SUPPLEMENTARY MATERIAL

Relationship between Evolutionary Constraint and Genome Function in 1% of the Human Genome

Elliott H. Margulies[*†], Gregory M. Cooper[*], George Asimenos[*], Daryl J. Thomas[*], Colin N. Dewey[*], Adam Siepel, Ewan Birney, *et al*.

[*] These authors contributed equally
[†] Corresponding Author: elliott@nhgri.nih.gov

Section 1. Generation of 'comparative-grade' BAC-based sequence data for 14 mammalian species.

The sequence dataset summarized in the table below represents the source of all analyses performed in this manuscript and is available at:

http://hgdownload.cse.ucsc.edu/goldenPath/hg17/encode/alignments/SEP-2005/

All BAC-based data was identified, sequenced to comparative-grade (Blakesley et al. 2004) and assembled by the NISC Comparative Sequencing Program (**http://www.nisc.nih.gov**), using previously established methods for comparative mapping (Thomas et al. 2002; Thomas et al. 2003) and shotgun sequencing (Wilson and Mardis 1997). The only exception to this is the cow sequence, which was generated and assembled at the Baylor College of Medicine's Genome Sequencing Center.

Note that this is an on-going project, and that the status of this effort can be found at **http://www.nisc.nih.gov/projects/encode/**. In addition, a bulk download of the most-complete sequence data assemblies can be obtained with the search string 'NISC AND ENCODE [keyword]' at **http://www.ncbi.nlm.nih.gov/**.

All Genome-Wide data were obtained from assemblies provided by the UCSC Genome Browser (Kent et al. 2002) (**http://genome.ucsc.edu**) unless otherwise specified.

Table S1. Summary of the sequence data used.

Species	Bases	Contigs
BAC-Based		
Baboon	34,874,023	45
Colobus Monkey	1,937,088	1
Dusky Titi	2,094,628	1
Galago	33,489,981	44
Hedgehog	3,423,366	3
Marmoset	35,639,002	44
Mouse Lemur	1,573,127	1
Owl Monkey	2,050,914	1
Cow	25,031,782	41
rfbat	25,126,285	44
BAC-Based and Genome-Wide[1]		
Armadillo[2]	25,683,871	92
Elephant[2]	26,974,410	668
Platypus[3]	17,218,288	534
Rabbit[2]	23,811,937	955
Shrew[2]	29,136,919	1,395
Genome-Wide		
Chicken	11,037,529	92
Chimpanzee	28,249,919	81
Dog	26,173,357	52
Fugu	3,085,623	175
Human	29,948,058	44
Macaque	25,533,478	506
Monodelphis	37,346,362	146
Mouse	30,554,370	56
Rat	31,447,713	58
Tenrec[2]	18,455,867	2,499
Tetraodon	4,243,463	185
Xenopus	10,847,995	75
Zebrafish	9,517,444	247
Total	**554,506,799**	**8,085**

[1] These totals represent combined BAC-based and Genome-Wide sequence data.

[2] Sequence obtained in part from the low-redundancy whole-genome shotgun sequencing effort (Margulies et al. 2005) at the Broad Institute (see **http://www.broad.mit.edu/mammals/**).

[3] Sequence obtained from a preliminary phusion assembly (Mullikin and Ning 2003) of traces deposited into the NCBI Trace Repository by the Washington University Genome Sequencing Center.

Section 2. Orthology predictions from whole-genome assemblies using MERCATOR and UCSC alignments.

2.1 Genome-wide assemblies (starting material)

The following whole genome assemblies were used:

Chicken (CGSC_Feb._2004, galGal2)

The February 2004 chicken (*Gallus gallus*) draft assembly was produced by the Genome Sequencing Center at the Washington University School of Medicine in St. Louis.

Chimp (NCBI_Build_1_v1, panTro1)

The 13 Nov. 2003 chimpanzee (*Pan troglodytes*) Arachne assembly -- NCBI Build 1 Version 1 -- was produced by the Chimpanzee Genome Sequencing Consortium.

Dog (Broad_Institute_v._1.0, canFam1)

The July 2004 dog (*Canis familiaris*) whole genome shotgun (WGS) assembly v1.0 was sequenced and assembled by the Broad Institute of MIT/Harvard and Agencourt Bioscience.

Fugu (IMCB/JGI, fr1)

The *Takifugu rubripes* v.3.0 (Aug. 2002) whole genome shotgun assembly was provided by the US DOE Joint Genome Institute (JGI) as part of the International Fugu Genome Consortium, led by JGI and the Singapore Institute of Molecular and Cell Biology (IMCB). Note that Fugu predictions were based on the hg16 / NCBI Human build 34 regions using hg16ToFr1.chain.

Macaque (BCM, rheMac1)

The Jan. 2005 Rhesus monkey or Rhesus macaque (*Macaca mulatta*) preliminary assembly, Mmul_0.1, was obtained from the Baylor College of Medicine Human Genome Sequencing Center (BCM HGSC).

Monodelphis (Broad_Institute, monDom1)

The Oct. 2004 opossum (*Monodelphis domestica*) preliminary assembly was produced by The Broad Institute.

Mouse (NCBI_Build_33, mm6)

The March 2005 mouse (*Mus musculus*) draft genome data was obtained from the Build 34 assembly by NCBI.

Rat (Baylor_HGSC_v3.1, rn3)

The June 2003 rat (*Rattus norvegicus*) genome assembly is based on version 3.1 produced by the Atlas group at Baylor Human Genome Sequencing Center (HGSC) as part of the Rat Genome Sequencing Consortium.

Tetraodon (Genoscope_V7, tetNig1)

The *Tetraodon nigroviridis* V7 assembly (February 2004) was provided by Genoscope, Evry, France. The assembly is the result of a collaboration between Genoscope and the Eli and Edythe L. Broad Institute of MIT and Harvard, MA, USA. The sequence data, which were assembled using the Arachne program, were generated by both institutes. The project was supported by the Consortium National de Recherche en Genomique and the National Human Genome Research Institute (NHGRI).

Xenopus (JGI, xenTro1)

The October 2004 frog (*Xenopus tropicalis*) whole genome shotgun (WGS) assembly version 3.0 was sequenced and assembled by the DOE Joint Genome Institute (JGI).

Zebrafish (Sanger_Zv4, danRer2)

The June 2004 zebrafish (*Danio rerio*) Zv4 assembly was produced by The Wellcome Trust Sanger Institute in collaboration with the Max Planck Institute for Developmental Biology in Tuebingen, Germany, and the Netherlands Institute for Developmental Biology (Hubrecht Laboratory), Utrecht, The Netherlands.

2.2 Blastz/chain/net/liftOver orthology

For non-human species with genome-wide assemblies supported by a browser at UCSC, orthology predictions were generated by the liftOver program. Individual chain and net files (Kent et al. 2003) were prepared from NCBI Human build 35 (UCSC hg17) to each of the other assemblies in the previous section.

The minimum size in the other species was chosen empirically, based on the quality of the assembly and its evolutionary distance from human. LiftOver was modified to allow multiple orthologous region predictions, such as synteny breaks. Parameters for liftOver were as follows:

```
minMatch=0.01 [minimum match ratio]
minSizeT=4000 [minimum human size]
minSizeQ=     [minimum size in the other species]
      {
            1000  for fr1, danRer2, galGal2, tetNig1
            5000  for monDom1 10000 for panTro1, rheMac1
            20000 for mm6, bosTau1, canFam1, rn3
      }
```

```
mergeGap=20000
```

Orthology predictions less than 20000 bp apart were merged together with liftOverMerge. In addition, orthology predictions were made for cow (bosTau1) and mouse (mm6) to assist in finishing efforts at the sequencing centers.

Source Downloads for liftOver and liftOverMerge:
http://hgdownload.cse.ucsc.edu/downloads.html#source_downloads

Online hgLiftOver tool:
http://genome.ucsc.edu/cgi-bin/hgLiftOver

Over.chain files used for the predictions:
http://hgdownload.cse.ucsc.edu/goldenPath/hg17/liftOver/

```
hg17ToBosTau1.over.chain.gz 18-Mar-2005 06:25   84M
hg17ToCanFam1.over.chain.gz 07-Jan-2005 14:40   90M
hg17ToDanRer2.over.chain.gz 03-Mar-2005 17:03  7.2M
hg17ToGalGal2.over.chain.gz 01-Mar-2005 20:22  7.8M
hg17ToMm6.over.chain.gz     07-Dec-2005 12:48   73M
hg17ToMonDom1.over.chain.gz 01-Mar-2005 20:19   36M
hg17ToPanTro1.over.chain.gz 20-Jan-2005 17:53 18M
hg17ToRheMac1.over.chain.gz 17-Mar-2005 18:36   48M
hg17ToRn3.over.chain.gz     01-Mar-2005 20:12   75M
hg17ToTetNig1.over.chain.gz 01-Mar-2005 20:34  3.0M
hg17ToXenTro1.over.chain.gz 05-Jul-2005 16:23  7.4M
```

2.3 Orthology predictions using Mercator

A second set of orthology predictions was generated independently by the Mercator program (Dewey and Pachter, in preparation). For each species, Mercator used as input the following gene annotation tracks as made available from the UCSC Genome Browser: Ensembl, Geneid, Genscan, Known Genes, MGC Gene, N-SCAN, RefSeq, SGP and Twinscan. Gene annotations were processed to produce a non-overlapping set of coding exons in each species. The amino acid sequences coded for by each exon in the resulting sets were compared to each other in an all-vs-all fashion with BLAT (Kent 2002). From the pairwise exon hits, Mercator produced a one-to-one orthology map between the twelve species. Sets of orthologous segments identified by the map that overlapped with ENCODE regions were put into multiple alignments with MAVID (Bray and Pachter 2004). The resulting multiple alignments were then used to map the ENCODE region intervals in human to their orthologous intervals in the other genomes.

2.4 Merging the two sets of orthology predictions

Orthology predictions within 20,000 bases from the chain/net/liftOver process and from Mercator were merged with liftOverMerge to produce the final dataset of non-overlapping sequences.

Section 3. Constrained sequence overlap with alignment-based predictions of coding potential.

We compared the positions of all the constrained sequences that currently have no functional annotation with predictions of coding potential made by exoniphy (Siepel and Haussler 2004a) on the basis of whole-genome multiple sequence alignments of human, mouse, rat, and dog. Specifically, we downloaded all hg17 exoniphy predictions in the ENCODE regions from the UCSC genome browser. We find that ~0.64% of the constrained bases without experimental annotation overlap these predictions. Exoniphy predictions cover about 1% of the genome, but only 0.15% of the portion of the genome outside of known CDSs. Thus, there is a roughly fourfold enrichment for exoniphy predictions in the unannotated constrained sequence. However, this enrichment can at least partially be attributed to a substantially elevated false positive rate for exoniphy in constrained regions (Siepel and Haussler 2004a), implying that 0.64% is an over-estimate. Thus, we find that the vast majority (> 99%) of the unannotated constrained sequences do not code for proteins.

Section 4. Enrichment of constrained sequence in experimentally identified elements.

For each set of experimental annotations (see Box S1), we determined the extent of evolutionary constraint by two different approaches (Fig. S1). The fraction of bases in each experimental annotation that are under constraint is shown in yellow, and the fraction of experimentally annotated regions that contain at least one constrained base is shown in blue. The expected overlap due to random chance is plotted within each bar (error bars correspond to confidence bounds at p=0.002; see Supplementary Information). Most, but not all, annotation classes exhibit significant enrichment levels, especially with respect to the fraction of annotated regions that are at least partially constrained (yellow bars). Note that ancestral repeats (ARs) are significantly devoid of constrained bases, as expected (International Mouse Genome Sequencing Consortium 2002).

Section 5. Trimmed annotations and their overlaps with constrained sequences.

In order to assess the relative specificity of overlap between constrained sequences and experimentally-identified annotations, we determined the extent to which annotations could be "enriched" for constrained sequences if they were artificially lengthened or trimmed. The details of this process are depicted in Figure 8A in the main text. Figure S2 depicts this analysis for many of the experimental annotations provided to us from the other ENCODE analysis groups (see Box S1).

Section 6. Longest and most constrained sequences in the ENCODE targets.

Table S2.

				Functional	PhastOdds	Score *
Human Chr	hg17 start	hg17 stop	Length	Annotation	Score (per-bp)	Length
chr7	113882729	113884236	1508	NONE	4.53	6828
chr7	113926150	113927410	1261	UTR	4.53	5710
chrX	122960433	122961911	1479	UTR	3.81	5642
chr7	113894705	113895999	1295	OTHER	3.97	5139
chr14	98707669	98708902	1234	UTR	3.85	4755
chr7	113650922	113651860	939	NONE	4.87	4570
chr7	113924154	113925168	1015	UTR	4.37	4433
chr7	113975626	113976563	938	NONE	4.65	4358
chr18	23785122	23786324	1203	CDS	3.62	4352
chr7	126607319	126608721	1403	CDS	3.05	4275
chr7	26997066	26998196	1131	CDS	3.69	4172
chr7	113889163	113890131	969	NONE	4.07	3943
chr7	113648944	113649962	1019	UTR	3.73	3799
chr7	113893716	113894702	987	OTHER	3.84	3787
chr7	113659412	113660386	975	NONE	3.88	3781
chr20	33704701	33706135	1435	CDS	2.53	3635
chr7	113922379	113923162	784	OTHER	4.44	3484
chr7	113856110	113856853	744	OTHER	4.52	3362
chr7	113736377	113737119	743	OTHER	4.40	3270
chr14	98711215	98712309	1095	CDS	2.95	3234
chr7	113921131	113921913	783	NONE	4.06	3176
chr7	113989692	113990409	718	NONE	4.32	3101
chr6	108848793	108849721	929	NONE	3.28	3047
chr7	113664996	113665742	747	NONE	4.06	3030
chr5	56226952	56227749	798	UTR	3.60	2874
chr6	108608040	108608861	822	CDS	3.48	2863
chr7	113853794	113854651	858	NONE	3.30	2835
chr7	113648124	113648872	749	OTHER	3.76	2815
chr15	41600955	41601862	908	CDS	3.07	2790
chr7	90539649	90540602	954	CDS	2.81	2676
chr7	26955875	26956726	852	CDS	3.10	2645
chr7	26915039	26915754	716	CDS	3.67	2631
chr8	119191591	119192704	1114	CDS	2.27	2524
chr7	113864398	113865006	609	OTHER	4.13	2516

chr5	56251156	56251850	695	UTR	3.49	2426
chr7	113646505	113647280	776	OTHER	3.11	2411
chr7	113891069	113891693	625	OTHER	3.82	2390
chr7	114520695	114521470	776	NONE	3.06	2377
chr7	26913513	26914457	945	CDS	2.51	2375
chr7	113925174	113925750	577	UTR	4.04	2330
chr21	33870431	33871295	865	CDS	2.68	2322
chr22	30676669	30677335	667	CDS	3.45	2299
chr7	114507091	114507643	553	NONE	4.15	2294
chr6	108850745	108851373	629	NONE	3.57	2246
chr7	126769630	126770348	719	NONE	3.09	2225
chr14	98710239	98710975	737	CDS	3.01	2220
chr7	113652081	113652550	470	OTHER	4.71	2215
chr7	113857046	113857596	551	NONE	3.99	2198
chr7	113656855	113657380	526	NONE	4.07	2140
chr7	113891725	113892350	626	CDS	3.41	2134

Section 7. Detection of lineage-specific constrained sequences.

Lineage-specific constrained sequences (LCSs) were identified with the program called DLESS (Siepel et al. 2006) (Detection of LinEage-Specific Selection). DLESS predicts three types of sequences: ones constrained in all species, ones that have been "gained" (i.e., that have come under purifying selection) on some branch of the phylogeny, and ones that have been "lost" (i.e., released from selection) on some branch of the phylogeny. The program is based on a phylogenetic hidden Markov model with states for neutrally-evolving sequences, fully constrained sequences, gains on each branch of the tree, and losses on each branch of the tree. DLESS takes as input a phylogeny with branch lengths, a model of neutral substitution, and a multiple alignment, and it outputs a General Feature Format (GFF) file with one line per predicted element, indicating its coordinates in a designated reference sequence, its type ("conserved", "gain", or "loss"), the branch in question (if "gain" or "loss"), and a log-odds score. DLESS uses indels as well as substitutions in identifying sequences under selection. Details are given in Siepel et al. (2006) (Siepel et al. 2006).

DLESS was run on the TBA multiple alignments for all ENCODE regions. Only the 17 mammals that were well represented across all regions were included in the analysis (human, chimp, baboon, macaque, marmoset, galago, rat, mouse, rabbit, cow, dog, rfbat, armadillo, elephant, tenrec, monodelphis, and platypus). The tree topology from Figure 1 was used, and the branch lengths and substitution model were estimated from fourfold degenerate sites in coding regions, using the REV model. The parameters that define the program's indel model and HMM transition probabilities were estimated by maximum likelihood from the entire ENCODE data set. The following values were estimated: --expected-length 20 --target-coverage 0.055 --phi 0.261 --indel-model 0.0334,0.0533,0.0529,0.0117,0.0206,0.0654.

Using the program phyloP (Siepel et al. 2005), each DLESS prediction was assigned a p-value indicating the probability that the observed number of substitutions or fewer would occur under the neutral model. Since the number of substitutions is not actually observed, the expected value of the posterior distribution was used in its place. In the case of lineage-specific elements, the p-values reported by phyloP indicate the probability of the "observed" (posterior expected) number of substitutions or fewer in the subtree beneath the branch in question given the substitutions in the rest of the tree. In this case, they can be interpreted as measures of "acceleration" or "deceleration" of substitution rate. See Siepel et al. (2006) for details. Only predictions with p-values of less than 0.05 were retained. About 10% of predictions were discarded.

After filtering by p-value, a total of 22728 elements remained, of which 75.7% are fully constrained, 8.7% are gains, and 15.7% are losses. The predicted elements span 5.0% of human bases; 59.4% of the bases in these predictions are in fully constrained elements, 14.0% are in gains, and 26.6% are in losses. The fully constrained elements are somewhat shorter (median length 24bp) than the gains (80bp) and losses (85bp), primarily because weaker power for detecting gains and losses produces an ascertainment bias for long elements. Most of the fully constrained predictions overlap sequences predicted as constrained by our other methods so we will not comment further on this set. In addition, the losses predicted to occur on branches leading to a single external node of the tree (e.g., the branch leading to rat) seem to be enriched for sequencing, assembly, and alignment errors---as might be expected, errors in the sequences tend show up as predictions of accelerated evolution in a single species. Therefore, we excluded these predictions from subsequent analyses.

The remaining 3610 lineage-specific predictions consist of 1972 gains and 1638 losses, covering 0.7% and 0.5% of bases in the human sequence, respectively. They include gains and losses on most internal branches of the phylogeny, but favor longer branches (on which more events are expected to have occurred) and branches near the root of the tree (where we have more power). The most common type of LCS is a gain on the branch to the eutherian (placental) mammals, which may reflect extensive gain-of-function evolution on this branch, but probably also reflects ascertainment biases due to alignability and detection power. As noted in the text, while our methods generally have fairly weak power to detect primate-specific elements, we did find 94 such elements. Two examples are shown in Figures S3 and S4.

The predicted LCSs overlap heavily with our separate predictions of fully constrained sequences, but 46% of bases within the LCSs (roughly 50% in gains and 40% in losses) fall outside of the predictions of fully constrained sequences. Compared with fully constrained sequences, these "novel" LCSs are depleted for coding regions (which tend to be fully constrained), and are significantly depleted for 3'UTRs but slightly enriched for 5'UTRs. Apparently, turnover of constrained sequences is more likely in 5'UTRs than in 3'UTRs or coding regions. In addition, 5'UTRs are strongly enriched for gains and depleted for losses compared to the set of LCSs as a whole. The novel LCSs overlap with the other experimental annotations (Box S1) at about the same rate as observed with fully constrained sequences. Thus, LCSs do account for some experimentally annotated bases not included in fully constrained elements, but do not dramatically change the fractions of experimentally annotated bases that show evidence of constraint. For example, as noted in the text, about 7% of TUFs and TARs/Transfrags fall in fully constrained sequences, and an additional 1% fall in novel LCSs. Conversely, while experimental annotations overlap a substantial fraction (41%) of bases in novel LCSs, the majority of bases in these sequences remain unannotated.

The DLESS predictions (after filtering with phyloP) are displayed in the UCSC Genome Browser (see "DLESS" track for Human May 2004 assembly). The p-values computed by phyloP are shown on the details page for each prediction.

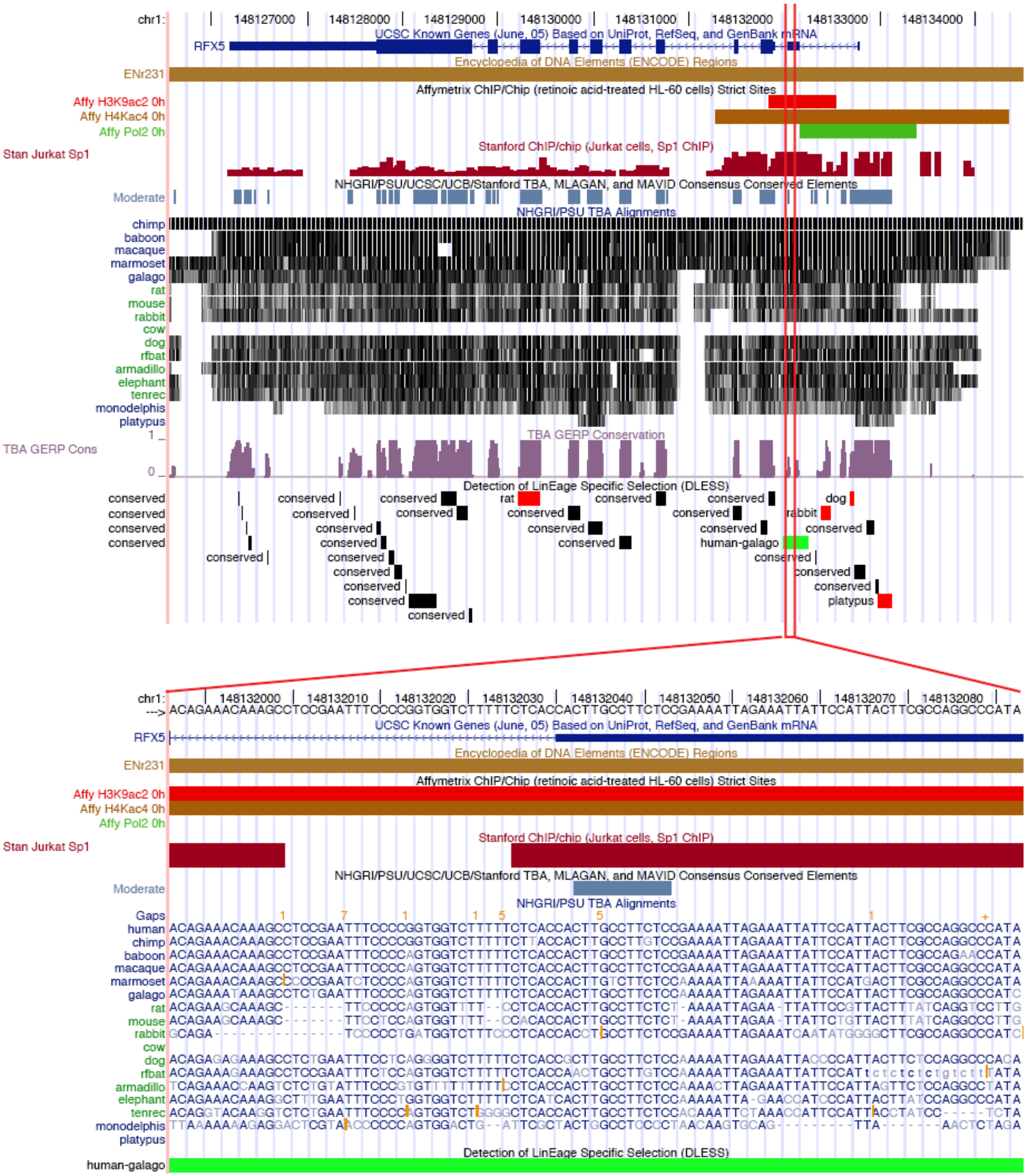
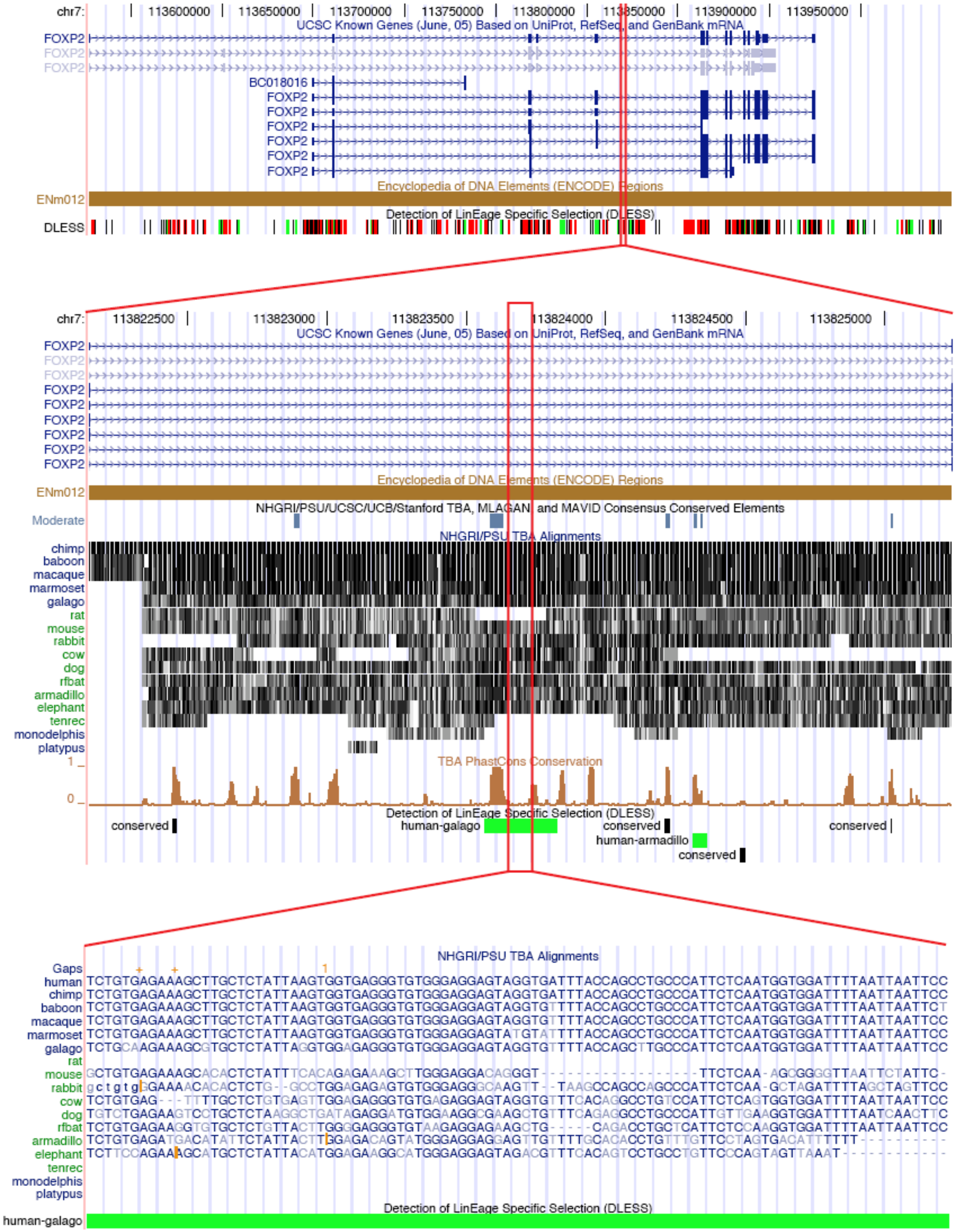

chr1:
148127000
148128000
148129000
148130000
148131000
148132000
148133000
148134000
UCSC Known Genes (June, 05) Based on UniProt, RefSeq, and GenBank mRNA
RFX5
Encyclopedia of DNA Elements (ENCODE) Regions
ENr231
Affymetrix ChIP/Chip (retinoic acid-treated HL-60 cells) Strict Sites
Affy H3K9ac2 0h
Affy H4Kac4 0h
Affy Pol2 0h
Stanford ChIP/chip (Jurkat cells, Sp1 ChIP)
Stan Jurkat Sp1
NHGRI/PSU/UCSC/UCB/Stanford TBA, MLAGAN, and MAVID Consensus Conserved Elements
Moderate
NHGRI/PSU TBA Alignments
chimp
baboon
macaque
marmoset
galago
rat
mouse
rabbit
cow
dog
rfbat
armadillo
elephant
tenrec
monodelphis
platypus
TBA GERP Conservation
TBA GERP Cons
1
0
Detection of LinEage Specific Selection (DLESS)
conserved
rat
dog
rabbit
human-galago
platypus
chr1:
148132000
148132010
148132020
148132030
148132040
148132050
148132060
148132070
148132080
---> ACAGAAACAAAGCCTCCGAATTTCCCCGGTGGTCTTTTTCTCACCACTTGCCTTCTCCGAAAATTAGAAATTATTCCATTACTTCGCCAGGCCCATA
Gaps
human
Detection of LinEage Specific Selection (DLESS)
human-galago

Figure S3. Primate-specific constrained sequence overlapping the 5' UTR of *RFX5*.

UCSC Genome Browser view of RFX5, a gene on human chromosome 1 that encodes a a subunit of the RFX nuclear protein complex and is implicated in bare lymphocyte syndrome type II. The Known Genes, Consensus Conserved Elements, NHGRI/PSU TBA Alignments, TBA GERP Conservation, and DLESS tracks, as well as selected ChIP/chip tracks are shown. A constrained element predicted to have been gained on the branch leading to the most recent common ancestor of human and galago (i.e., a primate-specific constrained sequence) is highlighted (see green element in DLESS track). This sequence overlaps a 5'UTR exon of RFX5 and also overlaps several regions identified in ChIP/chip experiments. It could potentially contain primate-specific regulatory elements, but the overlap with the low resolution ChIP/chip annotations could also be coincidental. At bottom a section of the predicted sequence is shown at the base level. The entire predicted sequence, which is 258bp long, is significantly more conserved in primates than would be expected under a model of neutral substitution (p=7e-13) but is not significantly conserved in the other species (p=0.06). The conditional p-value for the number of substitutions in the primate subtree given the number in the rest of the tree is p=3e-10. The DLESS prediction includes a short constrained element (see Consensus Conserved Elements track) but generally does not appear to be constrained when all species are considered together (see TBA GERP Conservation plot).



```
Gaps               +    +              1
human       TCTGTGAGAAAGCTTGCTCTATTAAGTGGTGAGGGTGTGGGAGGAGTAGGTGATTTACCAGCCTGCCCATTCTCAATGGTGGATTTTAATTAATTCC
chimp       TCTGTGAGAAAGCTTGCTCTATTAAGTGGTGAGGGTGTGGGAGGAGTAGGTGATTTACCAGCCTGCCCATTCTCAATGGTGGATTTTAATTAATTCC
baboon      TCTGTGAGAAAGCTTGCTCTATTAAGTGGTGAGGGTGTGGGAGGAGTAGGTGTTTTACCAGCCTGCCCATTCTCAATGGTGGATTTTAATTAATTCT
macaque     TCTGTGAGAAAGCTTGCTCTATTAAGTGGTGAGGGTGTGGGAGGAGTAGGTGTTTTACCAGCCTGCCCATTCTCAATGGTGGATTTTAATTAATTCC
marmoset    TCTGTGAGAAAGCTTGCTCTATTAAGTGGTGAGGGTGTGGGAGGAGTATGTATTTTACCAGCCTGCCCATTCTCAATGGTGGATTTTAATTAATTCC
galago      TCTGCAAGAAAGCGTGCTCTATTAGGTGGAGAGGGTGTGGGAGGAGTAGGTGTTTTACCAGCTTGCCCATTCTCAATGGTGGATTTTAATTAATTCC
rat
mouse       GCTGTGAGAAAGCACACTCTATTTCACAGAGAAAGCTTGGGAGGACAGGGT------------------TTCTCAA-AGCGGGGTTAATTCTATTC-
rabbit      gctgtgGGAAAACACACTCTG--GCCTGGAGAGAGTGTGGGAGGGCAAGTT--TAAGCCAGCCAGCCCATTCTCAA-GCTAGATTTTAGCTAGTTCC
cow         TCTGTGAG---TTTTGCTCTGTGAGTTGGAGAGGGTGTGAGAGGAGTAGGTGTTTCACAGGCCTGTCCATTCTCAGTGGTGGATTTTAATTAATTCC
dog         TGTCTGAGAAGTCCTGCTCTAAGGCTGATAGAGGATGTGGAAGGCGAAGCTGTTTCAGAGGCCTGCCCATTGTTGAAGGTGGATTTTAATCAACTTC
rfbat       TCTGTGAGAAGGTGTGCTCTGTTACTTGGGGAGGGTGTAAGAGGAGAAGCTG-----CAGACCTGCTCATTCTCCAAGGTGGATTTTAATTAATTCC
armadillo   TCTGTGAGATGACATATTCTATTACTTGGAGACAGTATGGGAGGAGGAGTTGTTTTGCACACCTGTTTGTTCCTAGTGACATTTTTT----------
elephant    TCTTCCAGAAAGCATGCTCTATTACATGGAGAAGGCATGGGAGGAGTAGACGTTTCACAGTCCTGCCTGTTCCCAGTAGTTAAAT-----------
tenrec
monodelphis
platypus
```

Figure S4. Primate-specific constrained sequence in an intron of *FOXP2*.

UCSC Genome Browser view of the FOXP2 gene, highlighting a predicted primate-specific constrained element in an intron of that gene. As with the previous example, this sequence contains a short constrained element but is generally not well conserved across all mammals (see TBA PhastCons Conservation and Consensus Conserved Elements tracks). The entire 258bp sequence is significantly conserved in primates (p=3.3e-10) but not in the other mammals (p=0.73), and has a conditional p-value of p=1.5e-12. This sequence does not overlap any experimental annotations.

Supplemental Methods

Estimating rates of evolution at neutral sites

We first generated a tree on the basis of aligned four-fold degenerate sites within coding exons (taken from the longest transcript if there was more than one at a given locus). For any given non-human sequence, sites that fell within gaps or that were no longer synonymous (because of changes in the first two bases) were treated as missing data. Substitution rates were estimated by maximum likelihood with the PHAST package (Siepel and Haussler 2004b). A generally accepted tree topology for the analyzed species was used (Murphy et al. 2001). The most general reversible substitution model (REV) was used, and no molecular clock was assumed.

Substitution rates in ancestral repeat alignments were estimated using the XRATE program (Holmes and Rubin 2002; Klosterman et al. 2006), which uses a version of the Expectation-Maximization algorithm to obtain fast estimates of the maximum likelihood parameterizations for continuous-time Markov chains (substitution models) on phylogenetic trees. A phylogenetic tree topology for all species under consideration was estimated from fourfold degenerate sites in coding sequence alignments, calibrated to one expected substitution per site per unit of time (see above). XRATE was then applied independently to each alignment of ancestral repeat elements, in order to fit the following models: "REV", the general reversible model of point substitution. Use of a general irreversible model and a dinucleotide model (largely to account for CpG effects) gave similar results (data not shown). Rates were calibrated to the reference phylogeny, i.e. a rate of 2.0 would indicate that substitutions were occurring twice as fast as at the fourfold degenerate sites. All of these evolutionary models are available as default presets in XRATE and/or its companion program XGRAM (Klosterman et al. submitted). As a control, XRATE was used to estimate REV matrices for the fourfold degenerate sites of coding regions, using the reference phylogeny (which was computed from these same sites). If the methods used (PHAST and XRATE) are perfectly consistent, one would expect this experiment to yield a substitution rate of 1.0. The actual rates as measured by XRATE were 0.98 (REV), indicating that the methods are consistent within a small margin of error.

Box S1: Experimental annotations by the ENCODE Consortium.

ENCODE Abbreviation	Description
CDS	**C**o**D**ing **S**equence: Well-characterized transcribed regions with an annotated protein-coding Open Reading Frame (ORF) (ENCODE Project Consortium 2007).
5' UTR	**5' U**n**T**ranslated **R**egion: Portions of CDS-containing transcripts prior to the start codon (ENCODE Project Consortium 2007). For the analyses reported here, 5’UTRs overlapping alternatively-transcribed CDS annotations were removed from this dataset.
3' UTR	**3' U**n**T**ranslated **R**egion: Portions of CDS-containing transcripts after the stop codon (ENCODE Project Consortium 2007). For the analyses reported here, 3’UTRs overlapping alternatively-transcribed 5’UTRs were removed from this dataset.
TUFs	**T**ranscripts of **U**nknown **F**unction: Well-characterized transcribed regions with no annotated protein-coding ORF (ENCODE Project Consortium 2007). For the analyses reported here, portions TUFs overlapping any of the above CDS, 5’ or 3’ UTR annotations were removed from this dataset.
TransFrags	**T**ranscriptionally **A**ctive **R**egions/**Trans**cribed **Frag**ments as determined by analyses of RNA hybridizations to multiple microarray platforms (ENCODE Project Consortium 2007). For the analyses reported here, portions of TransFrags overlapping any of the above CDS, 5’ or 3’ UTR annotations were removed from this dataset.
RACEfrags	Transcribed regions identified from 5' Rapid Amplification of cDNA Ends (RACEs) using primers anchored in well-characterized transcripts and followed by hybridization to high-density resolution tiling arrays (ENCODE Project Consortium 2007).
Pseudoexons	Regions representing exons from pseudogenes (ENCODE Project Consortium 2007).
DHS	**D**NAse **H**ypersensitive **S**ites: Regions of open chromatin detected by through quantitative chromatin profiling and novel microarray-based methods (Crawford et al. 2006; Dorschner et al. 2004; Sabo 2006). In addition to the complete set of DHSs (all), we also analyzed a set only overlapping non-repetitive sequence (no repeat).
FAIRE-sites	**F**ormaldehyde **A**ssisted **I**solation of **R**egulatory **E**lements: a procedure used to isolate chromatin that is resistant to the formation of protein-DNA crosslinks (Giresi et al. 2006; Nagy et al. 2003).
Seq. Specific Factors	Regions of DNA determined to be bound by sequence-specific transcription factors through **Ch**romatin **I**mmuno**P**recipitation

followed by microarray **chip** hybridization (so-called “ChIP-Chip”) analyses (ENCODE Project Consortium 2007)..

General Factors Regions of DNA determined to be bound by proteins with little sequence specificity (i.e., histones) using ChIP-Chip analyses (ENCODE Project Consortium 2007)..

All Motifs Computationally-identified short sequence motifs found to be over-represented in the Sequence-Specific Factors dataset (ENCODE Project Consortium 2007)..

All TR Data The union of all Seq. Specific Factors, General Factors, and All Motifs.

TSSs **T**ranscriptional **S**tart **S**ites (ENCODE Project Consortium 2007).

ARs **A**ncestral **R**epeats: ancient relics of transposable elements that inserted into the ancestral genome prior to the mammalian radiation (International Mouse Genome Sequencing Consortium 2002) (see Methods).

ALL Datasets The union of all the above datasets.

ALL non-exonic The union of all the above datasets, excluding CDSs, UTRs and the features that overlap them (ENCODE Project Consortium 2007).

RepSeg Regions undergoing replication at different times in the cell cycle, noted by Early, Mid, Late, or PanS (ENCODE Project Consortium 2007).

Predicted Origins Predicted origins of replication (ENCODE Project Consortium 2007).

Supplemental References

Blakesley RW, Hansen NF, Mullikin JC, Thomas PJ, McDowell JC, Maskeri B, Young AC, Benjamin B, Brooks SY, Coleman BI et al. 2004. An intermediate grade of finished genomic sequence suitable for comparative analyses. *Genome Res* **14:** 2235-2244.

Bray N and Pachter L. 2004. MAVID: constrained ancestral alignment of multiple sequences. *Genome Res* **14:** 693--699.

Crawford GE, Davis S, Scacheri PC, Renaud G, Halawi MJ, Erdos MR, Green R, Meltzer PS, Wolfsberg TG, and Collins FS. 2006. DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays. *Nat Methods* **3:** 503-509.

Dorschner MO, Hawrylycz M, Humbert R, Wallace JC, Shafer A, Kawamoto J, Mack J, Hall R, Goldy J, Sabo PJ et al. 2004. High-throughput localization of functional elements by quantitative chromatin profiling. *Nat Methods* **1:** 219--225.

ENCODE Project Consortium. 2007. The ENCODE pilot project: Functional annotation of 1% of the human genome. **Submitted**.

Giresi PG, Kim J, McDaniell RM, R. IV, and Lieb JD. 2006. FAIRE (Formaldehyde Assisted Isolation of Regulatory Elements) isolates nucleosome-depleted DNA from human chromatin. *Genome Res* **Submitted**.

Holmes I and Rubin GM. 2002. An expectation maximization algorithm for training hidden substitution models. *J Mol Biol* **317:** 753-764.

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

Kent WJ. 2002. BLAT--the BLAST-like alignment tool. *Genome Res* **12:** 656--664.

Kent WJ, Baertsch R, Hinrichs A, Miller W, and Haussler D. 2003. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc Natl Acad Sci U S A* **100:** 11484--11489.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, and Haussler D. 2002. The human genome browser at UCSC. *Genome Res* **12:** 996--1006.

Klosterman PS, Uzilov AV, Bendana YR, Bradley RK, Chao S, Kosiol C, Goldman N, and Holmes I. 2006. XRate: a fast prototyping, training and annotation tool for phylo-grammars. *BMC Bioinformatics* **7:** 428.

Margulies EH, Vinson JP, NISC Comparative Sequencing Program, Miller W, Jaffe DB, Lindblad-Toh K, Chang JL, Green ED, Lander ES, Mullikin JC et al. 2005. An initial strategy for the systematic identification of functional elements in the human genome by low-redundancy comparative sequencing. *Proc Natl Acad Sci U S A* **102:** 4795--4800.

Mullikin JC and Ning Z. 2003. The phusion assembler. *Genome Res* **13:** 81--90.

Murphy WJ, Eizirik E, O'Brien SJ, Madsen O, Scally M, Douady CJ, Teeling E, Ryder OA, Stanhope MJ, de Jong WW et al. 2001. Resolution of the early placental mammal radiation using Bayesian phylogenetics. *Science* **294:** 2348-2351.

Nagy PL, Cleary ML, Brown PO, and Lieb JD. 2003. Genomewide demarcation of RNA polymerase II transcription units revealed by physical fractionation of chromatin. *Proc Natl Acad Sci U S A* **100:** 6364--6369.

Sabo PJ. 2006. Genome-scale mapping of DNaseI sensitivity in vivo using tiling DNA microarrays. *Nat Methods* **Submitted**.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15:** 1034--1050.

Siepel A and Haussler D. 2004a. Computational identification of evolutionarily conserved exons. *Proc. 8th Annual Int'l Conf. on Research in Computational Biology* **RECOMB'04:** 177-186.

Siepel A and Haussler D. 2004b. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Mol Biol Evol* **21:** 468-488.

Siepel A, Pollard K, and Haussler D. 2006. New methods for detecting lineage-specific selection. In *Proceedings of the 10th Annual International Conference on Research in Computational Biology*.

Thomas JW, Prasad AB, Summers TJ, Lee-Lin S-Q, Maduro VVB, Idol JR, Ryan JF, Thomas PJ, McDowell JC, and Green ED. 2002. Parallel construction of orthologous sequence-ready clone contig maps in multiple species. *Genome Res* **12:** 1277-1285.

Thomas JW, Touchman JW, Blakesley RW, Bouffard GG, Beckstrom-Sternberg SM, Margulies EH, Blanchette M, Siepel AC, Thomas PJ, McDowell JC et al. 2003. Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* **424:** 788-793.

Wilson RK and Mardis ER. 1997. Shotgun Sequencing. In *Genome Analysis: A Laboratory Manual, Vol. 1 Analyzing DNA* (eds. B. Birren E.D. Green S. Klapholz R.M. Meyers, and J. Roskams), pp. 397-454. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.