Genome-wide inference of natural selection on human transcription factor binding sites

L. Arbiza, I. Gronau, B.A. Aksoy, M.J. Hubisz, B. Gulko, A. Keinan, A. Siepel

SUPPLEMENTARY INFORMATION

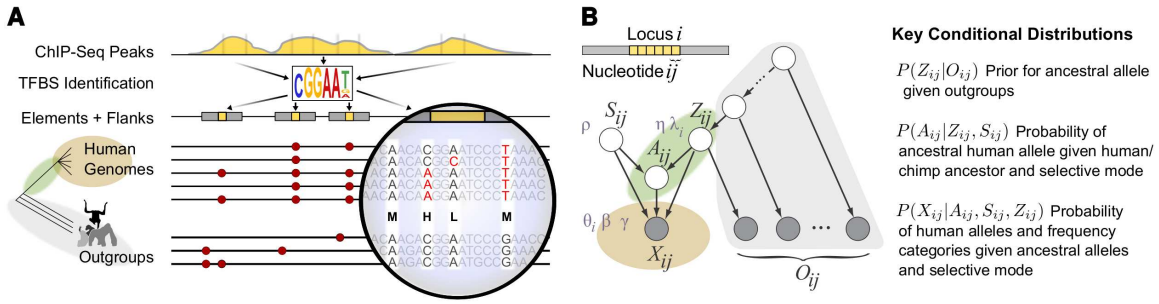
Supplementary Figures

Supplementary Tables

Supplementary Note

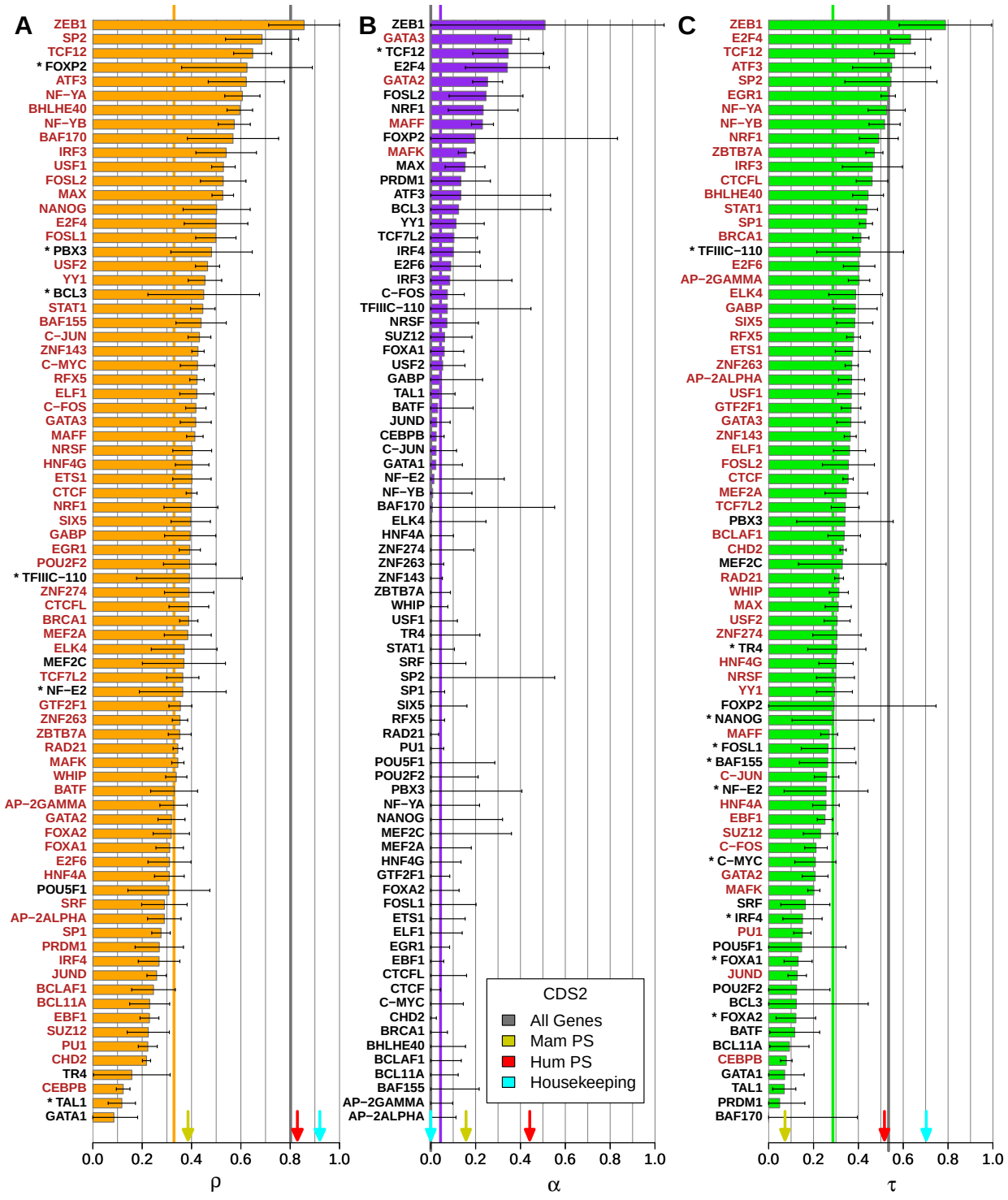
Nature Genetics: doi:10.1038/ng.2658

Nature Genetics: doi:10.1038/ng.2658



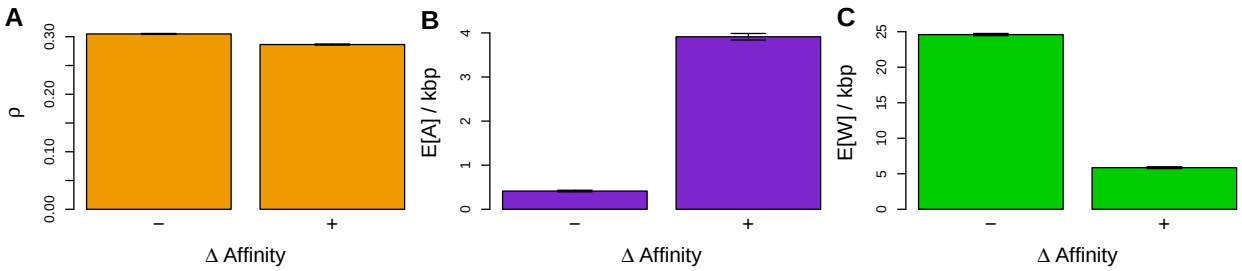
Supplementary Figure 1: Illustration of the Inference of Natural Selection from Interspersed Genomically coHerent elemenTs (INSIGHT) method. (**A**) Transcription factor binding sites (TFBSs) are identified genome-wide by a pipeline that combines de novo motif finding and binding site prediction at ChIP-seq peaks from ENCODE (Dunham et al., 2012). Each predicted element (yellow) is analyzed together with carefully filtered flanking neutral sites (gray) in a 20 kbp block containing the TFBS. Publicly available individual human genome sequences from Complete Genomics (Drmanac et al., 2010) ($n = 54$) are summarized at each position as displaying low- (L) or high- (H) frequency minor alleles, or being monomorphic (M). We experimented with various thresholds for low-frequency alleles but used 15% for most analyses. The orthologous nucleotides in nonhuman primate outgroup sequences (chimpanzee, orangutan, and rhesus macaque) are also summarized. Selection parameters for a collection of TFBSs are inferred by contrasting patterns of polymorphism and divergence at TFBSs with those at flanking neutral sites. (**B**) The probabilistic graphical model used by INSIGHT. The model assumes conditional independence of loci (TFBSs and their flanks) given a set of global parameters, and conditional independence of polymorphism and divergence data at each nucleotide site given locus-specific parameters (see Table 1, Online Methods). It allows for uncertainty in the alleles ancestral to humans and chimpanzees (Z_{ij}) and the human individuals (A_{ij}), and in the selective mode at each site (S_{ij}; fixed at neutral for flanking sites). Observed variables include the outgroup nucleotides (O_{ij}) and the human alleles and frequency categories (X_{ij}). The model is fitted to the data by maximum likelihood using an expectation maximization algorithm.

Nature Genetics: doi:10.1038/ng.2658



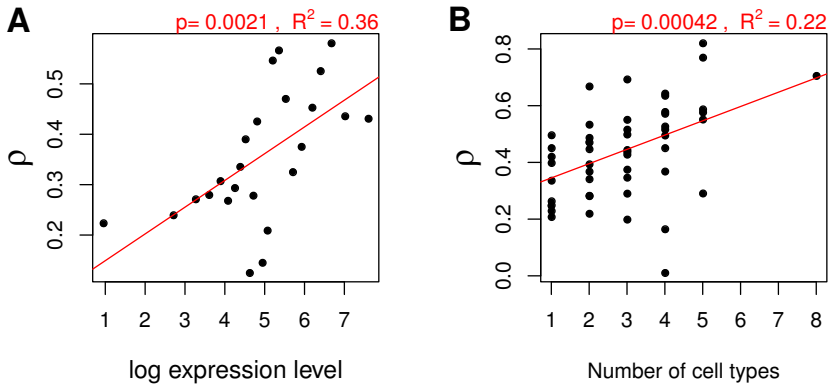
Supplementary Figure 2: Estimates of (**A**) ρ, (**B**) α, and (**C**) τ for the binding sites of all transcription factors (TFs). Compare with Fig. 2.

Nature Genetics: doi:10.1038/ng.2658



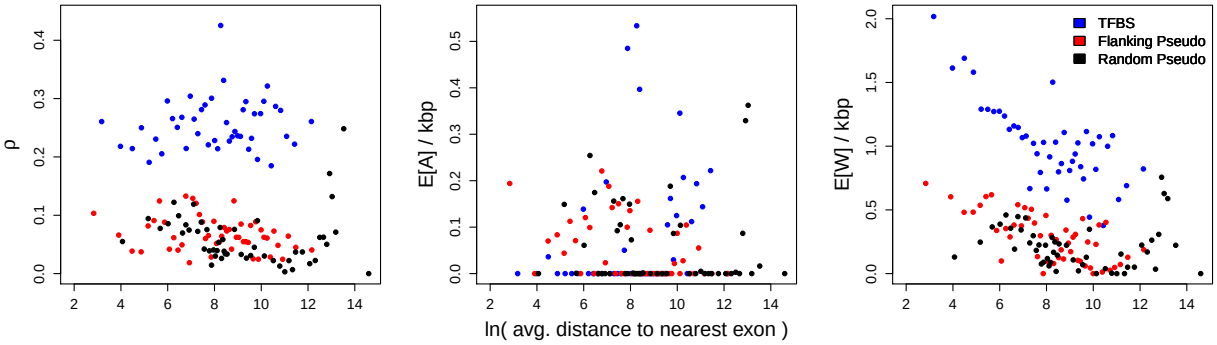
Supplementary Figure 3: Posterior expected values of (**A**) fractions of nucleotides under selection (ρ), (**B**) numbers of adaptive substitutions per kilobase ($\mathbb{E}[A]$/kbp), and (**C**) numbers of weakly deleterious polymorphisms per kilobase ($\mathbb{E}[W]$/kbp). In each case, binding sites containing mutations that produce a net decrease ($-$) or increase ($+$) in predicted binding affinity are compared (Section 8.1). All TFs were pooled for this analysis. Error bars indicate one standard error. We see a clear enrichment for adaptive substitutions in binding sites with affinity-increasing mutations, and a clear enrichment for weakly deleterious polymorphisms in binding sites with affinity-decreasing mutations. However, it should be emphasized that these observations could be influenced in part by an ascertainment bias favoring relatively high-affinity binding sites in the human reference genome, due to our stringent methods for binding site identification. For example, fixed differences that increase binding affinity are more likely to be sampled than those that decrease binding affinity. Similarly, segregating alleles that decrease binding affinity are more likely to be present in low frequencies than in high frequencies, because the reference genome will tend to reflect common alleles (see, e.g., Spivakov et al., 2012).

Nature Genetics: doi:10.1038/ng.2658



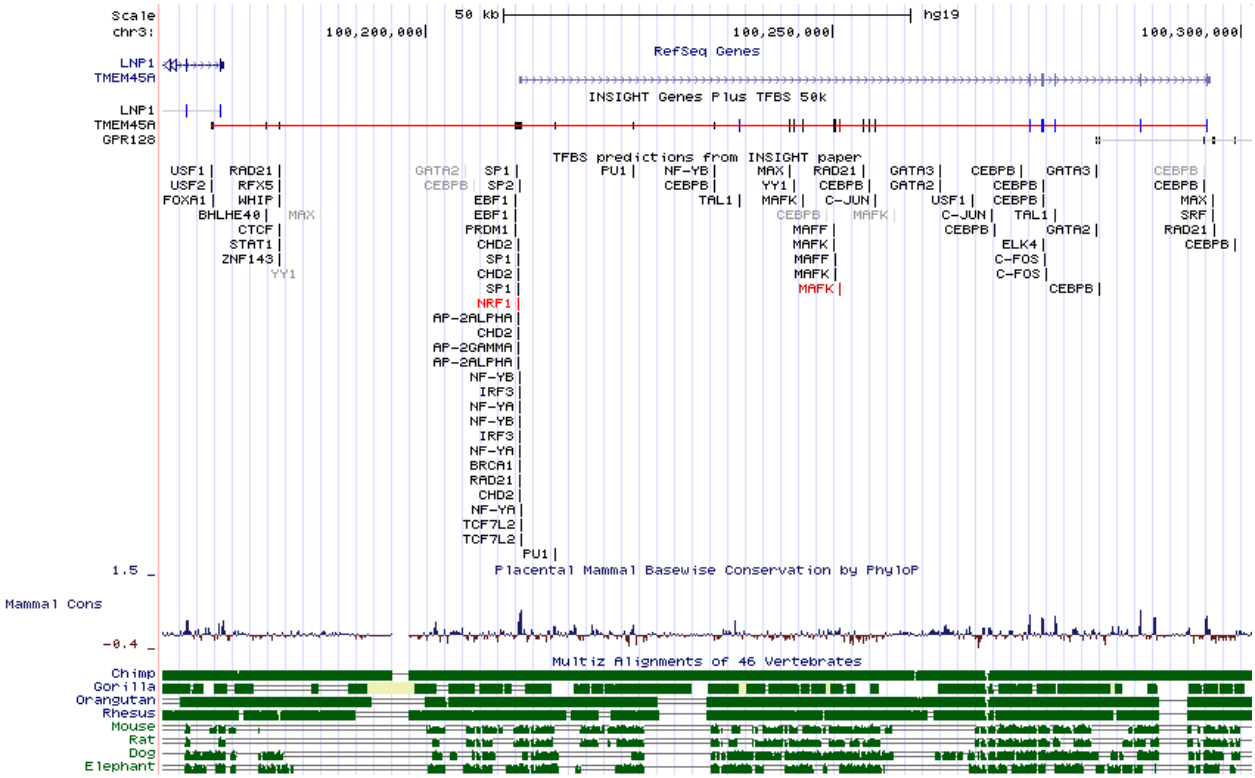
Supplementary Figure 4: Covariates of selection. (**A**) Log median expression of predicted target gene vs. average fraction of sites under selection (ρ) for all TFBSs. Binding sites were ranked by the expression levels (Wang et al., 2008) of associated genes (see Section 8.3) and partitioned into 25 groups. (**B**) Number of cell types in which a binding site is found vs. ρ for all TFs with ChIP-seq data in at least 5 cell types.

Nature Genetics: doi:10.1038/ng.2658



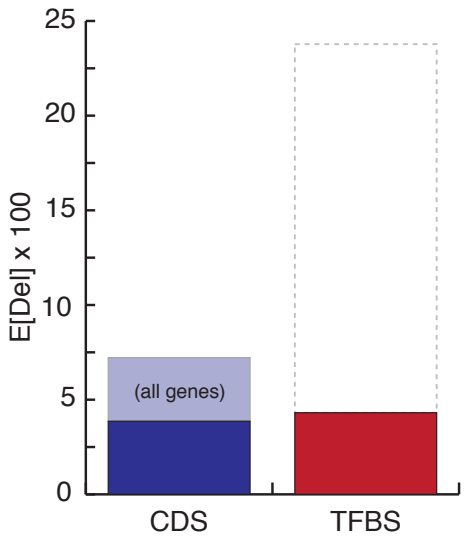
Supplementary Figure 5: Estimated values of ρ, $\mathbb{E}[A]$/kbp, and $\mathbb{E}[W]$/kbp as a function of distance from the nearest coding exon (note log scale). All TFBSs were pooled then partitioned into 50 equally sized groups by log distance, then INSIGHT was applied to each group of TFBSs. Shown are results for the true TFBSs and for two sets of "pseudo-TFBSs" with similar properties to the true TFBSs but located in neutral sites (see Section 8.2).

Nature Genetics: doi:10.1038/ng.2658



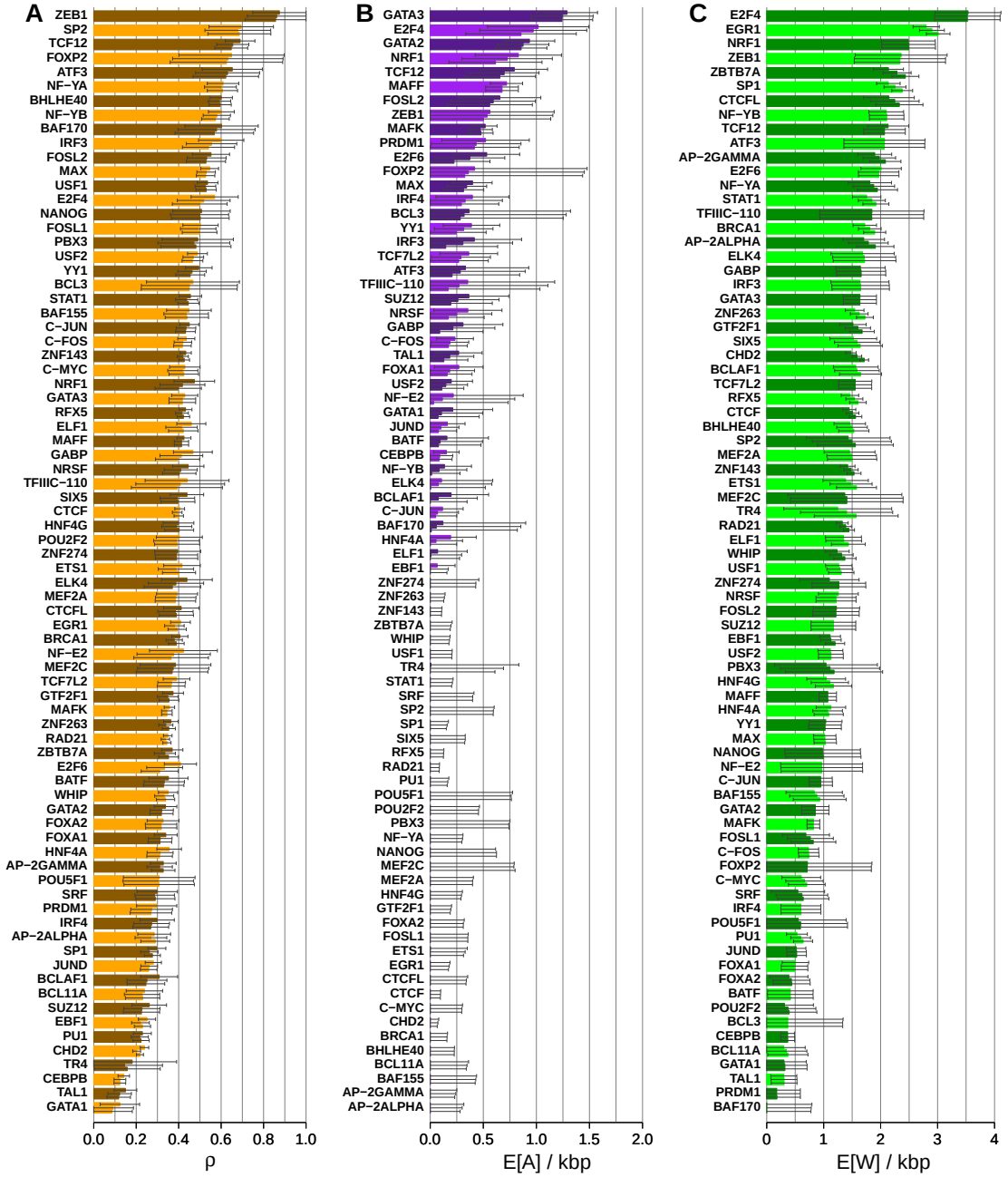
Supplementary Figure 6: Screenshot showing UCSC Genome Browser track based on INSIGHT predictions. The displayed region includes the transmembrane protein 45A (*TMEM45A*) gene and nearby TFBSs. The RefSeq annotation of *TMEM45A* is shown at top. The INSIGHT track set (immediately below) contains two tracks: one displaying TFBSs grouped according to nearby genes and another displaying the individual TFBSs genome-wide. Here, two individual TFBSs are colored red because they each have an expected number of adaptive substitutions exceeding 0.2. This track can be browsed at http://genome-mirror.bscb.cornell.edu (hg19 assembly).

Nature Genetics: doi:10.1038/ng.2658



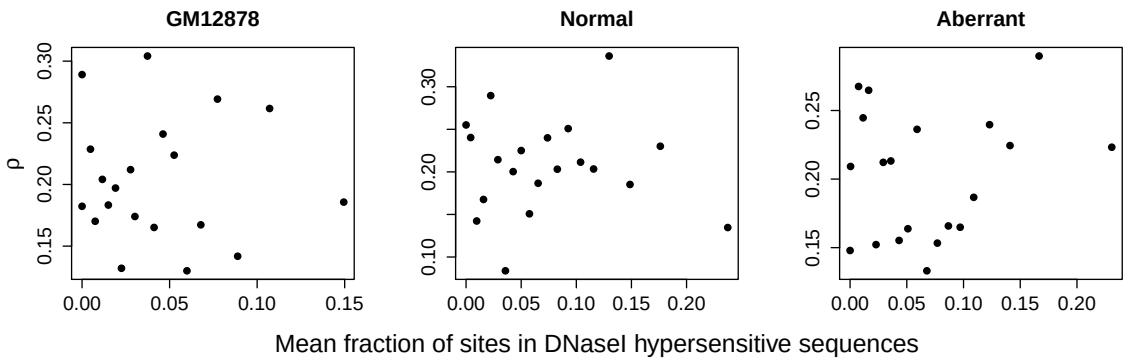
Supplementary Figure 7: Expected numbers of deleterious alleles per haploid genome ($\mathbb{E}[D]$) in CDSs and TFBSs.

Nature Genetics: doi:10.1038/ng.2658



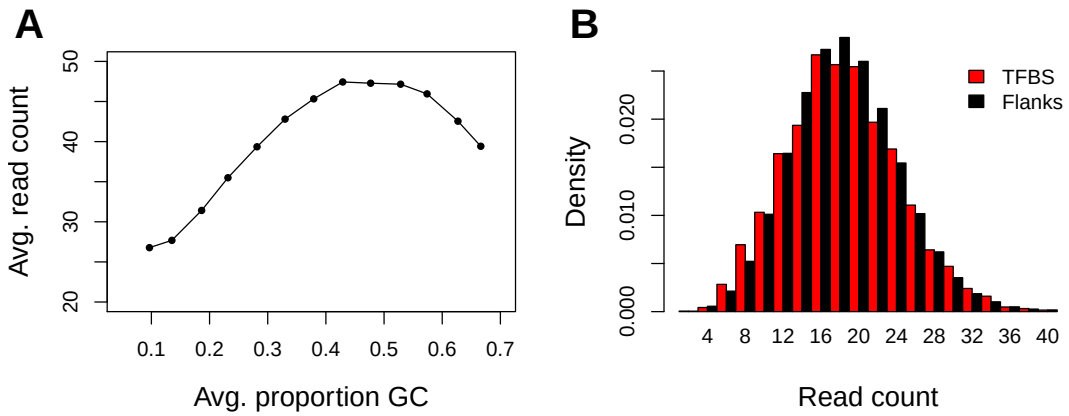
Supplementary Figure 8: Parameter estimates for all TFs based on alternative sets of neutral sites. Three bars are shown for each TF $\times$ parameter. The top bar is the same as that shown in Fig. 2 of the main paper. The middle bar reflects a matching strategy based on local G+C content, and coarse-grained (20 kbp) estimates of neutral diversity (θ) and divergence (λ) (see Section 10.1). The bottom bar reflects an extended matching strategy that also makes use of local estimates of recombination rate and genetic distance from the nearest coding exon.

Nature Genetics: doi:10.1038/ng.2658



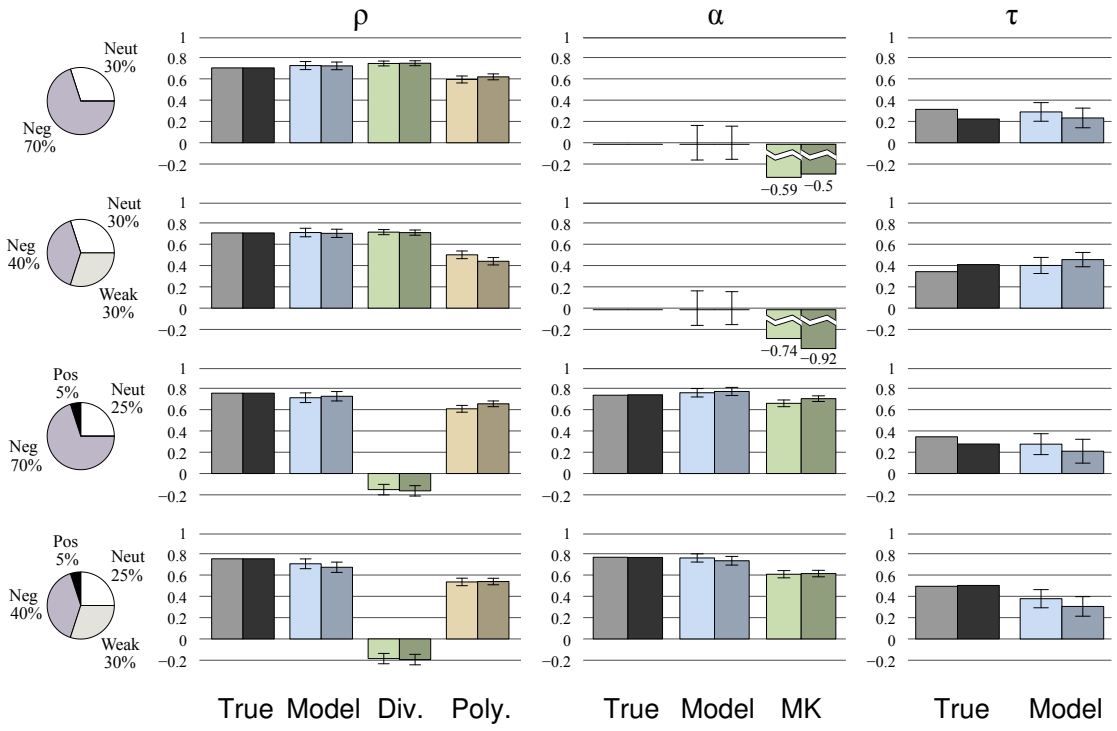
Supplementary Figure 9: Estimates of ρ as a function of the fraction of flanking neutral regions that fall in DNase-I hypersensitive sequences of matched cell types. Results are shown for cell type GM12878, for a collection of four cell types having normal karyotypes ("Normal"), and for a collection of six cancer cell types ("Aberrant"). No significant correlation was observed, indicating no major bias in parameter estimates stemming from differences between binding sites and flanking sequences in chromatin accessibility. See Section 10.3 for details.

Nature Genetics: doi:10.1038/ng.2658



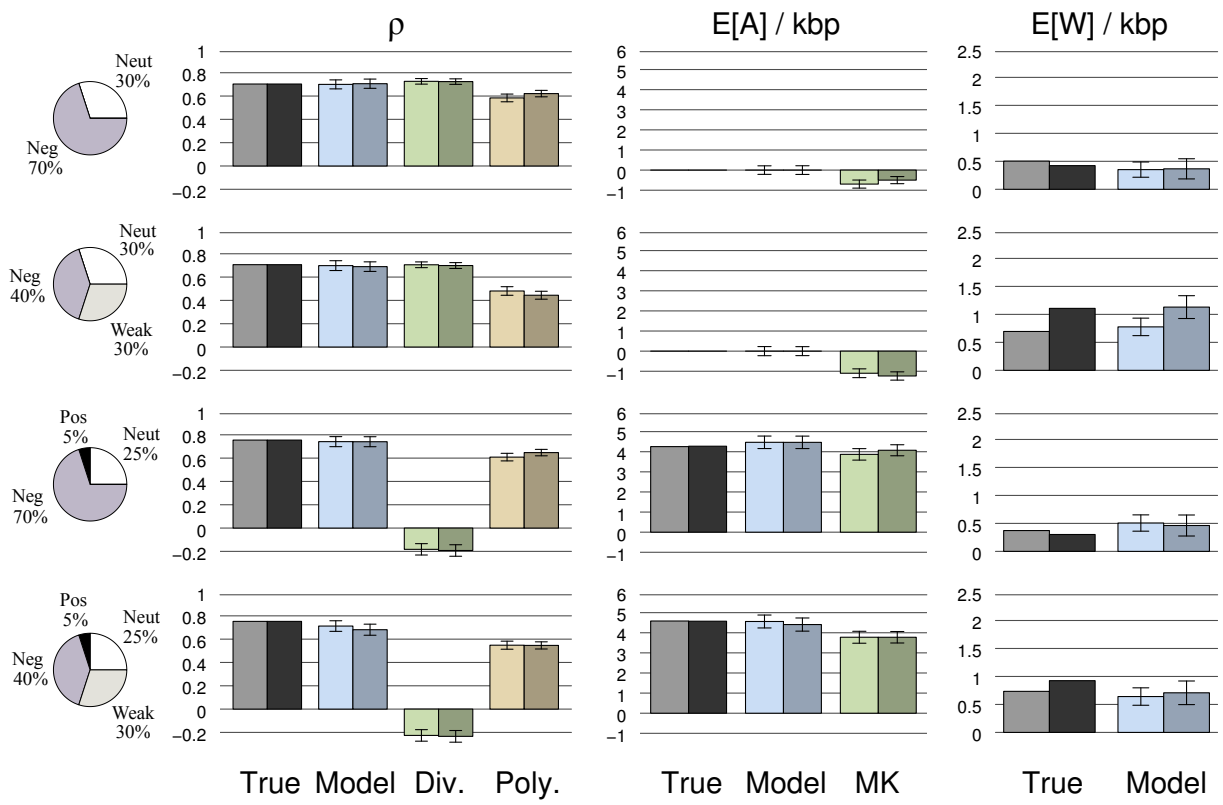
Supplementary Figure 10: Relationships between G+C content and sequencing coverage. **A:** Average read count as a function of G+C content for $\sim$1M putatively neutral 50 bp regions. Sequencing depth shows a clear dependency on G+C content, but is quite high across G+C levels. **B:** Histogram of read counts for unfiltered SNPs in TFBS (red) and flanks (black). The x-axis is truncated at 40. Read counts are based on a single individual (NA12878). Only minor differences in G+C content are observed between TFBSs and flanking sequences. Together, these plots suggest that it is unlikely that G+C-associated differences in coverage significantly influence our inferences of natural selection.

Nature Genetics: doi:10.1038/ng.2658



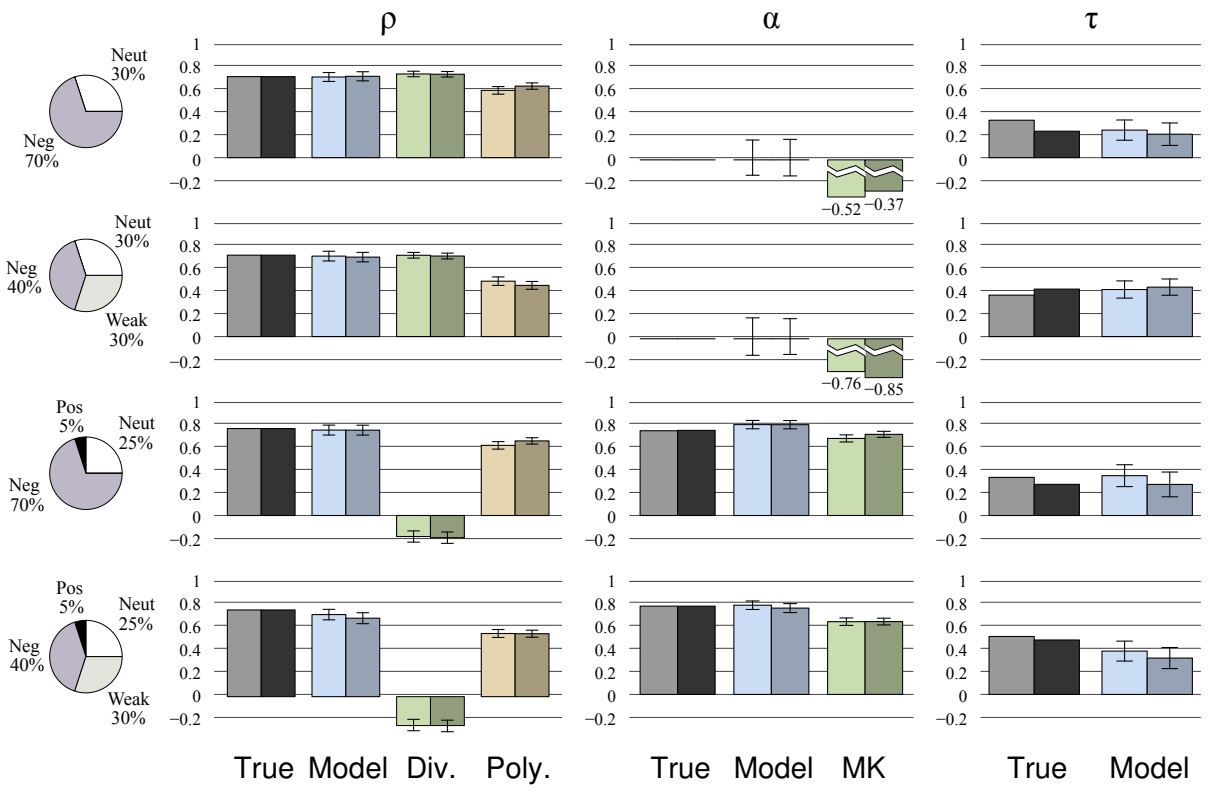
Supplementary Figure 11: Full simulation results including loss of functional constraint in chimpanzee. Model parameter estimates and simple estimators are shown for a full set of simulations (otherwise the same as the one depicted in Fig. 1 & Supplementary Fig. 13) where 20% of the TFBS modeled switch from the specified selective modes to completely neutral evolution in the chimpanzee lineage. See Section 12 of the Supplementary Note for discussion.

Nature Genetics: doi:10.1038/ng.2658



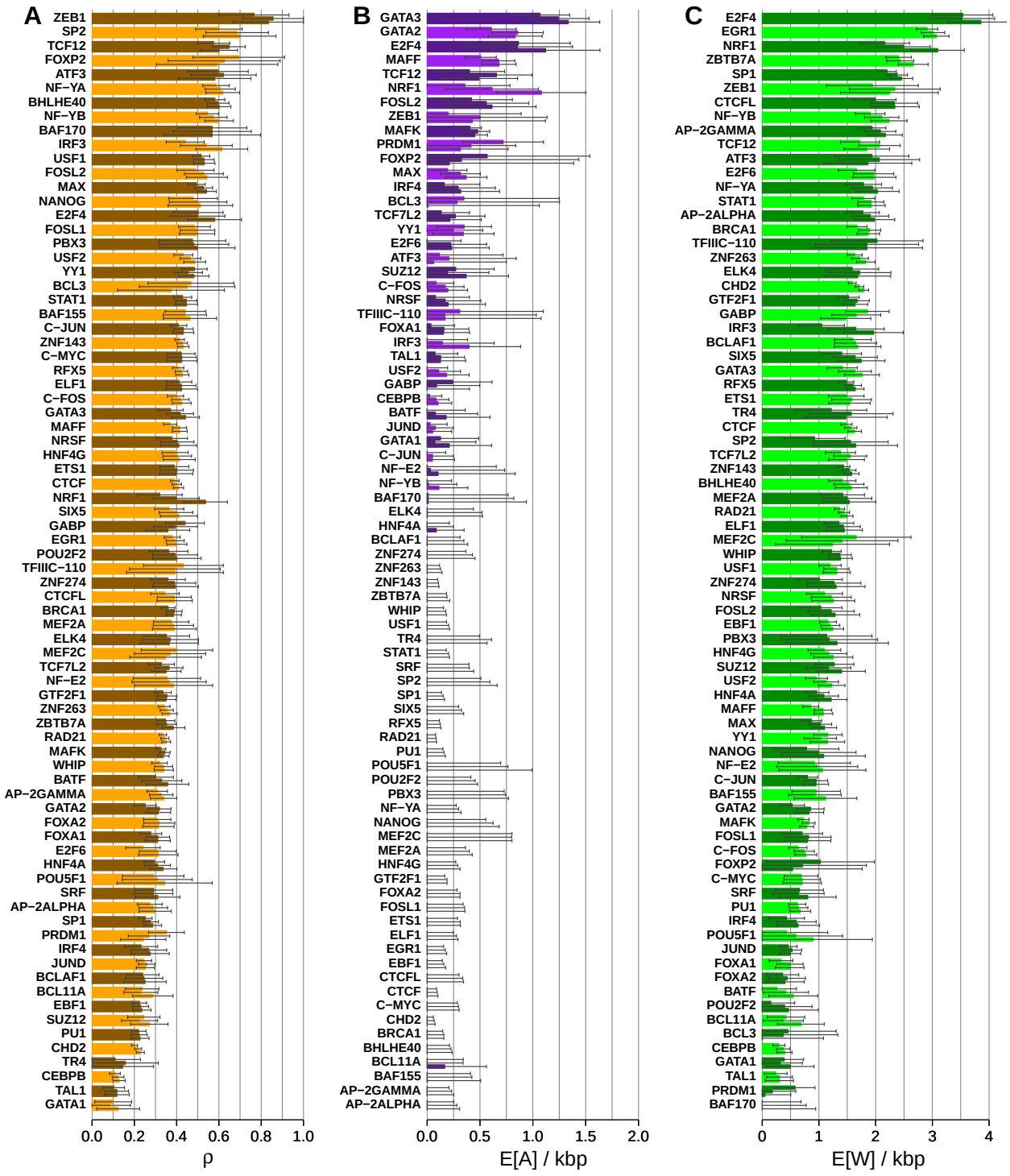
Supplementary Figure 12: Full simulation results with expected values. This is an expanded version of Fig. 1 in the main paper.

Nature Genetics: doi:10.1038/ng.2658



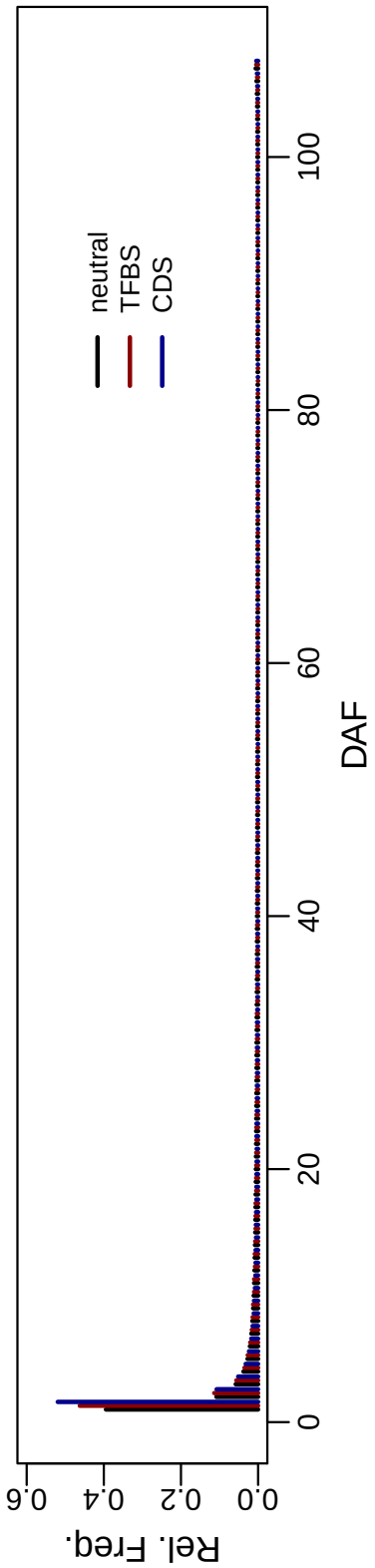
Supplementary Figure 13: Full simulation results with α and τ. Compare with Fig. 1 and Supplementary Fig. 12.

Nature Genetics: doi:10.1038/ng.2658



Supplementary Figure 14: Estimates of key parameters under three different choices of f, the threshold distinguishing high- and low-frequency polymorphisms. For each transcription factor (alternating colors), parameter estimates are shown as in Fig. 2, based on frequency thresholds of 10% 15%, and 20% (top to bottom within each group). The middle bars (15%) are the same as in Fig. 2 and are used to determine the order of the TFs. Error bars indicate one standard error (see Section 1.4). Notice that the parameter estimates are generally fairly insensitive to f relative to their variance.

Nature Genetics: doi:10.1038/ng.2658



Supplementary Figure 15: Full site frequency spectra for TFBSs, CDSs, and neutral flanks

Nature Genetics: doi:10.1038/ng.2658

Supplementary Table 1: Summary of TFBSs by transcription factor

TF	nTFBS	nbases	nunfilt[a]	diffs[b]	L[c]	H[d]
AP-2ALPHA	10346	93114	85313	204.9	372.1	74.9
AP-2GAMMA	14357	129213	118298	266.1	522.2	100.8
ATF3	1719	15471	14618	22.4	49.0	5.0
BAF155	4285	34280	33158	62.2	96.0	23.0
BAF170	895	6265	6077	8.9	8.0	3.0
BATF	7979	63832	62172	155.2	174.2	47.8
BCL11A	5728	80192	75703	206.6	235.7	70.3
BCL3	1164	9312	8970	20.4	21.1	4.9
BCLAF1	7288	94744	76434	192.6	313.9	61.1
BHLHE40	17729	124103	83676	112.4	249.1	47.9
BRCA1	14545	218175	176635	330.6	701.6	155.4
C-FOS	36656	293248	285088	655.7	806.9	192.1
C-JUN	28778	230224	223853	481.0	673.5	149.5
C-MYC	11439	68634	60402	115.1	164.2	46.8
CEBPB	96137	961370	921911	3152.3	3363.1	971.9
CHD2	106113	1910034	1543000	3954.4	6659.7	1425.3
CTCF	56954	740402	654915	1346.0	2433.9	524.1
CTCFL	4448	57824	44248	85.8	193.9	36.1
E2F4	5418	43344	29384	73.2	149.0	14.0
E2F6	11480	103320	85574	213.4	356.4	60.6
EBF1	44196	397764	374987	1009.8	1466.0	336.0
EGR1	28108	281080	214944	421.1	1077.5	151.5
ELF1	15120	105840	95615	185.0	320.6	61.4
ELK4	6063	42441	32426	67.9	122.1	20.9
ETS1	5885	70620	62976	123.4	224.4	44.6
FOSL1	4192	33536	32392	49.9	82.1	23.9
FOSL2	6127	49016	47585	107.6	136.9	25.1
FOXA1	19751	197510	190216	501.7	557.6	150.4
FOXA2	7793	77930	75305	179.4	214.5	62.5
FOXP2	743	5201	4939	8.1	10.0	1.0
GABP	7139	64251	51873	110.0	189.2	32.8
GATA1	12100	84700	82758	267.9	288.4	82.6
GATA2	27475	192325	188477	632.6	629.6	148.4
GATA3	15617	124936	122180	416.6	462.4	83.6
GTF2F1	14524	217860	184369	394.0	712.4	141.6
HNF4A	13863	166356	159368	389.4	555.3	124.7
HNF4G	6130	73560	70192	142.0	227.1	53.9
IRF3	3296	32960	29388	50.5	91.0	13.0
IRF4	6770	94780	88535	256.0	279.6	73.4
JUND	45537	409833	396734	1127.4	1290.1	350.9
MAFF	50389	453501	424178	1236.9	1386.6	301.4
MAFK	101925	917325	859673	2575.3	2824.0	684.0
MAX	40230	241380	206424	421.4	567.3	115.7
MEF2A	4878	48780	46543	100.1	169.2	33.8
MEF2C	582	5820	5591	9.4	20.0	6.0
NANOG	1103	16545	15809	29.4	45.0	10.0
NF-E2	2586	23274	18451	40.9	57.0	12.0
NF-YA	5823	52407	48406	61.7	158.0	24.0
NF-YB	10382	93438	88246	133.8	317.0	42.0
NRF1	6826	81912	47186	123.5	209.7	29.3
NRSF	8729	87290	81348	191.0	275.2	56.8
PBX3	799	7191	6865	10.5	20.0	5.0
POU2F2	3153	25224	24534	48.5	61.1	19.9
POU5F1	873	11349	10958	27.7	35.0	10.0
PRDM1	6031	66341	62812	193.2	176.5	53.5
PU1	33041	363451	344805	942.2	1162.1	315.9
RAD21	70042	910546	814762	1834.4	3120.9	738.1
RFX5	29408	382304	326396	608.5	1175.4	254.6
SIX5	2364	49644	43546	79.7	158.0	35.0
SP1	24154	314002	254879	592.2	1204.6	223.4
SP2	1720	12040	11038	11.3	28.0	3.0
SRF	2546	33098	31467	73.5	98.1	30.9
STAT1	14414	187382	156178	289.1	592.2	99.8
SUZ12	9149	100639	85658	267.5	352.5	80.5
TAL1	16908	253620	242604	783.6	825.9	243.1
TCF12	8061	56427	52997	100.4	174.2	19.8
TCF7L2	14558	131022	126439	326.3	482.3	92.7
TFIIIC-110	874	13984	11896	27.0	46.0	7.0
TR4	1211	16954	15120	40.4	64.8	14.2
USF1	21270	170160	148693	247.8	447.1	85.9
USF2	27288	218304	188939	385.5	579.2	117.8
WHIP	19370	251810	222583	497.4	815.1	175.9
YY1	10868	119548	105439	228.2	304.6	63.4
ZBTB7A	14827	148270	130287	253.7	585.7	112.3
ZEB1	1199	7194	6751	6.6	19.0	0.0
ZNF143	44716	581308	508187	982.6	1814.3	380.7
ZNF263	50634	405072	378538	801.7	1491.4	306.6
ZNF274	1582	39550	36529	78.2	127.0	27.0

[a]Total number of bases after filtering.

[b]Expected number of fixed differences (allowing for uncertainty in ancestral alleles).

[c]Expected number of low-frequency derived polymorphisms.

[d]Expected number of high-frequency derived polymorphisms.

Nature Genetics: doi:10.1038/ng.2658

Supplementary Table 2: Summary of parameter estimates by transcription factor

TF	ρ	α	τ	$\mathbb{E}[A]$	$\mathbb{E}[W]$	$\mathbb{E}[A]$/kbp	$\mathbb{E}[W]$/kbp
AP-2ALPHA	0.290	0.000	0.370	0.000	162.785	0.000	1.908
AP-2GAMMA	0.327	0.000	0.403	0.000	246.600	0.000	2.085
ATF3	0.622	0.134	0.548	3.014	30.197	0.206	2.066
BAF155	0.439	0.000	0.263	0.000	30.849	0.000	0.930
BAF170	0.568	0.007	0.000	0.062	0.000	0.010	0.000
BATF	0.329	0.031	0.116	4.824	25.887	0.078	0.416
BCL11A	0.230	0.000	0.093	0.000	28.096	0.000	0.371
BCL3	0.449	0.124	0.124	2.536	3.322	0.283	0.370
BCLAF1	0.246	0.000	0.337	0.000	125.941	0.000	1.648
BHLHE40	0.596	0.000	0.443	0.000	127.948	0.000	1.529
BRCA1	0.389	0.000	0.411	0.000	335.864	0.000	1.901
C-FOS	0.418	0.074	0.211	48.940	210.674	0.172	0.739
C-JUN	0.432	0.024	0.258	11.421	212.695	0.051	0.950
C-MYC	0.424	0.000	0.208	0.000	42.418	0.000	0.702
CEBPB	0.123	0.024	0.079	76.781	340.347	0.083	0.369
CHD2	0.218	0.000	0.332	0.000	2643.700	0.000	1.713
CTCF	0.400	0.000	0.354	0.000	1022.510	0.000	1.561
CTCFL	0.389	0.000	0.461	0.000	103.106	0.000	2.330
E2F4	0.499	0.342	0.633	25.082	103.570	0.854	3.525
E2F6	0.311	0.089	0.403	19.145	168.346	0.224	1.967
EBF1	0.229	0.000	0.251	0.000	450.049	0.000	1.200
EGR1	0.393	0.000	0.533	0.000	647.325	0.000	3.012
ELF1	0.422	0.000	0.361	0.000	137.090	0.000	1.434
ELK4	0.370	0.002	0.388	0.132	55.718	0.004	1.718
ETS1	0.402	0.000	0.375	0.000	99.301	0.000	1.577
FOSL1	0.499	0.000	0.264	0.000	26.441	0.000	0.816
FOSL2	0.528	0.247	0.356	26.624	57.858	0.559	1.216
FOXA1	0.312	0.060	0.131	30.277	92.976	0.159	0.489
FOXA2	0.318	0.000	0.121	0.000	33.044	0.000	0.439
FOXP2	0.625	0.197	0.291	1.608	3.507	0.326	0.710
GABP	0.395	0.042	0.387	4.624	86.055	0.089	1.659
GATA1	0.085	0.023	0.071	6.105	26.448	0.074	0.320
GATA2	0.319	0.253	0.207	160.933	160.908	0.854	0.854
GATA3	0.417	0.362	0.367	151.096	199.963	1.237	1.637
GTF2F1	0.355	0.000	0.368	0.000	308.933	0.000	1.676
HNF4A	0.310	0.000	0.255	0.041	173.508	0.000	1.089
HNF4G	0.403	0.000	0.300	0.000	82.282	0.000	1.172
IRF3	0.541	0.085	0.463	4.297	48.500	0.146	1.650
IRF4	0.269	0.101	0.151	25.968	53.223	0.293	0.601
JUND	0.259	0.027	0.127	30.672	208.642	0.077	0.526
MAFF	0.414	0.231	0.270	286.375	454.990	0.675	1.073
MAFK	0.345	0.159	0.201	411.030	705.242	0.478	0.820
MAX	0.527	0.153	0.310	64.742	211.725	0.314	1.026
MEF2A	0.385	0.000	0.346	0.000	69.794	0.000	1.500
MEF2C	0.369	0.000	0.328	0.000	7.869	0.000	1.407
NANOG	0.502	0.000	0.287	0.000	15.674	0.000	0.991
NF-E2	0.364	0.014	0.256	0.582	17.855	0.032	0.968
NF-YA	0.606	0.000	0.526	0.000	94.099	0.000	1.944
NF-YB	0.573	0.009	0.517	1.162	185.953	0.013	2.107
NRF1	0.398	0.234	0.491	28.953	117.632	0.614	2.493
NRSF	0.403	0.072	0.298	13.862	99.053	0.170	1.218
PBX3	0.481	0.000	0.340	0.000	8.117	0.000	1.182
POU2F2	0.393	0.000	0.124	0.000	9.669	0.000	0.394
POU5F1	0.308	0.000	0.147	0.000	6.499	0.000	0.593
PRDM1	0.269	0.134	0.049	26.070	11.266	0.415	0.179
PU1	0.223	0.000	0.150	0.000	220.125	0.000	0.638
RAD21	0.345	0.000	0.314	0.000	1175.120	0.000	1.442
RFX5	0.422	0.000	0.379	0.000	522.959	0.000	1.602
SIX5	0.397	0.000	0.384	0.000	71.474	0.000	1.641
SP1	0.276	0.000	0.434	0.000	606.600	0.000	2.380
SP2	0.686	0.000	0.545	0.000	17.196	0.000	1.558
SRF	0.290	0.000	0.164	0.000	20.196	0.000	0.642
STAT1	0.446	0.000	0.438	0.000	299.468	0.000	1.917
SUZ12	0.225	0.062	0.232	16.590	100.306	0.194	1.171
TAL1	0.118	0.039	0.069	31.013	74.063	0.128	0.305
TCF12	0.648	0.345	0.562	34.801	109.611	0.657	2.068
TCF7L2	0.365	0.103	0.341	33.842	196.407	0.268	1.553
TFIIIC-110	0.392	0.074	0.408	2.001	21.956	0.168	1.846
TR4	0.158	0.000	0.304	0.000	23.766	0.000	1.572
USF1	0.529	0.000	0.368	0.000	194.622	0.000	1.309
USF2	0.465	0.053	0.305	20.592	212.847	0.109	1.127
WHIP	0.338	0.000	0.313	0.000	305.599	0.000	1.373
YY1	0.455	0.113	0.293	25.840	108.053	0.245	1.025
ZBTB7A	0.352	0.000	0.471	0.000	317.287	0.000	2.435
ZEB1	0.856	0.510	0.788	3.385	15.789	0.501	2.339
ZNF143	0.426	0.000	0.364	0.000	779.046	0.000	1.533
ZNF263	0.353	0.000	0.370	0.000	652.790	0.000	1.725
ZNF274	0.390	0.000	0.305	0.000	46.307	0.000	1.268

Nature Genetics: doi:10.1038/ng.2658

Supplementary Table 3: Estimates for TFBS, CDS, and CDS2 elements under various partitioning schemes

Set[a]		Type[b]	ρ	$\mathbb{E}[A]$/kbp	$\mathbb{E}[W]$/kbp	N[c]	bp[d]
Global		TFBS	0.25 (0.01)	0 (0.03)	1.06 (0.03)	78	13285443
		CDS	0.65 (0.00)	0 (0.02)	0.90 (0.02)	15864	19518616
		CDS2	0.80 (0.00)	0 (0.02)	1.00 (0.02)	15816	6576609
Selection Class	CDS	Mam Other	0.67 (0.00)	0 (0.02)	0.93 (0.02)	13248	13811470
		Mam PS	0.26 (0.05)	4.2e-4 (0.22)	0.04 (0.20)	202	254451
		Hum Housekeeping	0.73 (0.03)	0 (0.10)	0.77 (0.11)	436	319418
		Hum PS	0.68 (0.04)	0.56 (0.18)	0.90 (0.17)	161	183336
		Hum NS	0.68 (0.02)	0 (0.07)	1.23 (0.09)	489	553012
		Hum Other	0.68 (0.00)	0 (0.02)	0.90 (0.02)	11133	11118275
	CDS2	Mam Other	0.82 (0.01)	0 (0.02)	1.04 (0.03)	13126	4601518
		Mam PS	0.39 (0.08)	0.45 (0.35)	0.23 (0.33)	201	85018
		Hum Housekeeping	0.92 (0.02)	0 (0.09)	0.78 (0.12)	436	107425
		Hum PS	0.83 (0.05)	0.50 (0.23)	0.78 (0.22)	161	61352
		Hum NS	0.87 (0.02)	0 (0.08)	1.61 (0.12)	489	185493
		Hum Other	0.84 (0.01)	0 (0.02)	0.98 (0.03)	11131	3729330
By TF		TFBS	0.33 (0.02)	0.12 (0.02)	1.28 (0.03)	78	13285443
By Gene		CDS	0.56 (0.01)	0.24 (0.01)	0.86 (0.02)	15864	19581186
By Gene		CDS2	0.51 (0.02)	0.18 (0.03)	0.73 (0.07)	15816	6576609

Standard errors are shown in parentheses following estimates of ρ, $\mathbb{E}[A]$/kbp, and $\mathbb{E}[W]$/kbp.

[a] Partition over which the model is fit.

[b] Transcription factor binding sites (TFBS), coding sequences (CDS), or second codon positions (CDS2). Subclasses of CDS/CDS2 categories indicate housekeeping genes, genes under positive selection (PS), genes under negative selection (NS), or other genes. These are shown separately for classifications based on mammals (Mam) or human populations (Hum) (see Section 6).

[c] Number of TFs or genes in set.

[d] Total number of bases analyzed across all elements.

Nature Genetics: doi:10.1038/ng.2658

Supplementary Table 4: Descriptive labels for transcription factors

TF	Labels[a]	TF	Labels[a]
AP-2ALPHA	activator, repressor, development	JUND	apoptosis, repressor, oncogene
AP-2GAMMA	activator, development, repressor	MAFF	oncogene, activator
ATF3	activator, repressor, stress-response	MAFK	leucine-zipper, repressor, activator
BAF155	chromatin, activator, repressor, development, neuronal, differentiation	MAX	helix-loop-helix, activator, repressor
BAF170	chromatin, activator, repressor, neuronal, development, differentiation	MEF2A	differentiation, activator, development
BATF	leucine-zipper, repressor	MEF2C	differentiation, activator, development
BCL11A	zinc-finger, oncogene, differentiation	NANOG	homeobox, differentiation, activator, repressor, development
BCL3	immune-response, oncogene, activator	NF-E2	leucine-zipper, differentiation, activator
BCLAF1	apoptosis, stress-response	NF-YA	activator
BHLHE40	helix-loop-helix, differentiation	NF-YB	activator
BRCA1	oncogene, DNA-repair	NRF1	activator, development
CEBPB	immune-response, activator, differentiation	NRSF	repressor, zinc-finger, Kruppel-type, neuronal
C-FOS	oncogene, activator, leucine-zipper, development, differentiation	PBX3	activator, homeobox
CHD2	helicase	POU2F2	activator, development, immune-response
C-JUN	oncogene	POU5F1	homeobox, development
C-MYC	differentiation, oncogene	PRDM1	immune-response, repressor, oncogene
CTCFL	zinc-finger, chromatin, oncogene	PU1	differentiation, activator, oncogene
CTCF	zinc-finger, repressor, activator	RAD21	DNA-repair, apoptosis
E2F4	activator	RFX5	activator
E2F6	repressor	SIX5	homeobox, development
EBF1	activator, helix-loop-helix	SP1	activator, repressor, apoptosis, immune-response, development
EGR1	zinc-finger, differentiation	SP2	zinc-finger, immune-response, activator, repressor
ELF1	activator, repressor	SRF	apoptosis, differentiation, development
ELK4	oncogene	STAT1	activator, immune-response
ETS1	activator, repressor, differentiation, apoptosis, oncogene	SUZ12	chromatin, repressor
FOSL1	leucine-zipper, immune-response	TAL1	differentiation, oncogene
FOSL2	leucine-zipper, differentiation	TCF12	helix-loop-helix, differentiation, neuronal
FOXA1	FOX, activator, differentiation, development, oncogene	TCF7L2	HMG-gox, activator, repressor, development
FOXA2	FOX, development, differentiation, activator	TFIIIC-110	activator
FOXP2	FOX, development	TR4	repressor, activator, development, differentiation
GABP	activator	USF1	helix-loop-helix
GATA1	GATA, zinc-finger, development, activator	USF2	helix-loop-helix
GATA2	GATA, zinc-finger, development, activator	WHIP	DNA-repair
GATA3	GATA, zinc-finger, development, activator	YY1	Kruppel-type, zinc-finger, activator, repressor, development, differentiation
GTF2F1	activator	ZBTB7A	zinc-finger, differentiation, repressor, development
HNF4A	nuclear-receptor, development	ZEB1	zinc-finger, repressor, activator, neuronal, differentiation
HNF4G	nuclear-receptor, development	ZNF143	zinc-finger, activator
IRF3	IRF, immune-response	ZNF263	zinc-finger, repressor, Kruppel-type
IRF4	IRF, immune-response, activator, differentiation, oncogene	ZNF274	repressor, zinc-finger, Kruppel-type

[a]Coarse-grained descriptive labels manually assigned based on GeneCards, UCSC Genes, RefSeq and other sources.

Nature Genetics: doi:10.1038/ng.2658

Supplementary Table 5: Label enrichments by ρ

Label[a]	Top[b]	Bottom[c]	p_1[d]	p_2[e]
activator	16	28	0.345	0.812
apoptosis	0	6	1.000	0.079
development	5	21	0.985	0.051
differentiation	9	16	0.462	0.728
helix-loop-helix	5	1	0.014	0.999
immune-response	5	5	0.199	0.937
leucine-zipper	2	4	0.658	0.685
oncogene	3	13	0.959	0.137
repressor	8	18	0.721	0.470
zinc-finger	4	11	0.818	0.388

[a] See Supplementary Table 4. Only labels that appear at least five times are considered.

[b] Number of TFs with this label ranked in top third by ρ.

[c] Number of TFs with this label ranked in bottom two thirds by ρ.

[d] p-value for enrichment in top third (one-sided Fisher's exact test).

[e] p-value for depletion in top third (one-sided Fisher's exact test).

Supplementary Table 6: Label enrichments by α

Label[a]	Top[b]	Bottom[c]	p_1[d]	p_2[e]
activator	17	27	0.188	0.916
apoptosis	0	6	1.000	0.079
development	8	18	0.721	0.470
differentiation	7	18	0.827	0.337
helix-loop-helix	3	3	0.315	0.909
immune-response	4	6	0.441	0.801
leucine-zipper	3	3	0.315	0.909
oncogene	6	10	0.453	0.759
repressor	10	16	0.333	0.825
zinc-finger	5	10	0.612	0.627

[a] See Supplementary Table 4. Only labels that appear at least five times are considered.

[b] Number of TFs with this label ranked in top third by α.

[c] Number of TFs with this label ranked in bottom two thirds by α.

[d] p-value for enrichment in top third (one-sided Fisher's exact test).

[e] p-value for depletion in top third (one-sided Fisher's exact test).

Nature Genetics: doi:10.1038/ng.2658

Supplementary Table 7: TF × GO category combinations with elevated ρ estimates

TF	GO Id	GO Term	$\lvert g\rvert^a$	$\lvert\bar{g}\rvert^b$	$\hat{\rho}_g{}^c$	$\hat{\rho}_{\bar{g}}{}^d$	Ratio^e	p^f
CEBPB	GO:0007265	Ras protein signal transduction	737	94236	0.7465	0.1218	6.13	0.038
CEBPB	GO:0071260	cellular response to mechanical stimulus	671	94302	0.7362	0.1222	6.03	0.027
CEBPB	GO:0001503	ossification	1046	93927	0.6877	0.1196	5.75	0.031
IRF4	GO:0007155	cell adhesion	373	6273	0.9999	0.2320	4.31	0.007
BCLAF1	GO:0044281	small molecule metabolic process	608	6263	0.8565	0.2178	3.93	0.018
CEBPB	GO:0007517	muscle organ development	1015	93958	0.4641	0.1231	3.77	0.028
SRF	GO:0007275	multicellular organismal development	148	2340	0.9995	0.2669	3.75	0.004
CHD2	GO:0000080	G1 phase of mitotic cell cycle	547	99424	0.7699	0.2080	3.70	0.006
CEBPB	GO:0032496	response to lipopolysaccharide	1894	93079	0.4438	0.1203	3.69	0.042
ZNF274	GO:0006355	regulation of transcription, DNA-dependent	1060	466	0.5090	0.1385	3.67	0.040
AP-2ALPHA	GO:0006810	transport	496	9430	0.9998	0.2734	3.66	0.010
BCLAF1	GO:0006955	immune response	311	6560	0.7623	0.2203	3.46	0.018
AP-2GAMMA	GO:0001701	in utero embryonic development	521	13242	0.9999	0.3024	3.31	0.004
GATA2	GO:0007219	Notch signaling pathway	567	26551	0.9998	0.3053	3.28	0.042
ZNF274	GO:0006351	transcription, DNA-dependent	1066	460	0.4950	0.1599	3.10	0.038
EBF1	GO:0006366	transcription from RNA polymerase II promoter	1604	41243	0.6738	0.2247	3.00	0.038
ZBTB7A	GO:0007268	synaptic transmission	521	13065	0.9998	0.3402	2.94	0.026
MAFK	GO:0034097	response to cytokine stimulus	678	99971	0.9999	0.3434	2.91	0.002
MAFK	GO:0030318	melanocyte differentiation	608	100041	0.9999	0.3437	2.91	0.021
MAFK	GO:0008104	protein localization	576	100073	0.9999	0.3439	2.91	0.037
MAFK	GO:0007416	synapse assembly	509	100140	0.9998	0.3445	2.90	0.001
TCF7L2	GO:0001525	angiogenesis	496	13845	0.9998	0.3548	2.82	0.036
JUND	GO:0006974	response to DNA damage stimulus	689	44162	0.7093	0.2520	2.81	0.006
JUND	GO:0045087	innate immune response	1323	43528	0.6916	0.2464	2.81	0.026
MEF2A	GO:0006355	regulation of transcription, DNA-dependent	764	4045	0.9999	0.3564	2.81	0.004
NRF1	GO:0010467	gene expression	484	5920	0.9997	0.3597	2.78	0.046
CHD2	GO:0007154	cell communication	911	99060	0.5689	0.2090	2.72	0.023
CHD2	GO:0007417	central nervous system development	1067	98904	0.5544	0.2048	2.71	0.031
NRSF	GO:0006915	apoptotic process	680	7820	0.9999	0.3712	2.69	0.019
CHD2	GO:0001570	vasculogenesis	1195	98776	0.5450	0.2053	2.65	0.005
WHIP	GO:0007267	cell-cell signaling	486	18249	0.8446	0.3187	2.65	0.018
RAD21	GO:0042472	inner ear morphogenesis	653	67358	0.8826	0.3386	2.61	0.041
EBF1	GO:0009952	anterior/posterior pattern specification	517	42330	0.6004	0.2308	2.60	0.000
ZNF263	GO:0007283	spermatogenesis	1235	46670	0.8969	0.3466	2.59	0.023
ZBTB7A	GO:0007275	multicellular organismal development	843	12743	0.7869	0.3060	2.57	0.006
C-MYC	GO:0042493	response to drug	541	10586	0.9997	0.3896	2.57	0.017
NRSF	GO:0008284	positive regulation of cell proliferation	406	8094	0.9998	0.3921	2.55	0.024
CTCF	GO:0009952	anterior/posterior pattern specification	721	54322	0.9999	0.3925	2.55	0.000
IRF4	GO:0006468	protein phosphorylation	392	6254	0.7008	0.2754	2.54	0.018
RAD21	GO:0001568	blood vessel development	506	67505	0.8550	0.3379	2.53	0.023
EBF1	GO:0001701	in utero embryonic development	1176	41671	0.5724	0.2284	2.51	0.036

[a] Number of TFBSs assigned to GO category.

[b] Number of TFBSs not assigned to GO category.

[c] Estimate of ρ for TFBSs assigned to GO category.

[d] Estimate of ρ for TFBSs not assigned to GO category.

[e] Ratio of $\hat{\rho}_g$ to $\hat{\rho}_{\bar{g}}$. Only TF × GO category combinations with $\hat{\rho}_g/\hat{\rho}_{\bar{g}} > 2.5$ are shown.

[f] Two-sided p-value for null hypothesis that $\rho_g = \rho_{\bar{g}}$ based on a likelihood ratio test. Only cases with $p < 0.05$ are shown. These are nominal p-values not corrected for multiple comparisons and are meant only as a guide for follow-up study.

Nature Genetics: doi:10.1038/ng.2658

Supplementary Table 8: TF $\times$ GO category combinations with elevated $\mathbb{E}[A]$ estimates

TF	GO Id	GO Term	$\lvert g\rvert^a$	$\lvert\bar{g}\rvert^b$	$\mathbb{E}[A]_g{}^c$	$\mathbb{E}[A]_{\bar{g}}{}^d$	Ratio^e	p^f
IRF3	GO:0051301	cell division	159	2914	0.3019	0.1411	2.14	0.049
TCF12	GO:0042493	response to drug	364	7448	1.2900	0.6186	2.09	0.021
USF2	GO:0016477	cell migration	527	26047	0.1990	0.1062	1.87	0.014
NRSF	GO:0007596	blood coagulation	467	8033	0.3070	0.1640	1.87	0.005
MAFK	GO:0007200	phospholipase C-activating G-protein coupled receptor signaling pathway	550	100099	0.8861	0.4782	1.85	0.002
CEBPB	GO:0007628	adult walking behavior	495	94478	0.1512	0.0830	1.82	0.001
C-JUN	GO:0006974	response to DNA damage stimulus	501	27901	0.0914	0.0508	1.80	0.018
FOSL2	GO:0008285	negative regulation of cell proliferation	347	5670	0.9586	0.5402	1.77	0.035
NRSF	GO:0042493	response to drug	417	8083	0.2917	0.1657	1.76	0.018
JUND	GO:0006974	response to DNA damage stimulus	689	44162	0.1350	0.0769	1.76	0.002
MAFK	GO:0019233	sensory perception of pain	748	99901	0.8196	0.4779	1.71	0.002
JUND	GO:0006486	protein glycosylation	502	44349	0.1310	0.0771	1.70	0.009
C-FOS	GO:0007568	aging	602	35543	0.2890	0.1705	1.69	0.015
C-FOS	GO:0030324	lung development	533	35612	0.2827	0.1709	1.65	0.027
JUND	GO:0007389	pattern specification process	540	44311	0.1253	0.0772	1.62	0.017
MAFK	GO:0007613	memory	627	100022	0.7767	0.4785	1.62	0.011
CEBPB	GO:0006887	exocytosis	580	94393	0.1299	0.0831	1.56	0.009
MAFK	GO:0006898	receptor-mediated endocytosis	559	100090	0.7373	0.4790	1.54	0.024
FOXA1	GO:0007399	nervous system development	779	18673	0.2383	0.1549	1.54	0.012
MAFK	GO:0070555	response to interleukin-1	524	100125	0.7330	0.4791	1.53	0.032
JUND	GO:0051216	cartilage development	542	44309	0.1178	0.0773	1.53	0.032
JUND	GO:0051090	regulation of sequence-specific DNA binding transcription factor activity	522	44329	0.1178	0.0773	1.52	0.034
MAFF	GO:0007218	neuropeptide signaling pathway	601	49094	1.0094	0.6729	1.50	0.032
FOXA1	GO:0051301	cell division	573	18879	0.2339	0.1560	1.50	0.036
MAFK	GO:0050728	negative regulation of inflammatory response	549	100100	0.7087	0.4791	1.48	0.041
MAFK	GO:0007610	behavior	658	99991	0.7061	0.4789	1.47	0.031
SUZ12	GO:0008285	negative regulation of cell proliferation	467	8352	0.2758	0.1885	1.46	0.034
MAFK	GO:0001569	patterning of blood vessels	596	100053	0.6871	0.4792	1.43	0.049
CEBPB	GO:0007157	heterophilic cell-cell adhesion	541	94432	0.1186	0.0832	1.43	0.037
C-FOS	GO:0006366	transcription from RNA polymerase II promoter	1282	34863	0.2417	0.1700	1.42	0.026
GATA1	GO:0006355	regulation of transcription, DNA-dependent	1740	10088	0.0996	0.0702	1.42	0.007
C-FOS	GO:0007420	brain development	953	35192	0.2414	0.1706	1.41	0.047
GATA1	GO:0006351	transcription, DNA-dependent	2217	9611	0.0975	0.0693	1.41	0.004
CEBPB	GO:0007169	transmembrane receptor protein tyrosine kinase signaling pathway	1157	93816	0.1162	0.0830	1.40	0.007

[a]Number of TFBSs assigned to GO category.

[b]Number of TFBSs not assigned to GO category.

[c]Estimate for TFBSs assigned to GO category (per kbp).

[d]Estimate for TFBSs not assigned to GO category (per kbp).

[e]Ratio of $\mathbb{E}[A]_g$ to $\mathbb{E}[A]_{\bar{g}}$.Only TF $\times$ GO category combinations with $\mathbb{E}[A]_g/\mathbb{E}[A]_{\bar{g}} > 1.4$ are shown.

[f]One-sided p-value for the null hypothesis that $\mathbb{E}[A]_g \leq \mathbb{E}[A]_{\bar{g}}$ indicating statistical significance of elevations in $\mathbb{E}[A]_g$ (see Section 8.3). Only cases with $p < 0.05$ are shown. These are nominal p-values not corrected for multiple comparisons and are meant only as a guide for follow-up study.

Nature Genetics: doi:10.1038/ng.2658

Supplementary Table 9: TF $\times$ GO category combinations with elevated $\mathbb{E}[W]$ estimates

TF	GO Id	GO Term	$\lvert g\rvert^a$	$\lvert\bar{g}\rvert^b$	$\mathbb{E}[W]_g{}^c$	$\mathbb{E}[W]_{\bar{g}}{}^d$	Ratio^e	p^f
ZEB1	GO:0044281	small molecule metabolic process	101	1033	5.7963	2.0031	2.89	0.038
BAF155	GO:0042493	response to drug	245	3912	2.0448	0.8384	2.44	0.002
ELK4	GO:0006412	translation	263	5449	3.5575	1.5942	2.23	0.012
EBF1	GO:0009952	anterior/posterior pattern specification	517	42330	2.6081	1.1869	2.20	0.000
GABP	GO:0015031	protein transport	293	6386	3.4200	1.6150	2.12	0.004
FOSL1	GO:0006810	transport	203	3861	1.6523	0.7882	2.10	0.023
BCL11A	GO:0007186	G-protein coupled receptor signaling pathway	279	5374	0.7223	0.3530	2.05	0.002
ELK4	GO:0044267	cellular protein metabolic process	325	5387	3.1901	1.5953	2.00	0.016
GATA2	GO:0019221	cytokine-mediated signaling pathway	491	26627	1.6736	0.8402	1.99	0.002
BHLHE40	GO:0006508	proteolysis	698	16458	2.9353	1.4822	1.98	0.004
FOSL2	GO:0006468	protein phosphorylation	357	5660	2.2754	1.1532	1.97	0.010
E2F4	GO:0006281	DNA repair	226	4825	6.2845	3.4372	1.83	0.032
C-MYC	GO:0007275	multicellular organismal development	722	10405	1.2185	0.6729	1.81	0.009
GATA2	GO:0043687	post-translational protein modification	546	26572	1.5086	0.8419	1.79	0.005
USF2	GO:0001525	angiogenesis	865	25709	1.9768	1.1056	1.79	0.002
C-JUN	GO:0030036	actin cytoskeleton organization	517	27885	1.6484	0.9367	1.76	0.009
E2F4	GO:0010467	gene expression	420	4631	5.7991	3.3603	1.73	0.014
JUND	GO:0006486	protein glycosylation	502	44349	0.8922	0.5223	1.71	0.006
C-JUN	GO:0019221	cytokine-mediated signaling pathway	640	27762	1.5962	0.9349	1.71	0.006
MAX	GO:0001822	kidney development	623	38717	1.7315	1.0181	1.70	0.034
ZNF274	GO:0007165	signal transduction	114	1412	2.0867	1.2277	1.70	0.033
JUND	GO:0019221	cytokine-mediated signaling pathway	935	43916	0.8798	0.5189	1.70	0.001
MAX	GO:0030182	neuron differentiation	551	38789	1.7158	1.0197	1.68	0.040
ZNF263	GO:0000082	G1/S transition of mitotic cell cycle	631	47274	2.8763	1.7126	1.68	0.004
GATA2	GO:0044267	cellular protein metabolic process	889	26229	1.3907	0.8372	1.66	0.003
EBF1	GO:0007156	homophilic cell adhesion	580	42267	1.9804	1.1931	1.66	0.005
NRSF	GO:0055085	transmembrane transport	417	8083	1.9157	1.1665	1.64	0.019
ETS1	GO:0044267	cellular protein metabolic process	292	5264	2.5391	1.5579	1.63	0.028
IRF4	GO:0007186	G-protein coupled receptor signaling pathway	309	6337	0.9508	0.5843	1.63	0.022
GATA2	GO:0006412	translation	522	26596	1.3594	0.8455	1.61	0.030
C-MYC	GO:0006468	protein phosphorylation	627	10500	1.1008	0.6852	1.61	0.038
FOXA2	GO:0055085	transmembrane transport	398	7286	0.6874	0.4295	1.60	0.025

[a] Number of TFBSs assigned to GO category.

[b] Number of TFBSs not assigned to GO category.

[c] Estimate for TFBSs assigned to GO category (per kbp).

[d] Estimate for TFBSs not assigned to GO category (per kbp).

[e] Ratio of $\mathbb{E}[W]_g$ to $\mathbb{E}[W]_{\bar{g}}$. Only TF $\times$ GO category combinations with $\mathbb{E}[W]_g/\mathbb{E}[W]_{\bar{g}} > 1.6$ are shown.

[f] One-sided p-value for the null hypothesis that $\mathbb{E}[W]_g \leq \mathbb{E}[W]_{\bar{g}}$ indicating statistical significance of elevations in $\mathbb{E}[W]_g$ (see Section 8.3). Only cases with $p < 0.05$ are shown. These are nominal p-values not corrected for multiple comparisons and are meant only as a guide for follow-up study.

Supplementary Table 10: Average selection coefficients of segregating deleterious mutations

Scenario[a]	poly(WN)[b]	poly(SN)[c]	derived(WN)[d]	derived(SN)[e]	$2N_e\bar{s}$ [f]
baseline	101	21	461	24	−14.5
WN=−0.5	196	10	4337	11	−0.8
WN=−2	163	17	2561	23	−2.9
WN=−5	118	20	848	25	−7.7
WN=−20	68	20	139	27	−33.0
SN=−20	100	32	393	88	−11.8
SN=−50	98	32	258	52	−16.7
SN=−200	100	9	326	9	−15.1
SN=−300	102	8	538	8	−14.2
SP=5	89	13	317	15	−14.1
SP=20	103	20	380	21	−14.7
SP=50	106	19	408	21	−14.4
SP=100	102	18	510	21	−13.6

[a] Simulation scenario used to estimate average selection coefficient. Baseline scenario involved $2N_es = -100$ for strong negative selection (SN), $2N_es = -10$ for weak negative selection (WN), and $2N_es = 10$ for strong positive selection (SP). Other scenarios reflect a change of a single population-scaled selection coefficient to the value indicated, while other values are held constant. In all cases, the mixture proportions were 20% neutral, 25% SN, 50% WN, and 5% SP. Realistic demographic parameters for human populations were assumed, as described in Section 2. Additional details are provided in Gronau et al. (2013).

[b] Observed number of polymorphic sites simulated under WN selection.

[c] Observed number of polymorphic sites simulated under SN selection.

[d] Observed number of alleles under WN selection, obtained by weighting each site by the number of derived alleles.

[e] Observed number of alleles under SN selection, obtained by weighting each site by the number of derived alleles.

[f] Average population-scaled selection coefficient for all segregating deleterious alleles. This number is a weighted average of the assumed WN and SN selection coefficients, with weights determined by the derived(WN) and derived(SN) counts.

Nature Genetics: doi:10.1038/ng.2658

Supplementary Note

1 INSIGHT model

The full mathematical details of the INSIGHT model and inference procedure are presented in a separate manuscript (Gronau et al., 2013), but we briefly describe the key features of the approach here.

1.1 The probabilistic model

The model has three components: a phylogenetic model of outgroup genomes (gray in Supplementary Fig. 1), a model of divergence along the branch of the phylogeny leading to the ancestor of modern humans (green), and a model of polymorphism in human populations (brown). The phylogenetic component of the model determines a prior distribution for the "deep" ancestral allele Z_{ij} (i.e., the allele at the most recent common ancestor of humans and the closest outgroup, chimpanzee, at site j of locus i) conditional on the bases observed in all outgroups. The divergence portion determines the conditional distribution for the "shallow" ancestral allele A_{ij} (i.e., the allele at the most recent common ancestor of all human samples) given Z_{ij}. The polymorphism portion determines the conditional distribution of the human data X_{ij} given A_{ij}. An additional binary latent variable S_{ij} indicates whether each site is under selection ("sel") or neutrally evolving ("neut") and influences the conditional distributions for A_{ij} and X_{ij}.

The human data X_{ij} is summarized in the following way. Each nucleotide position is classified as monomorphic (M), polymorphic with a low-frequency minor allele (L), or polymorphic with a high-frequency minor allele (H). In addition, the major and minor alleles are recorded (the minor allele is undefined in the case of monomorphic sites; sites with more than two alleles are discarded from the analysis). Formally, $X_{ij} = (X_{ij}^{\text{maj}}, X_{ij}^{\text{min}}, Y_{ij})$ where X_{ij}^{maj} and X_{ij}^{min} are the major and minor alleles, respectively (both from the set $\{\text{A, C, G, T}\}$) and $Y_{ij} \in \{\text{M, L, H}\}$. The distinction between L and H depends on a predefined threshold for low-frequency alleles, f, which can be set by the user (for most of this paper, we have assumed $f = 0.15$, but see Supplementary Fig. 14). The model considers uncertainty in the ancestral allele A_{ij} and therefore must distinguish between minor alleles and derived alleles. If Y_{ij} is M or H, then the derived allele can be assumed to belong to the same category, but if Y_{ij} is L, then the derived allele frequency can be low or high, depending on the value of A_{ij}. Our inference procedure integrates over these two cases.

The model is hierarchical, with a collection of global parameters (ρ, η, γ, β_1, β_2, β_3; see Table 1, Online Methods) and a collection of locus-specific parameters (θ_i, λ_i). The likelihood function assumes independence of loci and conditional independence of nucleotides within loci given the locus-specific parameters. The graphical model shown in Supplementary Fig. 1 applies at all sites in binding sites and flanks but S is fixed at the "neut" value in the flanking regions. The likelihood function can be written:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\zeta}\,; \mathbf{X}, \mathbf{O}) \equiv P(\mathbf{X} \mid \mathbf{O}, \boldsymbol{\zeta}) = \\ \prod_{i \in B} \left[\prod_{j \in F_i} \sum_z \sum_a P(X_{ij}, Z_{ij} = z, A_{ij} = a \mid S_{ij} = \text{neut}, O_{ij}, \boldsymbol{\zeta}) \right] \\ \times \left[\prod_{j \in E_i} \sum_{s \in \{\text{neut,sel}\}} P(S_{ij} = s \mid \boldsymbol{\zeta}) \sum_z \sum_a P(X_{ij}, Z_{ij} = z, A_{ij} = a \mid S_{ij} = s, O_{ij}, \boldsymbol{\zeta}) \right] \end{aligned} \quad (1)$$

where B is the set of loci being analyzed, $\boldsymbol{\zeta}$ is the set of all free parameters ($\boldsymbol{\zeta} = \{\rho, \eta, \gamma, \beta_1, \beta_2, \beta_3\} \cup \{\theta_i | i \in B\} \cup \{\lambda_i | i \in B\}$), E_i is the set of TFBS nucleotide positions in locus i, and F_i is the set of neutral flanking positions for locus i. We use $\mathbf{X}$ and $\mathbf{O}$ to denote the entire human and outgroup data sets, respectively.

According to the conditional independence assumptions of the graphical model (Supplementary Fig. 1),

Nature Genetics: doi:10.1038/ng.2658

each expression of the form $P(X_{ij}, Z_{ij}, A_{ij} \mid S_{ij}, O_{ij}, \boldsymbol{\zeta})$ can be further expressed as a product, as follows:

$$\begin{aligned} P(X_{ij},&Z_{ij}, A_{ij} \mid S_{ij}, O_{ij}, \boldsymbol{\zeta}) = \\ &P(Z_{ij} \mid O_{ij}, \boldsymbol{\zeta})\, P(A_{ij} \mid Z_{ij}, S_{ij}, \boldsymbol{\zeta})\, P(X_{ij} \mid Z_{ij}, A_{ij}, S_{ij}, \boldsymbol{\zeta}) \end{aligned} \tag{2}$$

The probability of selection is determined by a two-component mixture model, with coefficient ρ:

$$P(S_{ij} = s \mid \boldsymbol{\zeta}) = \begin{cases} \rho & s = \text{sel} \\ 1 - \rho & s = \text{neut} \end{cases} \tag{3}$$

The conditional distribution for Z_{ij} is determined by an ordinary statistical phylogenetic model. The conditional distributions for A_{ij} and X_{ij} capture our main modeling assumptions (as described in the section of the main paper entitled "Probabilistic Model") and make it possible to extract useful information about selection from the data. Closed form expressions can be derived for these distributions by assuming a Jukes-Cantor (1969) substitution model, and assuming an infinite sites model for human polymorphisms:

$$P(A_{ij} = a \mid S_{ij} = s, Z = z, \boldsymbol{\zeta}) = \begin{cases} \frac{1}{3}\lambda_i t & s = \text{neut},\ a \neq z \\ 1 - \lambda_i t & s = \text{neut},\ a = z \\ \frac{1}{3}\eta\lambda_i t & s = \text{sel},\ a \neq z \\ 1 - \eta\lambda_i t & s = \text{sel},\ a = z \end{cases} \tag{4}$$

$$P\left(X_{ij} = (x^{maj}, x^{min}, y) \middle| \, S_{ij} = s, Z_{ij} = z, A_{ij} = a, \boldsymbol{\zeta}\right) = \begin{cases} 1 - \theta_i a_n & s = \text{neut}, y = \text{M},\ a = x^{maj} \\ \frac{1}{3}\beta_1\theta_i a_n & s = \text{neut}, y = \text{L},\ a = x^{maj} \\ \frac{1}{3}\beta_3\theta_i a_n & s = \text{neut}, y = \text{L},\ a = x^{min} \\ \frac{1}{3}\beta_2\theta_i a_n & s = \text{neut}, y = \text{H},\ a \in \{x^{maj}, x^{min}\} \\ 1 - \gamma\theta_i a_n & s = \text{sel}, y = \text{M},\ z = a = x^{maj} \\ 1 & s = \text{sel}, y = \text{M},\ z \neq a = x^{maj} \\ \frac{1}{3}\gamma\theta_i a_n & s = \text{sel}, y = \text{L},\ z = a = x^{maj} \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where t is the neutral branch length to human, and $a_n = \sum_{k=1}^{n-1} \frac{1}{k}$ is Watterson's constant for n samples. See Gronau et al. (2013) for a complete derivation of these expressions.

1.2 The inference algorithm

All parameters can be estimated by a fairly straightforward expectation maximization (EM) algorithm, as detailed by Gronau et al. (2013). Because the flanking sites vastly outnumber sites in TFBSs, their contribution dominates the portion of the likelihood that is a function of the neutral parameters. We can thus obtain a good approximation of the MLEs of β_1, β_2, β_3 and the locus-specific neutral parameters θ_i and λ_i by pre-estimating them from the neutral flanks alone. In addition, the phylogenetic model can be pre-estimated using existing software (RPHAST; Hubisz et al., 2011). This allows us to run the EM algorithm on a collection of TFBS sites only, conditioning on pre-estimated neutral and phylogenetic parameters. This pre-estimation process is further streamlined by defining a single fixed collection of 20 kbp blocks, overlapping by 10 kbp, and subsequently associating each TFBS with the neutral block whose center is nearest to it. As a result, the pre-estimation of the neutral parameters can be performed once and re-used for all collections of binding sites. After the pre-estimation steps are complete and the necessary sufficient statistics have been extracted from the data, INSIGHT can estimate the remaining free parameters (ρ, η, and γ) from a typical collection of binding sites in minutes.

Nature Genetics: doi:10.1038/ng.2658

1.3 Posterior expected values

Once MLEs of all parameters are obtained, it is possible to compute posterior expected values for certain measures of interest. For instance, the expected number of adaptive substitutions in a given binding site, i, can be obtained by the following sum of posterior probabilities:

$$\mathbb{E}[A]_i \;=\; \sum_{j:Y_{ij}=\mathrm{M}} \sum_{z:z\neq X_{ij}^{\mathrm{maj}}} P(Z_{ij}=z, S_{ij}=\mathrm{sel} \,|\, X_{ij}, O_{ij}, \boldsymbol{\zeta}) \;, \tag{6}$$

where the first sum is over all (unfiltered) monomorphic nucleotides within binding site i, the second sum is over all possible ancestral alleles that do not match the observed human allele in that position (X_{ij}^{maj}), and $P(Z_{ij}=z, S_{ij}=\mathrm{sel} \,|\, X_{ij}, O_{ij}, \boldsymbol{\zeta})$ is the joint conditional probability, given the data and model parameters, that site (i,j) is under (positive) selection and has experienced divergence from a particular ancestral allele (z).

Similarly, the expected number of weakly deleterious polymorphisms in binding site i can be obtained by the following sum of posterior probabilities:

$$\mathbb{E}[W]_i \;=\; \sum_{j:Y_{ij}=\mathrm{L}} P(Z_{ij}=X_{ij}^{\mathrm{maj}}, S_{ij}=\mathrm{sel} \,|\, X_{ij}, O_{ij}, \boldsymbol{\zeta}) \;, \tag{7}$$

where $P(Z_{ij}=X_{ij}^{\mathrm{maj}}, S_{ij}=\mathrm{sel} \,|\, X_{ij}, O_{ij}, \boldsymbol{\zeta})$ is the joint conditional probability, given the data and model parameters, that polymorphic site j in binding site i is under selection and the deep ancestral allele equals the observed major allele. Note that, according to our modeling assumptions, the fact that a site is under selection and is observed to have a low-frequency minor allele ($Y_{ij}=\mathrm{L}$) implies that (1) it is under weak negative selection, (2) it has not experienced a divergence ($Z_{ij}=A_{ij}$), and (3) the minor allele is the derived allele ($A_{ij}=X_{ij}^{\mathrm{maj}}$) (see equation 5).

1.4 Variance estimates

Once MLEs of all parameters are obtained, approximate variances can be derived for the selection parameters (ρ, η, and γ) from an estimated Fisher information matrix. Specifically, we compute the 3×3 matrix of second derivatives (Hessian) of the log likelihood function for ρ, η, and γ at the maximum and negate it to estimate the Fisher information matrix. We then invert this matrix to estimate the variance/covariance matrix for all parameters, and we report square roots of elements on the diagonal as approximate standard errors for the three parameters.

To propagate variances in parameter estimates through to calculations of the posterior expected number of adaptive substitutions, we use the approximation $\mathbb{E}[A] \approx \rho\eta\bar{\lambda}t$, where $\bar{\lambda}$ is a weighted average of all λ_i values. Using a first-order Taylor approximation, we then estimate the variance of $\mathbb{E}[A]$ to be:

$$\mathrm{Var}[\mathbb{E}[A]] \approx (\eta\bar{\lambda}t)^2\mathrm{Var}[\rho] + (\rho\bar{\lambda}t)^2\mathrm{Var}[\eta] + 2(\eta\bar{\lambda}t)(\rho\bar{\lambda}t)\mathrm{Cov}[\rho,\eta]. \tag{8}$$

In this case, the necessary covariance term is also extracted from the inverted Fisher information matrix, and the variance in the pre-estimated λ_i values is ignored. The variance of $\mathbb{E}[W]$ can be approximated by a parallel, but slightly more complex, calculation as detailed in Gronau et al. (2013).

2 Simulation study

2.1 Simulation design

We tested the ability of INSIGHT to recover "true" evolutionary parameters under two simulation scenarios. The first scenario involved a single human population and three nonhuman primate populations: rhesus

Nature Genetics: doi:10.1038/ng.2658

macaque, orangutan, and chimpanzee (as with our real data). The effective population size was held constant at $N_e = 20{,}000$ in the ancestral and outgroup lineages, and at $N_e = 10{,}000$ on the human lineage. The rhesus macaque, orangutan, and chimpanzee outgroup populations were assumed to diverge from a lineage leading to modern humans at 30, 18, and 6.5 million years ago, respectively. The data were simulated as a collection of loci, each of which consisted of a TFBS of size 10 and 5,000 neutral bases on each side. The loci were assumed to be independent (i.e., in complete linkage equilibrium) but recombination was allowed within loci, at a constant rate of $r = 1.1 \times 10^{-8}$ recombinations per nucleotide position per generation (Kong et al., 2002). The mutation rate was assumed to have a mean value of $\mu = 1.8 \times 10^{-8}$ per nucleotide per generation (Sun et al., 2012), but was allowed to vary across loci. The specific rate at each locus was sampled from a normal distribution having this mean and a standard deviation equal to one tenth of the mean (Mouse Genome Sequencing Consortium, 2002).

The second scenario was similar to the first in most respects, but was designed to mimic our real data set as closely as possible with respect to human demographic history. In particular, we modeled distinct African, European, and East Asian populations, using the demographic parameters estimated by Gravel et al. (2011). This scenario starts, as above, with an ancestral primate population of 20,000 individuals, which decreases to 10,000 on the human lineage at the human/chimpanzee divergence; then, following Gravel et al. (2011), at 148 thousand years ago (kya) the human population increases to 14,474; at 51 kya, an ancestral Eurasian population diverges from the main African lineage with a population size of 1,861, while the African size remains constant at 14,474; and finally, at 23 kya the Eurasian lineage splits into European and East Asian populations and these populations immediately begin to grow exponentially, at rates of 152.0 and 192.0, respectively[1], and continue to grow at these rates until the present.

Under both scenarios, we assigned each nucleotide position in each simulated TFBS to one of four selective modes: positive selection ($2N_e s = 10$), strong negative selection ($2N_e s = -100$), weak negative selection ($2N_e s = -10$), and neutral evolution ($2N_e s = 0$). We considered four mixtures of these four modes, as shown in Supplementary Figs. 12 & 13. (Fig. 1 is similar but omits the simple case of only neutrality and strong negative selection.) The number of sites in each category at each TFBS was obtained by sampling from a multinomial distribution corresponding to the assumed mixture model.

Our strategy was to generate a data set roughly similar to our real data under each of these two demographic scenarios, with 10,000 loci, each of which is represented by 50 diploid human individuals (100 chromosomes) and one haploid genome from each of the three nonhuman primate outgroups. In the second scenario, the 50 human individuals were sampled in approximately the proportions observed in the Complete Genomics data: 15 Europeans, 10 East Asians, and 25 Africans.

2.2 Data generation

All simulated data was generated using SFS_CODE (Hernandez, 2008). To express times in units of $2N_e$ generations, as required by SFS_CODE, we divided times in years (reported above) by $2N_e g$, where N_e is the ancestral effective population size (20,000 in all cases) and g is the generation time (assumed to be 25 years for human populations and 15 years for outgroups). To save in computational cost, we performed forward simulations starting with $N_{\text{sim}} = 2{,}000$ individuals, and reducing to 1,000 individuals on the human lineage. The use of $N_{\text{sim}} < N_e$ should have little effect on results, provided N_{sim} is sufficiently large to limit sampling error, because all parameters are expressed in population-scaled form (thus, the mutation and recombination rates are implicitly scaled upward to compensate for the use of a reduced population size). We used the default "burn-in" of $5 \times 2N_{\text{sim}} = 20{,}000$ generations before initiating the specified demographic scenario. Each 10,010 bp locus was modeled in SFS_CODE as two neutral subloci of 5,000 bp each, and a sequence of 1–4 subloci with strong negative, weak negative, or positive selective modes, with lengths

[1] Exponential growth rates are expressed as $\log(N_e^{\text{final}}/N_e^{\text{initial}})$/time, where time is in units of $2N_e$ generations.

Nature Genetics: doi:10.1038/ng.2658

sampled from a multinomial distribution and a total length of 10 bp. All of these loci were contiguous but subject to recombination. At the end of each simulation, representatives of each population were stored for analysis, as described above.

Positive selection requires some special handling because we are primarily interested in modeling a scenario in which an element is initially under constraint but then comes under positive selection in one lineage, perhaps due to changing environmental conditions, and then stabilizes again once an advantageous allele is driven to fixation. The default behavior in SFS_CODE is instead to continually treat a position defined as being under positive selection as if it has a suboptimal allele, which tends to produce repeated fixation events. To address this problem, positively selected nucleotides are defined to be under weak negative selection ($2N_e s = -10$) in ancestral and outgroup populations, but then switch to positive selection ($2N_e s = 10$) on the human lineage immediately after the human/chimpanzee divergence. At a point in fairly recent human history, but prior to any human population size changes (at 300 kya), the site reverts to weak negative selection. This strategy ensures that positively selected sites have the opportunity of undergoing a selective sweep (although they are not guaranteed to have one), but mostly eliminates the signature of recurrent positive selection (in particular, an enrichment for high-frequency derived alleles) in present day populations.

An example command line for a simulation that uses the 25% neutral, 5% positive, 40% strong negative, 30% weak negative selection regime follows. In this case the calling code sampled a value of $\theta = 0.0015036231$ and numbers of sites evolving neutrally, under positive selection, under strong negative selection, and under weak negative selection of 1, 1, 3, and 5, respectively.

```
sfs_code 7 1 -n 50 -N 2000 -TE 50.0 -I –theta 0.0015036231 –rho 0.00088 \
    -L 6 5000 1 1 3 5 5000 -a N -W L 0 0 -W L 5 0 -W L 1 0 -W L 2 1 10 0 1 -TW \
    39.3666666667 P 3 L 2 1 10 1 0 -TW 45.3666666667 P 3 L 2 1 10 0 1 -TW \
    45.3666666667 P 4 L 2 1 10 0 1 -TW 45.3666666667 P 5 L 2 1 10 0 1 -W L 3 1 \
    100 0 1 -W L 4 1 10 0 1 -TS 0 0 1 -TS 20.0 1 2 -TS 39.1666666667 2 3 -Td \
    39.1666666667 P 3 0.5 -TE 45.6666666667 3 -TE 45.6666666667 4 -TE \
    45.6666666667 5 -TE 45.6666666667 6 -TS 45.5 3 4 -Td 45.5186666667 P 4 \
    1.4474 -TS 45.6156666667 4 5 -Td 45.6156666667 P 5 0.12857537654 -TS \
    45.6436666667 5 6 -Td 45.6436666667 P 5 0.554540569586 -Td 45.6436666667 P \
    6 0.297689414293 -Tg 45.6436666667 P 5 152.0 -Tg 45.6436666667 P 6 192.0
```

To mimic our handling of the real data (see below), triallelic sites were discarded (they are very rare). The full model (with probabilistic treatment of ancestral alleles) was applied to the simulated data.

2.3 "True" values of estimated quantities

The true fractions of sites under selection for each simulation scenario were simply defined as the non-neutral fractions (i.e., 70% in scenarios one and two, 75% in scenarios three and four). The true number of adaptive substitutions (A) was taken as the number of substitutions in nucleotide positions designated as being under positive selection. Fig. 1 & Supplementary Fig. 12 report this number divided by the number of kilobases analyzed (10 in these experiments). The true value of α was obtained by dividing A by the total number of substitutions on the human lineage.

The true number of weakly deleterious polymorphisms was taken as the number of polymorphic sites designated as being under selection. Notice that this number is nonzero in scenarios having no weak negative selection. The reason is that residual low-frequency polymorphisms in strong negative or positively selected sites are included in this category, as they will be in our inference procedure. In a sense, the "weak selection" category in the model is operationally defined to include all sites that are under selection but polymorphic.

Nature Genetics: doi:10.1038/ng.2658

Our simulations indicate that, in the absence of weak negative selection, strong negative selection will account for the majority of such sites (compare the second and fourth scenarios in Supplementary Fig. 12). Notably, estimates of ρ, $\mathbb{E}[A]$/kb, and α remain essentially unbiased despite that the model lumps polymorphic strongly selected sites together with weakly selected sites.

2.4 Simple estimators

As a comparison point for the model-based estimates from INSIGHT, we made use of estimators for the fraction of sites under selection (ρ), the number of adaptive substitutions ($\mathbb{E}[A]$), and the fraction of substitutions driven by positive selection (α) based on simple counts of nucleotide substitutions and/or polymorphic sites. As a divergence-based estimator for ρ, we used a quantity introduced by Kondrashov and Crow (1993):

$$\hat{\rho}_{\text{Div}} = 1 - \frac{f_{\text{TFBS}}}{f_{\text{N}}} \tag{9}$$

where f_{TFBS} is the rate at which substitutions occur in binding sites and f_{N} is the neutral substitution rate. We simply estimated f_{TFBS} as $D_{\text{TFBS}}/L_{\text{TFBS}}$, where D_{TFBS} is the total number of substitutions and L_{TFBS} is the total number of nucleotides, both in TFBSs. Similarly, we estimated f_{N} as $D_{\text{N}}/L_{\text{N}}$, where D_{N} and L_{TFBS} are the numbers of substitutions and nucleotide sites in neutral flanking regions, respectively. Thus,

$$\hat{\rho}_{\text{Div}} = 1 - \frac{D_{\text{TFBS}} L_{\text{N}}}{D_{\text{N}} L_{\text{TFBS}}} \tag{10}$$

This estimator ignores the phylogeny, recurrent mutations, and unequal rates of substitution (e.g., between transitions and transversions), but on the evolutionary time scale of human-chimpanzee divergence it performs nearly identically to full statistical phylogenetic methods such as phyloP (Pollard et al., 2010). The main differences with respect to INSIGHT are that $\hat{\rho}_{\text{Div}}$ ignores the effect of positive selection on divergence and it pools counts across binding sites in a manner that does not account for variable mutation rates across the genome.

As a simple polymorphism-based estimator we used an analogous quantity:

$$\hat{\rho}_{\text{Poly}} = 1 - \frac{P_{\text{TFBS}} L_{\text{N}}}{P_{\text{N}} L_{\text{TFBS}}} \tag{11}$$

where P_x is the number of polymorphisms in nucleotide sites of class x (TFBS or neutral flanks). The main differences with respect to INSIGHT are that $\hat{\rho}_{\text{Poly}}$ assumes selected sites cannot be polymorphic (hence it handles weak selection poorly) and, like $\hat{\rho}_{\text{Div}}$, it naively pools counts across loci.

As a simple estimator for α, we used an extension of the McDonald-Kreitman test due to Smith and Eyre-Walker (2002), again slightly adapted for our purposes:

$$\hat{\alpha}_{\text{MK}} = 1 - \frac{D_{\text{N}} P_{\text{TFBS}}}{D_{\text{TFBS}} P_{\text{N}}} \tag{12}$$

where D_x and P_x are the number of divergences and polymorphisms, respectively, in nucleotide sites of class x. (Notice that the L_x terms cancel in this equation and are not needed.) To estimate $\mathbb{E}[A]$, we multiplied all terms by D_{TFBS} (Smith and Eyre-Walker, 2002),

$$\mathbb{E}[\hat{A}]_{\text{MK}} = D_{\text{TFBS}} - \frac{D_{\text{N}} P_{\text{TFBS}}}{P_{\text{N}}} \tag{13}$$

Like $\hat{\rho}_{\text{Poly}}$, these estimators implicitly assume no polymorphisms occur in selected sites, and like both simple estimators for ρ, they naively pool counts across loci.

Nature Genetics: doi:10.1038/ng.2658

3 Pipeline for TFBS identification

Our TFBS identification pipeline involved three main steps: (1) data collection and preparation; (2) motif discovery; and (3) binding site identification. We describe each of these steps below.

3.1 Data collection and preparation

Our pipeline was based on chromatin immunoprecipitation and sequencing (ChIP-seq) data from the ENCODE project, which is publicly available via the UCSC Genome Browser ("ENCODE Transcription Factor Binding Tracks," hg19 assembly). We focused on the data sets provided by the Hudson Alpha Institute for Biotechnology (HAIB) and the Stanford / Yale / USC (previously, UC Davis) / Harvard (SYDH) consortium, which account for the majority of the available cell-type $\times$ transcription factor (TF) combinations. We excluded time-course experiments or cases in which cell types were chemically treated (e.g., by interferon gamma), control experiments, and any data sets with usage restriction release dates after June, 2012. We downloaded the genomic coordinates for the ENCODE ChIP-seq peak calls for the remaining data sets. In the end, we considered data for 122 TFs from a total of 32 different cell types, with an average of 2.6 cell types per TF. At this stage, data from the two providers (HAIB and SYDH) were kept separate, despite that some TFs were considered by both. We later merged these results (see below). We considered peaks on the autosomes only (chromosomes 1–22).

3.2 Motif discovery

Next we used the MEME program (Bailey and Elkan, 1994) for motif discovery in ChIP-seq identified regions, separately for each TF and for each of the two data providers. The goal of this step was to obtain a high-quality motif for each factor, making use of the abundant ChIP-seq data. We chose not to rely completely on the available motif databases, because of their uneven quality across TFs. However, we did cross-check our motifs with databases where available, as discussed below.

To reduce computational cost, we subsampled the ChIP-seq peaks, performing motif discovery for each TF with four sets of 1000 randomly sampled peaks. In addition, because the reported peaks tended to be quite broad, we truncated them to lengths of 100 bp, either centered on the designated peak position (SYDH) or simply by taking the central 100 nucleotides (HAIB). Note that the elimination of some true binding sites is not a problem in this analysis, provided that the remaining sequences are strongly enriched for binding sites. This strategy was inspired by MEME-ChIP (Machanick and Bailey, 2011). For each set of 1000 truncated peaks, we extracted the corresponding sequences from the reference genome and stored them in a FASTA file for analysis with MEME. MEME was run both on these four original FASTA files and on a second set of four files that had been processed with RepeatMasker, for a total of eight runs per TF. In each case, the program was configured to search for the three best scoring motifs (-nmotifs 3) between 6 and 25 bp long (-minw 6 -maxw 25), to consider both the forward and reverse strands (-revcomp), and to allow zero or one occurrence of the motif per sequence (-mod zoops).

The motifs returned by MEME for each TF were then manually inspected and compared with the JASPAR_2009 and UNIPROBE databases. No high-quality consensus motif could be identified for 28 of the TFs, and these were excluded from further analysis (E2F6, HDAC2, HEY1, PAX5C20, PAX5N19, Pol2, Pol2-4H8, RXRA, SIN3AK20, TAF1, TAF7, CCNT2, CtBP2, GTF2B, HA-E2F1, HMGN3, Ini1, KAP1, Mxi1, P300, Pol2, SETDB1, SMC3, SPT20, STAT3, TBP, ZNF217, ZZZ3). Reasons for discarding TFs included high variability in motif inferences across runs of the program, inconsistencies with previously annotated motifs, substantial differences between SYDH and HAIB data sets for the same TF, or inference of motifs with poor sequence specificity or low complexity (e.g., long mono- or di-nucleotide repeats). Not surprisingly, several of the discarded TFs are general DNA-binding proteins of various kinds (e.g.,

Nature Genetics: doi:10.1038/ng.2658

GTF2B, P300, Pol2, SMC3) rather than true sequence-specific transcription factors. This left 94 TFs with high-quality motifs. In each of these cases, we selected a single "best" motif by taking the highest-scoring case from the RepeatMasked set that either matched previously reported motifs or did not include simple sequence repeats.

3.3 Binding site identification

We then used the MAST program (also from the MEME Suite) and the selected motif for each TF to search the full set of peak regions for corresponding binding sites. Only hits with a p-value < 0.0001 and E-value < 10, both corrected for sequence composition (-comp), were retained. In this case, we allowed for multiple occurrences of the motif on both strands. This procedure was applied separately to the peaks for each cell type and, in the case of HAIB, each of two experimental replicates. First, results for HAIB replicates were combined by taking a strict intersection: predicted TFBSs were retained only if they were supported by peaks from both replicates (this reduced the number of binding sites by about 50%). The SYDH data also reflects multiple replicates but they were combined by the data providers in preprocessing. Next, a single set of TFBSs was identified for each TF by taking the union of the TFBSs for all cell types. At this stage, we also took the union of the HAIB and SYDH TFBSs for nine TFs that were examined by both providers (ATF3, YY1, CTCF, EBF1, BHLHE40, JUND, MAX, GATA2, and RAD21). When merging sets of TFBSs, the cell types and data providers supporting each binding site were recorded. After merging, we were left with a final set of 84 TFs with annotated binding sites across the genome. Six of these TFs (ZBTB33, THAP1, BDP1, GCN5, E2F1, and RPC155) had fewer than 500 TFBSs and these generally contained too few divergences or polymorphisms for reliable parameter inference, so they were also excluded, leaving 78 TFs. As a final step, degenerate positions were eliminated from the edges of all motifs by trimming away positions with information content <0.5. The corresponding positions were also eliminated from the edges of the TFBSs.

4 Genome sequence data and filters

Our source of human variation data was the "69 Genomes" data set released by Complete Genomics (CG) in 2011 (http://www.completegenomics.com/public-data/69-Genomes/). We considered 54 unrelated individuals from this set, excluding the child in each trio (YRI and PUR) and all but the four grandparents in the 17-member CEPH pedigree. For each individual considered, we recorded the diploid genotype call reported for each position in the hg19 (Genome Reference Consortium Human Build 37) reference genome using CG's 'masterVar' files[2]. We considered both "SNPs" and "length-preserving substitutions" in the masterVar file. We also recorded the positions at which CG could not confidently assign a genotype for subsequent masking (see below). All other positions were assumed to be homozygous for the allele reported in the reference genome. For divergence data, we used alignments from the UCSC Genome Browser of the human reference genome (hg19) with the chimpanzee (panTro2), orangutan (ponAbe2), and rhesus macaque (rheMac2) genomes. For each position in hg19, we recorded the aligned base from each of the three nonhuman primates, or an indication that no alignment was available at that position.

Borrowing from our previous work on demography inference (Gronau et al., 2011), we applied several filters to these data to reduce the impact of technical errors from alignment, sequencing, genotype inference, and genome assembly. We eliminated simple repeats, recent transposable elements, recombination hotspots

[2]The masterVar files are included in tar files available from ftp://ftp2.completegenomics.com. The tar files currently have URLs of the form ftp://ftp2.completegenomics.com/$GROUP/ASM_Build37_2.0.0/$SAMPLE-200-37-ASM-VAR-files.tar, where $GROUP is one of 'Diversity', 'Pedigree_1463', 'YRI_trio', 'PUR_trio' and $SAMPLE is the sample name. The enclosed masterVar files can be identified by names of the form masterVarBeta-$NAME-200-37-ASM.tsv.bz2.

Nature Genetics: doi:10.1038/ng.2658

(The 1000 Genomes Project Consortium, 2010), and also excluded all position pairs having a "CG" dinucleotide in any of the human samples or the outgroups, to avoid biases from hypermutable CpGs. As a further caution, we excluded position pairs with CX in an outgroup and YG in human, to avoid potential ancestral CpGs. We also excluded recent segmental duplications and regions in each outgroup genome not showing conserved synteny with the human reference sequence, also as described by Gronau et al. (2011). Finally, we excluded any additional regions found in the "black list" filter reported in Dunham et al. (2012).

When applying INSIGHT to the data we excludeed any site (in TFBS or neutral flanks) that had been eliminated by a filter or was otherwise missing in any of the human genomes. These filters excluded roughly 20% of nucleotide sites in the genome. For the outgroup sequences, we recorded an "N" at any site in an alignment gap, where confident synteny could not be established, or where the base had a phred quality score $<$20.

In addition, we applied further filters to the flanking regions of each TFBS (20,000 bp window surrounding the binding site) to eliminate sites likely to be under selection. As described by Gronau et al. (2011), we eliminated exons of annotated protein-coding genes and 1000 bp of flanking sequence and conserved noncoding elements (identified by phastCons) and 100 bp of flanking sequence. We also eliminated RNA genes (and 1000 bp flanks) included in the GENCODE set (version 11), and all transcription factor binding sites identified by our pipeline or in the subset of bound "known motifs" from Dunham et al. (2012).

Finally, we considered the autosomes only (chromosomes 1–22) in our analysis.

5 Application of INSIGHT to real TFBSs

5.1 Analysis setup

For our main analysis, we grouped all binding sites by TF, and ran INSIGHT separately on each of the 78 sets of TFBSs. Other analyses involved alternative partitions of the collection of all TFBSs (e.g., Section 8.3).

The human variation data was summarized by recording, for each position, the observed alleles and their counts. Sites with more than two observed alleles or with missing data in one of the human individuals were masked out by our filters (see Section 4). Nucleotide positions with a single observed allele across the 2×54=108 sampled chromosomes were labeled as monomorphic (M). Polymorphic sites were labeled "L" if the frequency of their minor allele was below the threshold for low-frequency alleles, f, and labeled "H" otherwise. We performed this preprocessing step for thresholds f of 10%, 15%, and 20% but used 15% for all analyses reported in the main text. Our results were fairly insensitive to the choice of low-frequency threshold (e.g., Supplementary Fig. 14).

Finally, to avoid the influence of sparse data in neutral flanking sites, we excluded any binding sites with fewer than 100 nucleotides of flanking sequence after filters had been applied. This filter eliminated about 3% of all TFBSs. Nevertheless, sparse data (particularly an absence of high-frequency polymorphisms) occasionally caused parameter estimates to go to extreme values (e.g., $\hat{\rho} = 1$) due to overfitting. To address this problem, we added a "pseudo-locus" to each set, consisting of a single high-frequency polymorphic position associated with a neutral polymorphism rate (θ_i) of 0.001 and a neutral divergence rate (λ_i) of 0.001. We settled on this strategy after some experimentation and found that it eliminated the overfitting problem but had almost no effect on the parameter estimates when adequate data was available (as in most cases).

5.2 Confidence intervals and posterior expected values

Error bars displayed in figures correspond to one standard error above and below the maximum likelihood estimates of the parameters. Standard errors were computed for ρ, $\mathbb{E}[A]$, and $\mathbb{E}[W]$ associated with a given

Nature Genetics: doi:10.1038/ng.2658

TF using the method described in Section 1.4.

The posterior expected values $\mathbb{E}[A]$ and $\mathbb{E}[W]$ were computed for a collection of binding sites, B, by summing over the expected counts $\mathbb{E}[A]_i$ and $\mathbb{E}[W]_i$ across all binding sites $i \in B$ (see equations 6-7 in Section 1.3 for derivation of these expressions):

$$\mathbb{E}[A] = \sum_{i \in B} \mathbb{E}[A]_i = \sum_{i \in B} \sum_{j:Y_{ij}=\mathrm{M}} \sum_{z:z \neq X_{ij}^{\mathrm{maj}}} P(Z_{ij} = z, S_{ij} = \mathrm{sel} \,|\, X_{ij}, O_{ij}, \boldsymbol{\zeta}(i)) \,. \tag{14}$$

$$\mathbb{E}[W] = \sum_{i \in B} \mathbb{E}[W]_i = \sum_{i \in B} \sum_{j:Y_{ij}=\mathrm{L}} P(Z_{ij} = X_{ij}^{\mathrm{maj}}, S_{ij} = \mathrm{sel} \,|\, X_{ij}, O_{ij}, \boldsymbol{\zeta}(i)) \,. \tag{15}$$

Because binding sites in B might correspond to different TFs, the notation $\boldsymbol{\zeta}(i)$ is used here to represent the parameter values inferred for the TF corresponding to binding site i. To allow comparisons between sets of different sizes, these posterior expected values were typically normalized by dividing them by the number of unfiltered nucleotides (in kilobases).

The posterior expected number of weakly deleterious mutations per haploid genome, $\mathbb{E}[D]$, was computed in a similar manner to $\mathbb{E}[W]$, but in this case additionally considering the number of derived alleles for each low-frequency polymorphism:

$$\mathbb{E}[D] = \frac{1}{n} \sum_{i \in B} \sum_{j:Y_{ij}=\mathrm{L}} M_{ij} \, P(Z_{ij} = X_{ij}^{\mathrm{maj}}, S_{ij} = \mathrm{sel} \,|\, X_{ij}, O_{ij}, \boldsymbol{\zeta}(i)) \,, \tag{16}$$

where n is the number of haploid samples (108 in our study), and M_{ij} is the minor allele frequency at site (i, j) (generally a number between 1 and 16).

5.3 Likelihood ratio tests

In several cases, we tested for significance using likelihood ratio tests (LRTs) based on our model. We performed two types of tests. The first type (used for Fig. 2) compared a null hypothesis that a parameter of interest (ρ, η, or γ) was equal to zero with an alternative hypothesis that it had a value greater than zero. This type of test was accomplished by using INSIGHT to find the maximum value of the log likelihood function with all parameters free (as in typical applications), and then using the program again to find the maximum value of the log likelihood function with the parameter of interest fixed at zero and other parameters free. Twice the difference between these values was then treated as a test statistic, in the usual way. In the case of η and γ this is a fairly straightforward LRT with one degree of freedom, except that the null hypothesis is at the boundary of the alternative hypothesis (because these parameters are bounded by zero). Therefore, we assumed an asymptotic distribution equal to a 50:50 mixture of a χ^2 distribution with one degree of freedom and a point mass at zero in computing approximate p-values (Self and Liang, 1987). The case of ρ is more complex, because, in addition to the boundary issue above, a value of $\rho = 0$ also causes η and γ to become undefined. In this case, we used a χ^2 distribution with three degrees of freedom for p-value calculations, which was shown to have good fit in an empirical evaluation (Gronau et al., 2013). In assessing the significance of $\mathbb{E}[A] > 0$ and $\mathbb{E}[W] > 0$ (Fig. 2), or $\alpha > 0$ and $\tau > 0$ (Supplementary Fig. 2), we used the LRTs of η and γ, respectively, because these are the critical parameters for measuring positive and weak negative selection, respectively.

The second type of LRT (used in the GO analysis for ρ) tests for significant differences in the influence of selection on two collections of binding sites. In this case, the log likelihood for the alternative model was obtained by applying INSIGHT separately to each data set then summing the maximized log likelihoods, and the log likelihood for the null model was estimated by applying INSIGHT to a combined data set. We calculated p-values for twice the difference in these log likelihoods using a χ^2 distribution with three degrees

Nature Genetics: doi:10.1038/ng.2658

of freedom. Note that, while we expect differences in ρ to be dominant in this test—and we use it as a heuristic way to test for differences in ρ in the GO analysis—it can in principle be influenced by differences in any of the parameters of the model.

6 Application to protein-coding genes

Because a single alignment or annotation error can contribute a large number of apparent divergence events, it was important to use a carefully curated set of human protein-coding genes for our analysis. To maximize the number of high-confidence genes in our set, we began with the protein-coding genes in GENCODE release v13 (Harrow et al., 2012) and designed our own filters to eliminate annotations likely to be problematic in our analysis. First, genes with any of the following labels were excluded from the set: *pseudogene, alternative 3 UTR, alternative 5 UTR, cds end NF, cds start NF, downstream ATG, mRNA end NF, mRNA start NF, non canonical genome sequence error, non canonical TEC, not best in genome evidence,* or *not organism supported*. We additionally excluded genes having coding sequence (CDS) lengths of <100 bp. Finally, we excluded individual exons for which the chimpanzee, orangutan, or macaque alignments either corresponded to multiple "chains" in the UCSC alignments (indicating a lack of conserved synteny) or contained frame shifts that were not compensated within 15 bp (similar to Kosiol et al., 2008). After these filters were applied, we were left with a set of coding exons corresponding to 15,799 genes and including 26,318,878 genomic positions.

We then ran these exons through our pipeline, treating exons as being analogous to binding sites. The flanking neutral regions were drawn from 10,000 bp on either side of each exon, similar to TFBSs, and the same filters were applied as with TFBSs. We removed all exons for which fewer than 100 flanking neutral sites remained after filtering, leaving at least one exon from each of 15,864 genes (a total of 19,581,186 coding bases and 143,093 exons out of the original 166,394, remained).

We also considered six smaller gene sets of interest. These were all subsets of a heavily filtered gene set that our research group had previously prepared for a study of positive selection in mammals (Kosiol et al., 2008). First, we considered (1) the 202 genes found by Kosiol et al. (2008) to show at least moderate evidence of positive selection on the human lineage (Mam PS, $p < 0.05$), and (2) the 13,248 genes that did not show evidence of positive selection on this branch (Mam Other, $p \geq 0.05$). In addition, we considered (3) 161 genes inferred to be under positive selection in human populations by Bustamante et al. (2005) (Hum PS), and (4) 489 genes with significant evidence of strong negative selection in the same study (Hum NS). Finally, we considered (5) 436 genes identified as housekeeping genes by Eisenberg and Levanon (2003) based on constitutive expression in human tissues, and (6) 11,133 genes not included in the Hum PS, Hum NS, or housekeeping sets (Hum Other).

We then applied INSIGHT both to the full set of genes and to these smaller sets. However, we faced a challenge in obtaining estimates that would be comparable between TFBSs and CDSs because these two data sets are quite different in structure. Both data sets can be analyzed in bulk by simply fitting the model to the set of all TFBSs or the set of all CDS exons, but this global strategy has the weakness that it fails to accommodate variation across elements in the influence of natural selection. For example, if a subset of genes shows strong evidence of adaptation, but this subset is fairly small and negative selection dominates in other genes, estimates of $\mathbb{E}[A]$ and α may still be driven to zero in a global analysis. Our sensitivity for natural selection may be increased by partitioning the elements and fitting the model separately to each partition. On the other hand, an overly fine-grained partitioning of elements may result in sparse data and overfitting. The TFBSs are naturally partitioned by transcription factor, leading to 78 groups each with a fairly large (>500) number of elements (as described in our main analysis), but it is less obvious how to partition the CDS elements. We settled on two partitioning schemes at opposite extremes: a "selection class" approach, where the model is fitted separately to the coarse-grained categories of genes described

Nature Genetics: doi:10.1038/ng.2658

above (Mam PS and other, or Hum PS, NS, housekeeping, and other); and a "by-gene" approach, where the model is fit separately to each gene, treating the individual exons of that gene as separate elements (like the individual TFBSs for a given TF). Notice that the selection class approach may have somewhat reduced sensitivity to selection, because each set of genes is still likely to be quite heterogeneous, while the gene-by-gene approach is likely to result in some over-fitting. All in all, we use two model-fitting approaches for TFBSs (global and by TF) and three model-fitting approaches for CDSs (global, selection class-wise, and by gene).

Results for these various model-fitting strategies for TFBSs and CDSs are presented in Supplementary Table 3. Overall, the results echo those presented in Fig. 2. The estimates of ρ are generally larger in genes, regardless of the partitioning scheme. Positive selection is evident in TFBSs under all partitioning schemes, except for the global one. Positive selection is not evident in the global CDS analysis. In the selection class analysis, it is apparent only for the groups of genes previously identified as being under positive selection. The by-gene analysis (last two rows) yields a fairly strong signal of positive selection across all genes. This estimate is probably somewhat generous, being influenced to a degree by over-fitting on individual genes for which little data is available. Nevertheless, we have used it as our basis for comparison with TFBSs in Fig. 4, treating it as a rough upper limit for the true influence of adaptation in these regions.

7 Extrapolation of expected counts to the complete genome

To extrapolate our estimates for $\mathbb{E}[A]$, $\mathbb{E}[W]$, and $\mathbb{E}[D]$ for protein-coding genes to the complete genome, we simply multiply the numbers estimated for the analyzed subset of genes by the ratio A/B, where A is an estimate of the total number of coding bases in the genome (36,633,898, based on a union of the UCSC Genes and RefSeq tracks in the UCSC Genome Browser), and B is the total number of coding bases analyzed (19,581,186, after all filters have been applied). For the comparison with TFBSs, we extrapolate the expected counts for TFBSs to the complete genome by multiplying the estimated numbers by $2A/C$, where A is as above and B is the total number of TFBS bases analyzed (13,285,443 after all filters have been applied).

8 Correlates of selection

8.1 Information content and binding affinity

We consider a motif M to be defined by a position weight matrix (PWM) based on an underlying probability model. Formally, we assume binding sites for a given transcription factor T are characterized by an independent multinomial probability distribution over the four nucleotide bases at each of k positions, with base b ($b \in \{\text{A,C,G,T}\}$) occuring at position i ($1 \leq i \leq k$) with probability $p_b^{(i)}$ and $\sum_b p_b^{(i)} = 1$ for all i. In addition, we assume "background" (non-bound) sequences are characterized by an alternative multinomial distribution $\pi = (\pi_\text{A}, \pi_\text{C}, \pi_\text{G}, \pi_\text{T})$ such that $\sum_b \pi_b = 1$ (e.g., Wasserman and Sandelin, 2004).

A score indicating how likely a sequence of nucleotides $X = (x_1, \ldots, x_k)$ is to be bound by T is given by:

$$S(X) = \sum_i W_{x_i}^{(i)}, \quad \text{where } W_{x_i}^{(i)} = \log_2 \frac{p_{x_i}^{(i)}}{\pi_{x_i}}. \tag{17}$$

The $k \times 4$ possible values of $W_{x_i}^{(i)}$ can be summarized in a "position weight matrix."

In statistical terms, $S(X)$ is a log-odds score with desirable properties for distinguishing sequences generated by the motif model from background sequences. However, $S(X)$ also can be interpreted in biophysical terms as a measure of *binding affinity* (sometimes called "binding energy"), assuming the affinity

Nature Genetics: doi:10.1038/ng.2658

of a sequence is a sum of independent contributions from each nucleotide (Berg and von Hippel, 1987; Stormo, 2000). This is the measure of predicted binding affinity used in this paper.

By convention, the information content (IC) of the motif is defined to be (Schneider et al., 1986; Wasserman and Sandelin, 2004):

$$IC = 2 + \sum_{i=1}^{k} \sum_{b} \log_2 p_b^{(i)} \tag{18}$$

This is equivalent to the relative entropy of the distribution defined by the motif model with respect to a uniform distribution over the four bases per position, and can be thought of as a measure of departure from randomness.

It is worth noting that, if the IC were defined as the relatively entropy with respect to the background model π rather than the uniform distribution, then it would be equal to the expected value of $S(X)$ under the distribution defined by the motif model. Thus, provided π is reasonably close to a uniform distribution, the IC (as defined above) should be roughly equal to the average binding affinity of a large collection of true binding sites.

To examine the relationship between binding affinity and the fraction of sites under selection (ρ; Fig. 3C), we pooled the TFBSs for all TFs, calculated $S(X)$ for each one using the corresponding motif model, listed them in increasing order by $S(X)$, and partitioned this list into twenty equally sized sub-lists. We then used INSIGHT to estimate a full set of parameters for the binding sites in each of these twenty groups, allowing the binding sites in each group to represent diverse transcription factors. The maximum likelihood estimates of ρ are plotted against the average binding affinity of each group in Fig. 3C. The strong correlation between these quanities is partly a consequence of the correlation between IC and ρ shown in Fig. 3A, but the stronger relationship with binding affinity suggests that some of the variation in selective pressure across binding sites for the same TF is accounted for by variation in binding affinity.

To create Supplementary Fig. 3 we again pooled data from all TFs, but this time we sorted TFBSs into three groups: ones containing no mutations (i.e., no polymorphisms or divergences), ones containing at least one mutation such that the binding affinity of the mutated binding site is less than that of the unmutated (ancestral) binding site, and ones containing at least one mutation such that the binding affinity of the mutated binding sites is greater than that of the unmutated binding site. (Almost no mutations leave the binding affinity exactly unchanged.) The first group, with no mutations, was largely uninformative and was discarded. We then computed the posterior expected values: ρ, $\mathbb{E}[A]$, and $\mathbb{E}[W]$, for each TFBS in each of the two groups of mutated sites, using the parameters estimated for the corresponding TF in each case (see Section 5). These are the values plotted in Supplementary Fig. 3.

8.2 Distance from coding exons

To examine the relationship between the estimated selection parameters and the distances of TFBSs from known functional elements, we pooled all TFBSs and partitioned them into 50 groups by physical distance from the nearest annotated coding exon. We excluded TFBSs that overlapped exons of protein-coding or RNA genes. We then applied INSIGHT separately to each partition. Note that these sets of TFBS likely differ in their distributions of associated TFs, because some TFs are more likely to bind to distal binding sites, while others tend to bind sequences in proximal promoters or introns. However, our data set was not sufficiently large to perform this analysis separately for each TF. Importantly, we also expect differences among partitions in patterns of neutral variation, with reduced diversity and divergence near coding exons owing to the effects of selection on linked sites (background selection and/or hitchhiking; McVicker et al. (2009); Gottipati et al. (2011)). A secondary goal of this analysis was therefore to gain insight into possible biases in our parameter estimates stemming from such distortions in neutral variation.

Nature Genetics: doi:10.1038/ng.2658

To distinguish between the influences of selection and neutral variation on our parameter estimates, we defined two sets of "pseudo-TFBSs," similar in length to the true TFBSs (we fixed their lengths at 20 bp) but located in nearby sites that had passed our neutral filters. The first set, the *flanking pseudo-TFBSs*, were located immediately next to the true TFBSs (no more than 100 bp away), while the second set, the *random pseudo-TFBSs*, were randomly distributed in the neutral blocks containing the TFBSs. Note that the flanking pseudo-TFBSs will be better surrogates for the true TFBSs in the presence of significant fine-scale variation in patterns of neutral diversity. However, they are also more likely than the random pseudo-TFBSs to be influenced by cryptic (unannotated) binding sites or other functional elements, because functional elements tend to cluster in the genome.

The results of this experiment (Supplementary Fig. 5) indicate that, for true TFBSs, distance from exons is at most weakly correlated with ρ and $\mathbb{E}[A]$/kbp, but it displays a fairly striking negative correlation with $\mathbb{E}[W]$/kbp. By itself, this negative correlation could either indicate that weak negative selection is substantially more prevalent in TFBS near to coding exons than in those distal from exons, or that selection from linked sites (or another covariate) creates a spurious, distance-dependent signal for weak negative selection, perhaps by altering the neutral ratio of low-frequency-to-high-frequency derived alleles.

Our pseudo-TFBSs allow us to distinguish between these possibilities. These sets display much reduced evidence of selection and little dependency on distance from exons. We do observe a slight negative correlation of distance with $\mathbb{E}[W]$/kbp for the flanking pseudo-TFBSs, but this correlation is essentially absent from the random pseudo-TFBSs. While we cannot rule out an influence from distance-dependent fine-scale variation in neutral variation, it seems much more likely that this weak correlation is driven by cryptic functional elements, which will tend to be densest near exons. Together, these results indicate that our parameter estimates are primarily driven by real signatures of selection, rather than fine-scale variation in patterns of neutral variation, and that our method for pairing TFBSs with flanking neutral sites adequately controls for the effects of selection on linked sites. They also suggest a genuine enrichment for segregating weakly deleterious alleles in TFBSs near coding exons as compared with more distal binding sites. However, this phenomenon appears to have little effect on the total fraction of sites under selection (ρ).

8.3 Gene Ontology enrichments

Each TFBS was associated with one or more genes using the Genomic Regions Enrichment of Annotations Tool (GREAT) (McLean et al., 2010). This was accomplished by extracting the lists of associated genes reported by the web-based interface. We used the "Basal plus extension" association rules, which define a gene's basal regulatory domain to include the regions 5 kb upstream and 1 kb downstream of the annotated transcription start site, regardless of other nearby genes, and then extend this domain in both directions until either another gene's basal regulatory domain is encountered, or a maximum distance of 1 Mb is reached. A small number of additional curated regulatory domains are also included (see http://great.stanford.edu).

We downloaded the 07/2012 release of the GO database from http://geneontology.org and obtained the full set of GO categories reported for each gene identified by GREAT, cross-referencing by gene symbol. We discarded categories in levels 1–2, which were generally too broad to be useful, focusing on subclasses of the "Biological Process" ontology. We then assigned to each TFBS in our set the union of the GO categories of all associated genes. Finally, we discarded all GO categories assigned to fewer than 100 binding sites. This left 3,551 GO category $\times$ TF combinations. For each such $g \times t$ combination, we computed $\mathbb{E}[A]$/kbp and $\mathbb{E}[W]$/kbp for all binding sites of t assigned to g (hereafter, the g set) and for all binding sites of t not assigned to g (the $\bar{g}$ set). For ρ we found the posterior expected values were uninformative, so we used INSIGHT to obtain maximum likelihood estimates of ρ for the g and $\bar{g}$ sets.

To assess statistical significance of differences in $\mathbb{E}[A]$/kbp and $\mathbb{E}[W]$/kbp, we obtained an approximate null distribution by randomly sampling 10,000 subsets of size $|g|$ from the set $\bar{g}$ and recomputing the statistic for each sample. We then computed empirical one-sided p values for our estimate for the g set with respect to

Nature Genetics: doi:10.1038/ng.2658

this distribution. In the case of ρ we performed a likelihood ratio test for a difference in the parameter values of the g and $\bar{g}$ sets (see Section 5). Because these 3,042 tests ($\times$ 3 parameters) were highly correlated and violate the assumptions of easily applied multiple testing corrections we simply report uncorrected, nominal p-values and caution the reader that they should not be taken to indicate family-wise statistical significance.

We observed elevations in ρ, $\mathbb{E}[A]$/kbp, and $\mathbb{E}[W]$/kbp for several $t \times g$ combinations (Supplementary Tables 7–9). The identified categories were generally diverse, but a few patterns of interest emerged. Several TFs had substantially elevated values of $\mathbb{E}[A]$/kbp, indicating adaptive evolution, near genes involved in neural processes (e.g., "memory" [MAFK], "nervous system development" [FOXA1], and "brain development" [C-FOS]) This trend is consistent with observations from an analysis of evolutionary rates in promoter regions of primate genomes (Haygood et al., 2007), but in our case they are based on particular transcription factors and binding sites, and reflect more recent evolution. In addition, several TFs had elevated values of ρ, indicating increased selective pressure, near genes involved in development or cellular differentiation (e.g., "muscle organ development" [CEBPB], "blood vessel development" [RAD21], "multicellular organismal development" [SRF], and "in utero embryonic development" [AP-2GAMMA]).

9 Deleterious mutations

9.1 Numbers of deleterious mutations in coding regions

By various methods, Fay et al. (2001) obtained estimates of 329, 513, and 500–1200 deleterious coding mutations per individual. However, they assumed a genome-wide total of 5×10^7 coding bases in these calculations, while we assume 3.66×10^7 based on current protein coding annotations. If we rescale Fay et al.'s estimates using our assumption for the true number of CDS bases, and then halve the scaled estimates to convert from diploid to haploid units, we obtain adjusted estimates of 120, 188, and 183–439, respectively.

Lohmueller et al. (2008) performed a slightly different analysis, estimating numbers of "damaging" alleles per individual, rather than numbers of weakly deleterious alleles. They used the PolyPhen program (Ramensky et al., 2002) to obtain these estimates. Lohmueller et al. estimated an average of 426.1 heterozygous and 91.7 homozygous damaging alleles per individual from 10,150 genes. This corresponds to $426.1/2 + 91.7 \approx 304$ damaging alleles per haploid genome. If we scale these estimates up to 22,500 genes, we obtain an estimate of 674 damaging coding alleles per haploid genome. A earlier analysis by Sunyaev et al. (2001) used related methods to predict about 2000 deleterious nonsynonymous variants per individual, assuming 45,000 genes. Rescaling to 22,500 genes results in an estimate of 1000 deleterious alleles per individual, or 500 per haploid genome.

Thus, our estimate of $\mathbb{E}[D] = 738.1$ weakly deleterious mutations per haploid genome for CDSs is reasonably concordant with the estimates of Lohmueller et al. (2008) and Sunyaev et al. (2001) despite the use of quite different methods. Our estimate is somewhat larger than the estimates obtained by Fay et al. (2001), but this study made use of two rather small collections of genes (consisting of 106 and 75 genes), which may not have been representative of the genome as a whole.

9.2 Genetic load

Under a multiplicative model, the relative fitness of a haploid genome containing D mutations with selection coefficients $s_1, \ldots, s_D$ is $w = (1 + s_1)(1 + s_2) \cdots (1 + s_D)$. If s_i is negative and close to 0 for all i, as expected for weakly deleterious mutations, then w is well approximated by $w \approx 1 + \sum_{i=1}^{D} s_i = 1 + D\bar{s}$, where $\bar{s}$ is the arithmetic mean of the selection coefficients, $\bar{s} = \frac{1}{D}\sum_{i=1}^{D} s_i$. The quantity $-D\bar{s}$ can be thought of as the average reduction in fitness per haploid genome, or equivalently, as the number of "lethal equivalents per gamete" (Morton et al., 1956; Bittles and Neel, 1994).

Nature Genetics: doi:10.1038/ng.2658

We can obtain approximate estimates of the average selection coefficient for segregating deleterious mutations ($\bar{s}$) in two ways. First, population genetic theory suggests that selection coefficients for segregating deleterious mutations should fall approximately in the range $-1/N_e \leq s \leq -1/4N_e$, where N_e is the effective population size, because genetic drift will dominate at smaller values of $|s|$, and mutations will be eliminated rapidly from the population at larger values of $|s|$ (Kondrashov, 1995). This is equivalent to a range for the population-scaled selection coefficient of $-2 \leq 2N_e s \leq -0.5$. However, these assumptions do not take into account the influence of population bottlenecks, expansions, and other demographic effects, which might allow sites under stronger selection to remain polymorphic.

We can obtain alternative estimates of $\bar{s}$ from a series of simulation experiments that were designed to test the sensitivity of our methods to assumptions about the true distribution of fitness effects (Gronau et al., 2013). Specifically, we can simply calculate the average selection coefficient of segregating deleterious mutations for the data set generated under each simulation scenario, to obtain empirical, simulation-based estimates of $\bar{s}$. Because these simulations were based on realistic assumptions about human demography, these estimates may be more accurate than the theoretical ones for our purposes. Note that, because we are interested in the average reduction in fitness per haploid genome, these averages must be computed across chromosomes, not across polymorphic sites—i.e., the selection coefficient at each polymorphic site must be weighted by the number of derived alleles at that site.

For the thirteen simulation scenarios considered, we obtain average population-scaled selection coefficients for segregating deleterious alleles ranging from -0.8 to -33.0 (Supplementary Table 10). However, the two scenarios that produced the smallest estimates (in absolute value) involved assumptions of quite weak negative selection ($2N_e|s| \leq 2$), for which our methods have limited power (Gronau et al., 2013), hence these are likely to be underestimates for detectable alleles. In addition, the scenario that produces the largest estimate is based on a somewhat implausible assumption of $2N_e s = -20$ for weak negative selection. If we discard these three outliers, we obtain a range for $2N_e\bar{s}$ of -16.7 to -7.7. These estimates naturally depend somewhat on the mixture proportions and selection coefficients assumed for the simulation experiments, but they appear to be reasonably insensitive to moderate differences in the assumed distribution of fitness effects. The reason is that selection acts as a filter, discarding mutations under strong selection and allowing only those under weak selection to persist as polymorphisms.

As expected, our estimate for $2N_e\bar{s}$ based on demographically realistic simulations (between -16.7 and -7.7) is considerably larger in absolute value than the theoretical prediction of -2 to -0.5. Given the rather extreme demographic scenario for human populations and the fact that our methods likely have a slight ascertainment bias favoring the detection of sites under stronger selection, we expect the simulation-based estimates to be more representative of the sites detected by our methods and use them in the calculation reported in the text.

The reported numbers are obtained by dividing the simulation-based estimates of $2N_e\bar{s} \in [-16.7, -7.7]$ by $2N_e$, assuming $N_e = 10{,}000$, and then multiplying the resulting estimates of $\bar{s}$ by $-\mathbb{E}[D]$, where $\mathbb{E}[D]$ is the expected number of deleterious mutations per haploid genome. We consider cases of $\mathbb{E}[D] = 722.3$ extrapolated CDS mutations and $\mathbb{E}[D] = 3099.8$ total extrapolated CDS and TFBS mutations (see Supplementary Fig. 7), obtaining approximate ranges of 0.3–0.6 and 1.2–2.6, respectively, for the number of lethal equivalents per gamete.

10 Robustness to choice of neutral sites

INSIGHT is designed to accommodate kilobase-scale genomic variation in features that influence patterns of neutral diversity and divergence, by measuring polymorphism and divergence rates in binding sites relative to the rates in nearby neutral regions. It is conceivable, however, that biases in our parameter estimates could result from finer-scale differences between binding sites and their immediate flanking regions in mutation

Nature Genetics: doi:10.1038/ng.2658

rates, fixation rates, or ascertainment of sequence variants. Possible sources of such differences include: (1) G+C content, which is well known to correlate with mutation rate (e.g., Mouse Genome Sequencing Consortium, 2002); (2) GC-biased gene conversion, which can produce elevated derived allele frequencies and/or substitution rates by favoring the fixation of G and C alleles (Duret and Galtier, 2009; Kostka et al., 2012); (3) chromatin accessibility, for which there is some recent evidence of a correlation with mutation rate (Thurman et al., 2012); and (4) sequence coverage, which correlates with G+C content, and may influence the power for detection of low-frequency polymorphisms. All of these features may differ to a degree between binding sites and their immediately flanking regions, hence they might contribute to systematic biases in parameter estimates.

Below we report the results of a series of experiments designed to test for biases of this kind. We show that, despite some differences between binding sites and their flanks, none of these features is likely to have a major influence on our parameter estimates. Notably, some of the features considered here—particularly G+C content and chromatin accessibility—are also likely correlated with distance from the nearest coding exon, another correlate of patterns of neutral variation that could differ between binding sites and their flanks. (In particular, because our neutral filters exclude coding exons, binding sites near exon boundaries will tend to be associated with neutral sites that are somewhat farther from the exons than they are.) Thus, the results in this section complement those reported in Section 8.2. Together, we believe these two sets of analyses demonstrate that our neutral sites are adequate proxies for neutral evolutionary processes in the analyzed binding sites, and are unlikely to contribute to significant biases in parameter estimates.

10.1 Alternative neutral sites

In our first experiment, we attempted to address the concern about systematic differences between binding sites and their flanking sequences in as general a manner as possible by making use of an alternative method for matching TFBSs with neutral sites. This method did not make use of local genomic location but instead explicitly matched TFBSs with neutral sites drawn from similar regions across the genome, based on several features known to be correlated with patterns of neutral variation.

First, we divided each 20 kbp neutral block into mini-blocks of 200 bp. We then annotated these mini-blocks with category labels reflecting their G+C content, recombination rate, and genetic distance from the nearest coding exon, relative to the other mini-blocks.[3] Specifically, for each measure of interest (G+C content, etc.), we ordered all mini-blocks by this measure, divided them into equal-sized "bins," and labeled each mini-block according to the bin in which it was placed. In addition, each mini-block was annotated with category labels reflecting the estimated neutral polymorphism (θ) and divergence (λ) rates for the containing 20-kbp block. We performed two versions of the experiment: one with all five covariates (G+C content, recombination rate, genetic distance from the nearest coding exon, θ, and λ) and one with three covariates (G+C content, θ, and λ). In the first case we used five bins per covariate and in the second case we used ten bins per covariate.

We then considered the Cartesian product of all category labels—that is, the set of $5^5 = 3,125$ possible joint labelings of mini-blocks in the five-covariate version, and the set of $10^3 = 1,000$ in the three-covariate version. We identified the collection of mini-blocks assigned each possible joint labeling and calculated an average θ and λ value for the neutral sites by pooling data from those blocks. These estimates of θ and λ were subsequently treated as the neutral rates of diversity and divergence, respectively, for binding sites having the same joint category labels as those neutral blocks. The labels for binding sites were inherited from the closest mini-block, and therefore reflected features within about a 100 bp radius of the binding site, on average.

[3]Recombination rates for each mini-block and genetic distance to the nearest exon were estimated using the hg19 release of the HapMapII genetic map: http://hapmap.ncbi.nlm.nih.gov/downloads/recombination/2011-01_phaseII_B37.

Nature Genetics: doi:10.1038/ng.2658

As a concrete example, consider the mini-blocks that fall in the 3rd G+C bin, the 2nd recombination rate bin, the 5th distance-from-coding-exon bin, the 1st coarse-grained λ bin, and the 3rd coarse-grained θ bin (assuming the five-covariate experiment). Call this the 3,2,5,1,3 set. The nucleotide positions falling in all of the mini-blocks in the 3,2,5,1,3 set were pooled, and global estimates of θ and λ—call them $\theta_{3,2,5,1,3}$ and $\lambda_{3,2,5,1,3}$—were obtained from these pooled sites. Subsequently, whenever a TFBS was assigned to the 3,2,5,1,3 set, based on the closest mini-block to that site, the neutral parameters used by INSIGHT for that binding site were taken to be $\theta_{3,2,5,1,3}$ and $\lambda_{3,2,5,1,3}$. These values were used in place of the θ and λ estimates for the neutral block in which the TFBS fell, as considered in our main analysis. The result of this strategy was to assign to binding sites neutral parameters that reflected not only (coarse-grained) local levels of diversity and divergence, but also the fine-scale influence of features such as G+C content and recombination rate. Note in particular that, under this scheme, a binding site that falls in a sequence that is high in G+C content will be assigned neutral parameters estimated from mini-blocks that are similarly high in G+C content.

We repeated our analysis of the binding sites of all TFs using each of these binning strategies in place of our original method. The estimated parameters and expected values were highly consistent with those from the original analysis, particularly when the uncertainty in the estimates is considered (Supplementary Fig. 8). These results indicate that our results are not highly dependent on the use of neutral sites that immediately flank the analyzed TFBSs, and that our estimates are unlikely to be strongly influenced by fine-scale variation in G+C content, recombination rate, distance-to-nearest-coding-exon, or correlated features.

We note that, while the experiment above does not directly address differences in local nucleotide "context" (e.g., at the di- or tri-nucleotide level), which could also contribute to local differences in mutation rates (Siepel and Haussler, 2004; Hwang and Green, 2004), nucleotide di- and tri-nucleotide composition tends to be fairly strongly correlated with features that we did consider (particularly G+C content and distance-from-coding exon) and therefore should be addressed indirectly, to a degree. In addition, our filtering procedure eliminates CpG dinucleotides (Section 4), the positions at which context effects are most pronounced. We did scan our main results for correlations between the proportion of particular dinucleotides in TFBSs and estimates of ρ (data not shown), but we found the results to be difficult to interpret, because of the strong dependency of dinucleotide content in TFBSs on the particular binding motif, and the pronounced differences between TFs in estimates of ρ (Fig. 2). Therefore, we cannot completely rule out an influence on our parameter estimates from context-dependent mutation rates, but we think it is unlikely that they are playing a major role in our analysis.

10.2 Biased gene conversion

We examined the possible impact of GC-biased gene conversion (gBGC) on our parameter estimates in two ways. First, we simply examined the ratio of weak-to-strong (A/T $\rightarrow$ G/C) vs. strong-to-weak (G/C $\rightarrow$ A/T) substitutions at positions that passed all of our filters and that were not polymorphic in the human chromosomes in our sample (i.e., apparent fixed differences). We observed a slight deficiency of weak-to-strong substitutions in TFBSs (weak-to-strong/strong-to-weak ratio of 0.92), suggesting that they are not strongly influenced by gBGC overall. The neutral blocks exhibited a similar deficiency of weak-to-strong mutations (ratio of 0.89). Because substitutions in the TFBSs are strongly constrained by the base preferences of the binding TFs, a slight difference between these ratios is not surprising.

Second, we compared our TFBSs and neutral blocks with genome-wide predictions of gBGC tracts based on a new program called phastBias (Capra et al., 2013). PhastBias uses a phylogenetic hidden Markov model to identify genomic tracts on the order of 1 kb in length that display a significant excess of weak-to-strong substitutions on the human lineages, given a pre-estimated model of neutral evolution. The model also considers evolutionary conservation in functional elements, which can be a confounding factor in the identification of gBGC tracts. PhastBias was applied to genome-wide alignments of the human, chimpanzee,

Nature Genetics: doi:10.1038/ng.2658

orangutan, and rhesus macaque genomes, and predicted a collection of gBGC tracts that cover about 0.3% of the human genome.

We found that TFBSs were somewhat enriched (1.7-fold) for overlap with the predicted gBGC tracts, but this enrichment disappeared when they were compared with a set of control regions matched by G+C content (which displayed a 1.9-fold enrichment). Thus, the observed enrichment appears to be simply a function of the G+C content of the binding sites. The total number of affected binding sites was fairly small (5,815 TFBSs, or 0.4% of all binding sites), and our average parameter estimates showed little difference when these binding sites were excluded (e.g., average ρ increased from 0.329 to 0.334 and $\mathbb{E}[A]$ decreased from 0.123 to 0.120). The neutral blocks also showed no significant enrichment or depletion for gBGC tracts, and excluding the tracts from the neutral blocks had very little impact on our parameter estimates.

It is worth noting that our filters exclude sites in recently identified recombination hotspots (The 1000 Genomes Project Consortium, 2010), which should help to diminish any impact from gBGC.

10.3 Chromatin accessibility

Next, we examined the influence of chromatin accessibility on our parameter estimates, using DNase-I hypersensitive sequences (DHSs) from the ENCODE project (Thurman et al., 2012) as a surrogate for regions of open chromatin. Of particular interest was the possibility that differences in mutation rates in DHSs could lead to biased estimates of selection parameters. For example, if some classes of DHSs display elevated mutation rates, as reported for cell types labeled as "pluripotent" or "malignant" by Thurman et al. (2012), and if our binding sites are more likely than their associated flanking sites to fall in such regions, as might be expected, we might observe a downward bias in estimates of ρ and/or an upward bias in estimates of $\mathbb{E}[A]$ due to an excess of polymorphism and divergence in elements relative to flanks. Conversely, if some DHSs display reduced mutation rates, we might observe an upward bias in ρ and/or a downward bias in $\mathbb{E}[A]$ in TFBSs associated with those DHSs.

We reasoned that our binding sites will generally fall in regions of open chromatin, while our flanking sites will display greater variation in their DHS overlap. Therefore, we examined the dependency of our parameter estimates on the fraction of neutral blocks that fall in DHSs associated with various cell types. If differences between TFBSs and flanking sites in chromatin accessibility were to have a significant influence on our estimates, we would expect to observe a correlation between this fraction and ρ, while if they were to have no influence we would expect little correlation.

We obtained DHSs from the ENCODE ftp site[4], keeping track of the cell types in which each DHS was observed. We then identified subsets of our TFBSs indicated as bound (via the presence of ChIP-seq peaks) in three subsets of cell types: (1) GM12878; (2) GM12878, HUVEC, GM12891, and GM12892 ("Normal"); and (3) HCT-116, HeLa-S3, HepG2, K562, PANC-1, and SK-N-SH ("Aberrant"). (These "Aberrant" cell types all fall in the group identified by Thurman et al. (2012) as "Malignant.") Next we partitioned each of these three sets of TFBSs into 15 bins of equal size, based on the fraction of flanking neutral sites that fell in DHSs associated with the same cell-types. To avoid confounding effects, we further excluded all elements that fell within gene transcripts or phylogenetically conserved noncoding sequences. Finally, we ran INSIGHT independently on the TFBSs in each bin. Consistent with our assumptions, we observed much higher rates of overlap of TFBSs with corresponding DHSs than for flanking regions: the fraction of nucleotides falling in matched DHSs was ∼45–61% for the TFBSs, and ∼0–20% for the various partitions of flanking sites.

We observed no significant correlation between the estimated values of ρ and the fraction of flanking sites in DHSs, for any of the three cell-type groups considered (Supplementary Fig. 9; $p > 0.45$, $R^2 < 0.03$), suggesting that differences between binding sites and flanking regions in chromatin accessibility have

[4] ftp://ftp.ebi.ac.uk/pub/databases/ensembl/encode/integration_data_jan2011/byDataType/openchrom/jan2011/combined_peaks/

Nature Genetics: doi:10.1038/ng.2658

not played a major role in our analysis. Of course, it is conceivable that our analysis is influenced in a complex manner by several variables, of which chromatin accessibility is one, and that this influence does not produce an easily detectable marginal correlation. Nevertheless, using the data we have available, we see no evidence of an obvious influence from chromatin accessibility on our parameter estimates.

10.4 Sequencing coverage

We examined the sequencing coverage of the TFBSs and flanks, looking for differences sufficiently pronounced to produce important differences in polymorphism detection rates. We found that the Complete Genomics sequencing data does exhibit a clear dependency of read depth on G+C content (Supplementary Fig. 10A), but that average read depths are quite high (>25 reads per site) across a broad range of G+C levels. There is only a very slight downward shift in the distribution of read depths in TFBSs relative to the flanking neutral regions (Supplementary Fig. 10B). Together with our direct controls for G+C content (above), this analysis suggests that it is unlikely that differences in coverage play a significant role in our analysis.

11 Robustness to choice of binding sites

To evaluate the sensitivity of our results to our pipeline for binding site identification, we examined two alternative sets of TFBSs identified for the ENCODE project (Dunham et al., 2012). Set #1 consists of binding sites based on either known motifs (from JASPAR) or motifs discovered from ENCODE ChIP-seq peaks (see http://www.broadinstitute.org/~pouyak/encode-motif-disc/), which were mapped to the genome and filtered to require intersection with ChIP-seq or DNase-seq peaks. Set #2 is a smaller collection of binding sites, identified by similar methods but with more stringent filtering for use in the satellite paper by Ward and Kellis (2012).

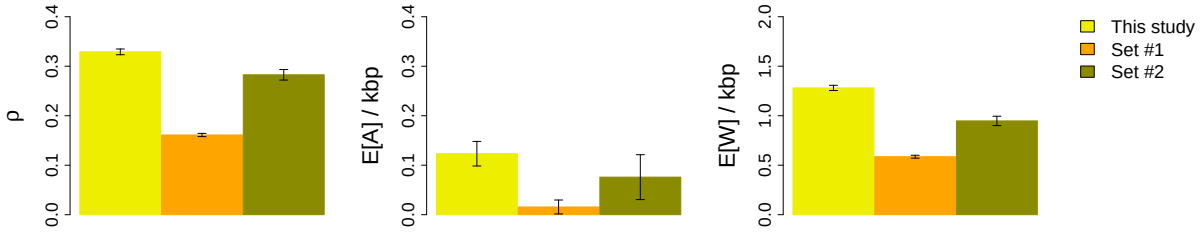
We encountered several difficulties in comparing these two data sets with ours. First, set #1 is extremely large—it contains 15.8 million TFBSs (more than 13 times as many as in our set, and more than twice as many as in our crude extrapolation to the complete genome)—and for many of the motifs based on de novo discovery we could find no match to motifs reported in the literature or present in the JASPAR or TRANSFAC databases. Even if the binding sites based only on de novo discovery are excluded, the data set still contains almost 6 million TFBSs. Second, motifs in this set were identified by TF family rather than the names of individual TFs, making it difficult to distinguish among, say, GATA1, GATA2, and GATA3 binding sites. The same was done for the FOXA, MEF2, BCL, and AP-2 families, among others. Third, some TFs included in our analysis (such as FOXP2) are not represented in the ENCODE set. Finally, the filtering criteria for binding sites used by the ENCODE group were quite different from ours. For example, their set includes predicted GATA binding sites that fall in DNase-seq peaks, even if they are not supported by a single GATA ChIP-seq experiment in a single cell type, while ours does not. In general, the filtering criteria used to define this set seem to have favored sensitivity over specificity, leading to high coverage but likely also to fairly high false discovery rates.

Set #2, on the other hand, is considerably smaller than ours, with about a fourth as many TFBSs. Limited information is available on how this set was defined, but the procedure seems to have involved a reduction of an initial collection of TFBSs roughly like set # 1 by choosing a single motif for each TF family followed by filtering based on ChIP-seq experiments for that TF family (see Ward and Kellis, 2012). Thus, the mismatch in the ChIP-seq filtering criteria described above is not an issue with set #2, but the problem of grouping of TFs by family remains. In addition, the actual motifs used in defining this set have not been made available, so we are unable to compare them with ours.

Despite these difficulties, we proceeded to apply INSIGHT to both ENCODE data sets after removing

Nature Genetics: doi:10.1038/ng.2658

any TF family with fewer than 500 TFBS. (In the case of set #1, we used the higher-confidence subset of 6 million TFBSs supported by JASPAR, rather than the larger set of 15.8 million TFBSs.) Given the differences in the criteria used, we observed reasonable overall concordance in parameter estimates between our data set and ENCODE set #2. In particular, the weighted average of the estimate of ρ for the 49 TFs in set #2 is 0.28, compared with a value of 0.33 for the 78 TFs considered in our analysis (see figure below). Similarly, the estimates for $\mathbb{E}[A]$ and $\mathbb{E}[W]$ are 0.08 and 0.95, respectively, compared with 0.12 and 1.28 for our data set. The agreement at the level of individual TFs is variable, with some agreeing well and others showing substantial differences. Inspection of individual cases suggests that the most divergent parameter estimates likely reflect differences in motif definitions, TF grouping, and filtering. By contrast, set #1 yields substantially lower estimates for the selection parameters, with ρ = 0.16, $\mathbb{E}[A]$ = 0.016, and $\mathbb{E}[W]$ = 0.59 (52 TFs). This difference is perhaps not surprising given the much more inclusive nature of this data set.



Average parameter estimates for alternative sets of TFBSs. Averages values (weighted by numbers of TFBSs per TF) of ρ, $\mathbb{E}[A]$, and $\mathbb{E}[W]$ are shown for our TFBSs and the two alternative sets described above.

We conclude from this analysis that the estimates obtained by INSIGHT are fairly highly dependent on the stringency used in TFBS identification, but given appropriately rigorous criteria, they are reasonably robust to the precise definitions used. It is of course possible that the reduced signature of selection in the ENCODE data sets results in part from the inclusion of real TFBSs that are under weaker selection than the ones in our set, but with currently available data, we think it is difficult to distinguish this effect from the influence of false positive TFBS predictions. Our results should therefore be interpreted as describing a subset of binding sites with strong experimental and bioinformatic support, and it should be understood that they may not be representative of all binding events genome wide. It will be an interesting topic for future work to measure the influence of selection on true, but weak, binding sites, in contrast with stronger, easier-to-identify TFBSs.

12 Robustness to binding site turnover

Transcription factor binding sites are known to be gained, lost, and altered in position over evolutionary time, by a variety of molecular evolutionary processes (e.g., Dermitzakis and Clark, 2002; Moses et al., 2006; Schmidt et al., 2010). We refer to these changes in binding site identity or position collectively as "turnover". Owing to turnover, some genuine human TFBSs likely do not correspond to functional binding sites at orthologous locations in other species. INSIGHT is designed to protect against turnover by using the outgroup genomes only to provide information about the ancestral allele Z_{ij}, and focusing on evolutionary processes along the branch to humans. However, turnover could still conceivably contribute to biases in our parameter estimates in various ways.

Perhaps the most serious concern is that losses of binding sites in one or more outgroup species (particularly the chimpanzee) could lead to misspecification of the ancestral allele Z_{ij} and produce inflated counts of divergences on the human branch, which could in turn lead to spurious inferences of adaptive evolution. For example, suppose an ancestral binding site (functional in the MRCA of humans and chimpanzees) is

Nature Genetics: doi:10.1038/ng.2658

lost on the chimpanzee lineage but remains functional in humans. Consider a nucleotide in this binding site that is under negative selection in humans. By our assumptions the nucleotide observed in the human genome must be the ancestral allele, but a substitution may occur in the chimpanzee genome due to drift. If the same substitution occurs on the orangutan or rhesus macaque lineages, say, because of a parallel loss of the binding site, a relaxation of constraint at that position, or persistence of a derived allele despite negative selection, then the ancestral allele may be misspecified and the human allele misinterpreted as a fixed difference. Notice that this scenario still requires parallel substitutions on at least two lineages, but such changes will be marginally more likely in the presence of binding site losses, and it is conceivable that high rates of loss across the genome could lead to a significant bias in parameter estimates in aggregate.

To test the effect of plausible rates of binding site loss on the chimpanzee branch, we carried out simulations like those described in Section 2 except that at 20% of the binding sites selection was "switched off" on the chimpanzee lineage immediately after its divergence from the human lineage (Supplementary Fig. 11). As in our other simulation experiments, we then applied INSIGHT to these data after inferring ancestral alleles, exactly as for real data. We observed almost no effect on our parameter estimates under this simulation scenario (considering the standard errors of the estimates), even with these fairly substantial rates of chimpanzee loss (compare Supplementary Fig. 11 with Supplementary Fig. 13, which describes an analogous case with no turnover). The results for a rate of loss of 30% (not shown) were nearly identical. Some bias could almost certainly be produced by pushing the rates of turnover or rates of neutral evolution to extreme values, but the method seems to be highly robust to plausible rates of binding site loss for primates.

The other turnover scenario of primary concern is the gain of a new binding site on the human branch at a location at which no functional binding site is present in orthologous outgroup sequences. This case is substantially more complex, because one can imagine various ways in which it might occur. For example, a point mutation could create a functional binding site from a nonfunctional sequence, causing a neutrally evolving sequence to come under selection. In such a case, it seems most likely that the new sequence would not initially be in an optimal configuration and the new binding site would experience positive selection for a time before coming under constraint. The model will tend to treat these substitutions as adaptive, because they will tend to occur at a rate that exceeds that observed in flanking sequences, and human polymorphisms will tend to be eliminated from the now-selected TFBS. This is a complex, difficult-to-model scenario, but the behavior of our simplified model in this case seems fairly reasonable—after all, these substitutions driving the formation of a new TFBS *are* adaptive in a sense, even if the first one occurred at a site that was not formerly under selection.

One can imagine an alternative scenario in which a neutral sequence rapidly switches to negative selection due to exogenous influences (say, a change in chromatin structure, a binding TF, or a nearby binding site). If a substitution has occurred at a position in this sequence prior to the emergence of negative selection, it may be misinterpreted by our model as an adaptive substitution, because negative selection will tend to eliminate polymorphisms. The problem here is that a switch from neutrality to negative selection along the human branch leads to a violation of one of the key assumptions of our model: that nucleotide positions currently under negative selection cannot contain fixed differences. Simulation results (not shown) indicate that this type of emergence of binding sites fully formed (with no intermediate adaptive regime), if it is fairly common (affecting, say, 20% of binding sites), can lead to a modest but significant over-estimation of $\mathbb{E}[A]$ and α. However, we know of no evidence suggesting that binding sites have emerged in this way at high rates during human evolution, and it seems unlikely that this phenomenon has a pronounced effect on our results.

Nature Genetics: doi:10.1038/ng.2658

References

Bailey, T. L. and Elkan, C., 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In *Proc. 6th Int'l Conf. on Intelligent Systems for Molecular Biology*, pages 28–36.

Berg, O. G. and von Hippel, P. H., 1987. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J Mol Biol*, **193**(4):723–750.

Bittles, A. H. and Neel, J. V., 1994. The costs of human inbreeding and their implications for variations at the DNA level. *Nat. Genet.*, **8**(2):117–121.

Bustamante, C. D., Fledel-Alon, A., Williamson, S., Nielsen, R., Hubisz, M. T., Glanowski, S., Tanenbaum, D. M., White, T. J., Sninsky, J. J., Hernandez, R. D., *et al.*, 2005. Natural selection on protein-coding genes in the human genome. *Nature*, **437**(7062):1153–1157.

Capra, J. A., Hubisz, M. J., Kostka, D., Pollard, K. S., and Siepel, A., 2013. A model-based analysis of GC-biased gene conversion in the human and chimpanzee genomes. *ArXiv e-prints*, .

Dermitzakis, E. T. and Clark, A. G., 2002. Evolution of transcription factor binding sites in mammalian gene regulatory regions: conservation and turnover. *Mol Biol Evol*, **19**(7):1114–1121.

Drmanac, R., Sparks, A. B., Callow, M. J., Halpern, A. L., Burns, N. L., Kermani, B. G., Carnevali, P., Nazarenko, I., Nilsen, G. B., Yeung, G., *et al.*, 2010. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science*, **327**:78–81.

Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., Epstein, C. B., Frietze, S., Harrow, J., Kaul, R., *et al.*, 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**(7414):57–74.

Duret, L. and Galtier, N., 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu Rev Genom Hum G*, **10**(1):285–311.

Eisenberg, E. and Levanon, E. Y., 2003. Human housekeeping genes are compact. *TIG*, **19**(7):362–365.

Fay, J. C., Wyckoff, G. J., and Wu, C. I., 2001. Positive and negative selection on the human genome. *Genetics*, **158**:1227–1234.

Gottipati, S., Arbiza, L., Siepel, A., Clark, A. G., and Keinan, A., 2011. Analyses of X-linked and autosomal genetic variation in population-scale whole genome sequencing. *Nat. Genet.*, **43**(8):741–743.

Gravel, S., Henn, B. M., Gutenkunst, R. N., Indap, A. R., Marth, G. T., Clark, A. G., Yu, F., The 1000 Genomes Project, Gibbs, R. A., and Bustamante, C. D., *et al.*, 2011. Demographic history and rare allele sharing among human populations. *Proc. Natl. Acad. Sci. U.S.A.*, **108**(29):11983–11988.

Gronau, I., Arbiza, L., Mohammed, J., and Siepel, A., 2013. Inference of natural selection from interspersed genomic elements based on polymorphism and divergence. *Mol. Biol. Evol., in press*, **doi: 10.1093/molbev/mst019**.

Gronau, I., Hubisz, M. J., Gulko, B., Danko, C. G., and Siepel, A., 2011. Bayesian inference of ancient human demography from individual genome sequences. *Nat. Genet.*, **43**(10):1031–1034.

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B. L., Barrell, D., Zadissa, A., Searle, S., *et al.*, 2012. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.*, **22**(9):1760–1774.

Nature Genetics: doi:10.1038/ng.2658

Haygood, R., Fedrigo, O., Hanson, B., Yokoyama, K.-D., and Wray, G. A., 2007. Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution. *Nat Genet*, **39**(9):1140–1144.

Hernandez, R. D., 2008. A flexible forward simulator for populations subject to selection and demography. *Bioinformatics*, **24**:2786–2787.

Hubisz, M. J., Pollard, K. S., and Siepel, A., 2011. PHAST and RPHAST: Phylogenetic analysis with space/time models. *Briefings in Bioinformatics*, **12**:41–51.

Hwang, D. and Green, P., 2004. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc Natl Acad Sci USA*, **101**:13994–14001.

Jukes, T. H. and Cantor, C. R., 1969. Evolution of protein molecules. In Munro, H., editor, *Mammalian Protein Metabolism*, pages 21–132. Academic Press, New York.

Kondrashov, A. S., 1995. Contamination of the genome by very slightly deleterious mutations: why have we not died 100 times over? *J. Theor. Biol.*, **175**(4):583–594.

Kondrashov, A. S. and Crow, J. F., 1993. A molecular approach to estimating the human deleterious mutation rate. *Hum. Mutat.*, **2**:229–234.

Kong, A., Gudbjartsson, D. F., Sainz, J., Jonsdottir, G. M., Gudjonsson, S. A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G., *et al.*, 2002. A high-resolution recombination map of the human genome. *Nat. Genet.*, **31**(3):241–247.

Kosiol, C., Vinar, T., da Fonseca, R., Hubisz, M., Bustamante, C., Nielsen, R., and Siepel, A., 2008. Patterns of positive selection in six mammalian genomes. *PLoS Genet.*, **4**:e1000144.

Kostka, D., Hubisz, M. J., Siepel, A., and Pollard, K. S., 2012. The role of GC-biased gene conversion in shaping the fastest evolving regions of the human genome. *Mol. Biol. Evol.*, **29**(3):1047–1057.

Lohmueller, K. E., Indap, A. R., Schmidt, S., Boyko, A. R., Hernandez, R. D., Hubisz, M. J., Sninsky, J. J., White, T. J., Sunyaev, S. R., Nielsen, R., *et al.*, 2008. Proportionally more deleterious genetic variation in European than in African populations. *Nature*, **451**:994–997.

Machanick, P. and Bailey, T. L., 2011. MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics*, **27**(12):1696–1697.

McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., Wenger, A. M., and Bejerano, G., 2010. GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.*, **28**(5):495–501.

McVicker, G., Gordon, D., Davis, C., and Green, P., 2009. Widespread genomic signatures of natural selection in hominid evolution. *PLoS Genet.*, **5**:e1000471.

Morton, N. E., Crow, J. F., and Muller, H. J., 1956. An estimate of the mutational damage in man from data on consanguineous marriages. *Proc. Natl. Acad. Sci. U.S.A.*, **42**(11):855–863.

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**(10):e130.

Nature Genetics: doi:10.1038/ng.2658

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

Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R., and Siepel, A., 2010. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.*, **20**:110–121.

Ramensky, V., Bork, P., and Sunyaev, S., 2002. Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.*, **30**:3894–3900.

Schmidt, D., Wilson, M. D., Ballester, B., Schwalie, P. C., Brown, G. D., Marshall, A., Kutter, C., Watt, S., Martinez-Jimenez, C. P., Mackay, S., *et al.*, 2010. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science*, **328**(5981):1036–1040.

Schneider, T. D., Stormo, G. D., Gold, L., and Ehrenfeucht, A., 1986. Information content of binding sites on nucleotide sequences. *J Mol Biol*, **188**(3):415–431.

Self, S. and Liang, K., 1987. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Am. Stat. Assoc.*, **82**:605–610.

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

Smith, N. G. and Eyre-Walker, A., 2002. Adaptive protein evolution in Drosophila. *Nature*, **415**:1022–1024.

Spivakov, M., Akhtar, J., Kheradpour, P., Beal, K., Girardot, C., Koscielny, G., Herrero, J., Kellis, M., Furlong, E. E., and Birney, E., *et al.*, 2012. Analysis of variation at transcription factor binding sites in Drosophila and humans. *Genome Biol.*, **13**(9):R49.

Stormo, G. D., 2000. DNA binding sites: representation and discovery. *Bioinformatics*, **16**(1):16–23.

Sun, J. X., Helgason, A., Masson, G., Ebenesersdottir, S. S., Li, H., Mallick, S., Gnerre, S., Patterson, N., Kong, A., Reich, D., *et al.*, 2012. A direct characterization of human mutation based on microsatellites. *Nat. Genet.*, **44**(10):1161–1165.

Sunyaev, S., Ramensky, V., Koch, I., Lathe, W., Kondrashov, A. S., and Bork, P., 2001. Prediction of deleterious human alleles. *Hum. Mol. Genet.*, **10**(6):591–597.

The 1000 Genomes Project Consortium, 2010. A map of human genome variation from population-scale sequencing. *Nature*, **467**:1061–1073.

Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., Sheffield, N. C., Stergachis, A. B., Wang, H., Vernot, B., *et al.*, 2012. The accessible chromatin landscape of the human genome. *Nature*, **489**(7414):75–82.

Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S. F., Schroth, G. P., and Burge, C. B., 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**(7221):470–476.

Ward, L. D. and Kellis, M., 2012. Evidence of abundant purifying selection in humans for recently acquired regulatory functions. *Science*, **337**(6102):1675–1678.

Wasserman, W. W. and Sandelin, A., 2004. Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.*, **5**:276–287.

Nature Genetics: doi:10.1038/ng.2658