiting biased substitution patterns. Bioinformatics. 2009; 25:i54–62. doi:btp190 [pii] 10.1093/bioinformatics/btp190. [PubMed: 19478016]

14. Chiaromonte F, et al. The share of human genomic DNA under selection estimated from human-mouse genomic alignments. Cold Spring Harb Symp Quant Biol. 2003; 68:245–254. [PubMed: 15338624]

15. Davydov EV, et al. Identifying a high fraction of the human genome to be under selective constraint using GERP++ PLoS Comput Biol. 2010; 6:e1001025. doi:10.1371/journal.pcbi. 1001025. [PubMed: 21152010]

16. Meader S, Ponting CP, Lunter G. Massive turnover of functional sequence in human and other mammalian genomes. Genome Res. 2010; 20:1335–1343. doi:gr.108795.110 [pii] 10.1101/gr. 108795.110. [PubMed: 20693480]

17. Siepel A, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 2005; 15:1034–1050. doi:gr.3715005 [pii] 10.1101/gr.3715005. [PubMed: 16024819]

18. Drake JA, et al. Conserved noncoding sequences are selectively constrained and not mutation cold spots. Nat Genet. 2006; 38:223–227. doi:ng1710 [pii] 10.1038/ng1710. [PubMed: 16380714]

19. Keinan A, Mullikin JC, Patterson N, Reich D. Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. Nat Genet. 2007; 39:1251–1255. doi:ng2116 [pii] 10.1038/ng2116. [PubMed: 17828266]

20. Clamp M, et al. Distinguishing protein-coding and noncoding genes in the human genome. Proc Natl Acad Sci U S A. 2007; 104:19428–19433. doi:0709013104 [pii] 10.1073/pnas.0709013104. [PubMed: 18040051]

21. Xie X, et al. Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. Nature. 2005; 434:338–345. doi:nature03441 [pii] 10.1038/nature03441. [PubMed: 15735639]

22. Wang ET, et al. Alternative isoform regulation in human tissue transcriptomes. Nature. 2008; 456:470–476. doi:nature07509 [pii] 10.1038/nature07509. [PubMed: 18978772]

23. Siepel A, et al. Targeted discovery of novel human exons by comparative genomics. Genome Res. 2007; 17:1763–1773. doi:gr.7128207 [pii] 10.1101/gr.7128207. [PubMed: 17989246]

24. Pruitt KD, et al. The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. Genome Res. 2009; 19:1316–1323. doi:gr.080531.108 [pii] 10.1101/gr.080531.108. [PubMed: 19498102]

25. Lin MF, Jungreis I, Kellis M. PhyloCSF: a comparative genomics method to distinguish protein-coding and non-coding regions. Nature Precedings. 2010 doi:http://hdl.handle.net/10101/npre.2010.4784.1.

26. Guttman M, et al. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat Biotechnol. 2010; 28:503–510. doi:nbt.1633 [pii] 10.1038/nbt.1633. [PubMed: 20436462]

27. Harrow J, et al. GENCODE: producing a reference annotation for ENCODE. Genome Biol. 2006; 7(Suppl 1):S4, 1–9. doi:gb-2006-7-s1-s4 [pii] 10.1186/gb-2006-7-s1-s4. [PubMed: 16925838]

28. Lin MF, et al. Revisiting the protein-coding gene catalog of Drosophila melanogaster using 12 fly genomes. Genome Res. 2007; 17:1823–1836. doi:gr.6679507 [pii] 10.1101/gr.6679507. [PubMed: 17989253]

29. Jungreis I, Lin MF, Chan CS, Kellis M. Evidence of abundant stop codon readthrough in Drosophila and other metazoa. Genome Research. 2011

30. Washietl S, Hofacker IL, Stadler PF. Fast and reliable prediction of noncoding RNAs. Proc Natl Acad Sci U S A. 2005; 102:2454–2459. doi:0409169102 [pii] 10.1073/pnas.0409169102. [PubMed: 15665081]

31. Lin MF, et al. Locating protein-coding sequences under selection for additional, overlapping functions in 29 mammalian genomes. Genome Research. 2011

32. Tumpel S, Cambronero F, Sims C, Krumlauf R, Wiedemann LM. A regulatory module embedded in the coding region of Hoxa2 controls expression in rhombomere 2. Proc Natl Acad Sci U S A. 2008; 105:20077–20082. doi:0806360105 [pii] 10.1073/pnas.0806360105. [PubMed: 19104046]

33. Lampe X, et al. An ultraconserved Hox-Pbx responsive element resides in the coding sequence of Hoxa2 and is active in rhombomere 4. Nucleic Acids Res. 2008; 36:3214–3225. doi:10.1093/nar/gkn148. [PubMed: 18417536]

34. Pedersen JS, et al. Identification and classification of conserved RNA secondary structures in the human genome. PLoS Comput Biol. 2006; 2:e33. doi:10.1371/journal.pcbi.0020033. [PubMed: 16628248]

35. Lee JT. Lessons from X-chromosome inactivation: long ncRNA as guides and tethers to the epigenome. Genes Dev. 2009; 23:1831–1842. doi:10.1101/gad.1811209. [PubMed: 19684108]

36. Maenner S, et al. 2-D structure of the A region of Xist RNA and its implication for PRC2 association. PLoS Biol. 2007; 8:e1000276. doi:10.1371/journal.pbio.1000276. [PubMed: 20052282]

37. Parker BJ, et al. New families of human regulatory RNA structures identified by comparative analysis of vertebrate genomes. Genome Research. 2011

38. Martinez-Chantar ML, et al. L-methionine availability regulates expression of the methionine adenosyltransferase 2A gene in human hepatocarcinoma cells: role of S-adenosylmethionine. J Biol Chem. 2003; 278:19885–19890. doi:10.1074/jbc.M211554200 M211554200 [pii]. [PubMed: 12660248]

39. Baek D, Davis C, Ewing B, Gordon D, Green P. Characterization and predictive discovery of evolutionarily conserved mammalian alternative promoters. Genome Res. 2007; 17:145–155. doi:gr.5872707 [pii] 10.1101/gr.5872707. [PubMed: 17210929]

40. Kheradpour P, Stark A, Roy S, Kellis M. Reliable prediction of regulator targets using 12 Drosophila genomes. Genome Res. 2007; 17:1919–1931. doi:gr.7090407 [pii] 10.1101/gr. 7090407. [PubMed: 17989251]

41. Ernst J, Kellis M. Discovery and characterization of chromatin states for systematic annotation of the human genome. Nat Biotechnol. 2010; 28:817–825. doi:nbt.1662 [pii] 10.1038/nbt.1662. [PubMed: 20657582]

42. Ernst J, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature. 2011; 473:43–49. doi:nature09906 [pii] 10.1038/nature09906. [PubMed: 21441907]

43. Pillas D, et al. Genome-wide association study reveals multiple loci associated with primary tooth development during infancy. PLoS Genet. 2010; 6:e1000856. doi:10.1371/journal.pgen.1000856. [PubMed: 20195514]

44. Lowe CB, Bejerano G, Haussler D. Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. Proc Natl Acad Sci U S A. 2007; 104:8005–8010. doi:0611223104 [pii] 10.1073/pnas.0611223104. [PubMed: 17463089]

45. Prabhakar S, et al. Human-specific gain of function in a developmental enhancer. Science. 2008; 321:1346–1350. doi:321/5894/1346 [pii] 10.1126/science.1159974. [PubMed: 18772437]

46. Pollard KS, et al. Forces shaping the fastest evolving regions in the human genome. PLoS Genet. 2006; 2:e168. doi:10.1371/journal.pgen.0020168. [PubMed: 17040131]

47. Schmidt D, et al. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. Science. 2010; 328:1036–1040. doi:science.1186176 [pii] 10.1126/science. 1186176. [PubMed: 20378774]

48. Mikkelsen TS, et al. Genome of the marsupial Monodelphis domestica reveals innovation in non-coding sequences. Nature. 2007; 447:167–177. doi:nature05805 [pii] 10.1038/nature05805. [PubMed: 17495919]

49. Genome10KCommunityOfScientists. Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. J Hered. 2009; 100:659–674. doi:esp086 [pii] 10.1093/jhered/esp086. [PubMed: 19892720]

50. A map of human genome variation from population-scale sequencing. Nature. 2010; 467:1061–1073. doi:nature09534 [pii] 10.1038/nature09534. [PubMed: 20981092]
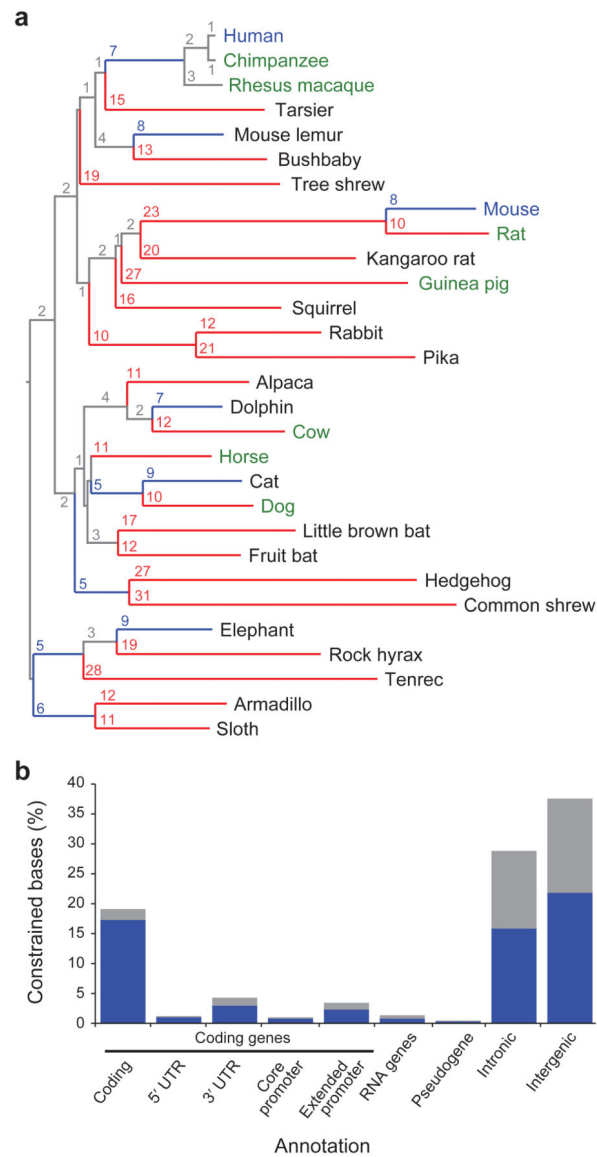
**Figure 1. Phylogeny and constrained elements from the 29 eutherian mammalian genome sequences**

**a,** A phylogenetic tree of all 29 mammals used in this analysis based on the substitution rates in the MultiZ alignments. Organisms with finished genome sequences are indicated in blue, high quality drafts in green and 2X assemblies in black. Substitutions per 100 bp are given for each branch, and branches with 10 substitutions are colored red, while blue indicates < 10 substitutions. **b,** At 10% FDR, 3.6 million constrained elements can be detected encompassing 4.2% of the genome, including a substantial fraction of newly detected bases (blue) compared to the union of the HMRD 50-bp + Siepel vertebrate elements[17] (see Figure S4b for comparison to HMRD elements only). The largest fraction of constraint can be seen in coding exons, introns and intergenic regions. For unique counts, the analysis was performed hierarchically: coding exons, 5′-UTRs, 3′-UTRs, promoters, pseudogenes, non-

coding RNAs, introns, intergenic. The constrained bases are particularly enriched in coding transcripts and their promoters (Supp Fig S4c).
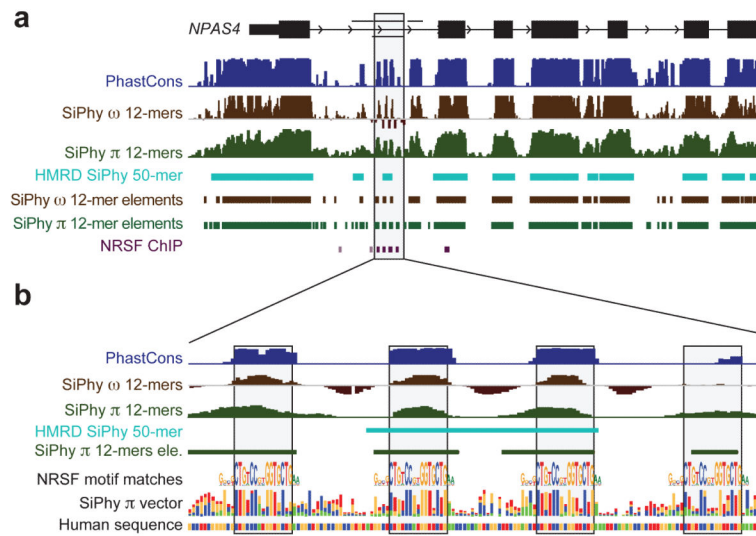
**Figure 2. Identification of four NRSF-binding sites in *NPAS4***

**a.** The neurological gene *NPAS4* has many constrained elements overlapping introns and the upstream intergenic region. The gray shaded box contained only one constrained element using HMRD, while analysis of 29 mammalian sequences reveals four smaller elements. **b,** These four constrained elements in the first intron correspond to binding sites for the NRSF transcription factor, known to regulate neuronal lineages.
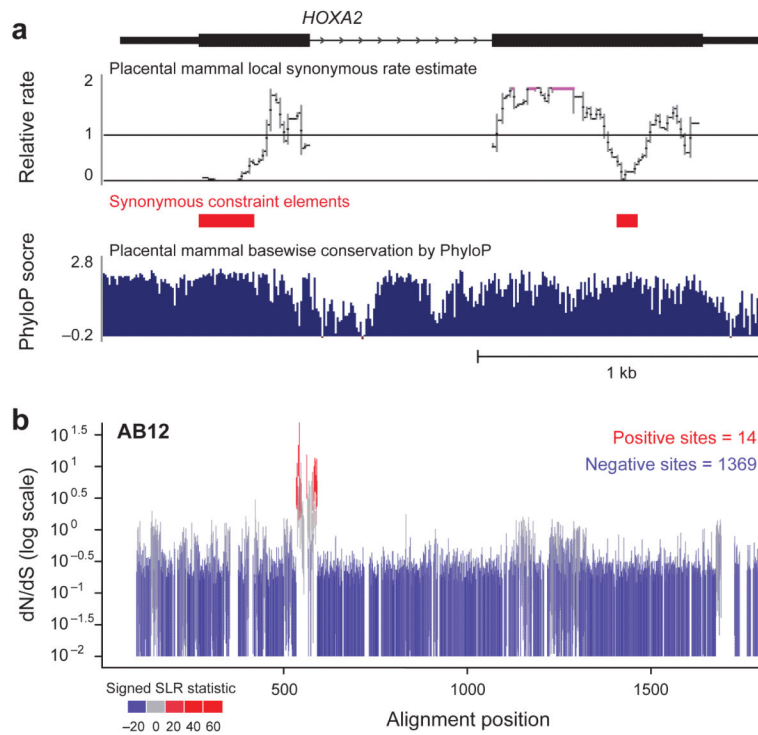
**Figure 3. Examination of evolutionary signatures identifies synonymous constrained elements (SCEs) and evidence of positive selection**

**a,** Two regions within the *HOXA2* open reading frame are identified as Synonymous Constraint Elements (red), corresponding to overlapping functional elements within coding regions. Note that the synonymous rate reductions are not obvious from the base-wise conservation measure (in blue). Both elements have been characterized as enhancers driving *Hoxa2* expression in distinct segments of the developing mouse hindbrain. The element in the first exon encodes Hox-Pbx binding sites and drives expression in rhombomere 4[33], while the element in the second exon contains Sox binding sites and drives expression in rhombomere 2[32]. Synonymous constraint elements are also found in most other *Hox* genes, and up to a quarter of all genes. **b,** While ~85% of genes show only negative (purifying) selection and 9 % of genes show uniform positive selection, the remaining 6% of genes, including *ABI2,* show only localized regions of positively-selected sites. Each vertical bar covers the estimated 95% confidence interval for dN/dS at that site (with values of 0 truncated to 0.01 to accommodate the log scaling), and bars are colored according to a signed version of the SLR statistic for non-neutral evolution: blue for sites under purifying selection, gray for neutral sites, and red for sites under positive selection.
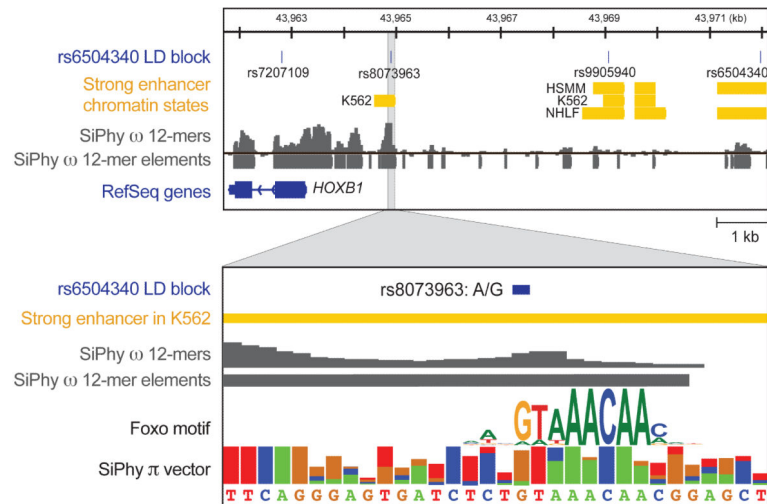
**Figure 4. Utilizing constraint to identify candidate mutations**

Conservation can help us resolve amidst multiple SNPs the ones that disrupt conserved functional elements and are likely to have regulatory roles. In this example, a SNP (rs6504340) associated with tooth development is perfectly linked to a conserved intergenic SNP, rs8073963, 7.1kb away, which disrupts a deeply conserved Forkhead-family motif in a strong enhancer. While the SNPs shown here stem from GWAs or HAPMAP data, the same principle should be applicable also to associated variants detected by resequencing the region of interest.