BMC Genomics

RESEARCH ARTICLE Open Access

Transcription-factor occupancy at HOT regions quantitatively predicts RNA polymerase recruitment in five human cell lines

Joseph W Foley[1,2*] and Arend Sidow[1,3]

Abstract

Background: High-occupancy target (HOT) regions are compact genome loci occupied by many different transcription factors (TFs). HOT regions were initially defined in invertebrate model organisms, and we here show that they are a ubiquitous feature of the human gene-regulation landscape.

Results: We identified HOT regions by a comprehensive analysis of ChIP-seq data from 96 DNA-associated proteins in 5 human cell lines. Most HOT regions co-localize with RNA polymerase II binding sites, but many are not near the promoters of annotated genes. At HOT promoters, TF occupancy is strongly predictive of transcription preinitiation complex recruitment and moderately predictive of initiating Pol II recruitment, but only weakly predictive of elongating Pol II and RNA transcript abundance. TF occupancy varies quantitatively within human HOT regions; we used this variation to discover novel associations between TFs. The sequence motif associated with any given TF's direct DNA binding is somewhat predictive of its empirical occupancy, but a great deal of occupancy occurs at sites without the TF's motif, implying indirect recruitment by another TF whose motif is present.

Conclusions: Mammalian HOT regions are regulatory hubs that integrate the signals from diverse regulatory pathways to quantitatively tune the promoter for RNA polymerase II recruitment.

Keywords: Transcription factor, ChIP-seq, HOT region, Gene regulation

Background

Transcription factors (TFs) are proteins that regulate the expression of genes by binding the DNA at their regulatory elements (promoters or enhancers) and either preventing or facilitating the recruitment, in eukaryotes, of the transcription preinitiation complex (PIC). The PIC in turn recruits RNA polymerase II (Pol II) to the transcription start site (TSS) to synthesize an RNA transcript. This is a primary mechanism for the regulation of gene expression in response to environmental stimuli or developmental programs.

Promoters must integrate a multitude of signals that converge on them in the form of activating or repressing transcription factors. In invertebrates, some regulatory regions ("high-occupancy target", or HOT, regions) are occupied by a large number of transcription factors [1-6]. However, less is known about the interactions among TFs at HOT regions and how these interactions contribute combinatorially to the regulation of transcription, and until recently, insufficient data existed to search for HOT regions in human cells.

The ENCODE data set [7,8] provides the first opportunity to study a large group of TFs together in human cells. These data come from the chromatin-immunoprecipitation sequencing (ChIP-seq) protocol: chromatin is crosslinked to preserve DNA-protein and protein-protein bonds, then a target-specific antibody is used to capture the DNA proximally associated with a given protein, and this DNA is sequenced and aligned to a

*Correspondence: joseph.foley@mail.mcgill.ca
[1]Department of Genetics, Stanford University, 300 Pasteur Drive, Stanford, California 94305, USA
[2]Current address: Douglas Mental Health University Institute, McGill University, 6875 Boulevard LaSalle, Verdun, Québec H4H 1R3, Canada
Full list of author information is available at the end of the article

reference assembly to create a genome-wide map of protein occupancy [9]. At each genome site occupied (though not necessarily bound directly) by a protein, ChIP-seq produces a tight cluster of read alignments, which can then be detected by software with high resolution.

Previous ChIP-seq analyses have generally considered a single experiment at once, and have treated TF occupancy as a binary signal—present vs. absent. However, the particular strength of the signal at any given site may represent important biological information, such as the persistence of occupancy within a cell or frequency across all cells in the sample. We developed a new software package, `UniPeak`, to analyze these data accordingly.

Using `UniPeak` to discover and quantify HOT regions, we performed a comprehensive analysis of these regulatory hubs. In particular, we characterized HOT regions with regard to other known markers of gene activity. We also compared the occupancy profiles of different TFs to predict novel interactions, and used mechanistic evidence to infer which complex members directly bind DNA. Finally, we quantified the relationship between TF occupancy and several measures of gene expression at HOT promoters.

Results

The human genome contains thousands of HOT regions

We obtained all publicly available ENCODE ChIP-seq data from the 5 most studied human cell lines [8], which assayed 96 DNA-associated proteins (Additional file 1: Table S1). These cell lines are derived from a variety of tissues and germ layers: GM12878 (lymphoblastoid/mesoderm), H1-hESC (embryonic stem cell), HeLa-S3 (epithelium/ectoderm), HepG2 (hepatic/endoderm), and K562 (leukocyte/mesoderm). We aligned the read sequences from each experiment to the `hg19` reference genome, standardizing the read length and removing low-confidence alignments in order to ensure accurate mapping without read-length bias.

`UniPeak` extends the peak-calling algorithm from `QuEST` [10] to the parallel analysis of multiple samples (Figure 1). Each aligned sequence read is considered one hit at the 5′ end of its alignment to the reference assembly. For each sample (i.e. a single replicate of a single experiment), `UniPeak` estimates the base-pair shift between strands, due to reading from opposite ends of sheared fragments, by selecting a shift value that maximizes strand correlations at the strongest regions. After shift correction of individual samples, kernel density estimation is used to compute a single smooth density profile for the combined signal of all samples. `UniPeak` identifies enriched regions where this profile exceeds a fixed threshold of fold enrichment relative to a uniform background distribution. The number of hits within each of these regions from each sample is reported, yielding a regions × samples matrix of hit counts. Unlike other peak-callers for ChIP-seq, `UniPeak` does not directly use "input" or other negative controls to filter enriched regions initially; rather, though these samples do not contribute to the region-calling step, negative-control reads (as well as histone-mark ChIP reads) are counted within the regions called from ChIP samples, and reported alongside read counts from TF ChIP samples. We normalized the peak intensities from discrete read counts to continuous occupancy values with the variance-stabilizing transformation in `DESeq` [11]. Performance validation of `UniPeak` is described in Additional File 2.

From the full set of 96 proteins in 5 cells, `UniPeak` detected 11,239 enriched regions (Table 1) of median size 136 bp (Additional file 1: Figure S3). Many of these appeared roughly evenly occupied by most proteins, with notable exceptions (Figure 2A). In particular, a large fraction of these regions were occupied only by the cohesin complex (CTCF, RAD21, SMC3), which, unlike canonical TFs, is known to bind insulator elements [12]. Cohesin-specific sites were less likely to be near a Pol II initiation site, and showed depletion of histone 3 lysine 4 trimethylation (H3K4me3), a chromatin mark associated with active promoters [13]. REST, a transcription repressor that binds the RE1 element to repress neuronal genes in non-neurons [9,14-16], similarly showed preferential occupancy in a large set of regions depleted for other TFs and for initiating Pol II.

To focus on only canonical TFs, which should be more functionally homogeneous and increase our specificity for HOT regions, we removed from the analysis four classes of proteins with different behaviors that confounded our goal of HOT region analysis. These were the cohesin complex, REST, chromatin remodelers and modifiers (e.g., p300 and SWI/SNF), and the preinitiation complex. The latter was later used to test functional predictions.

With this reduced set of 75 canonical TFs, `UniPeak` detected 7,227 regions (Table 1), of median size 171 bp (Additional file 1: Figure S3). Consistent with HOT regions, these regions were occupied by most or all TFs (Figure 2). Hierarchical clustering showed that the occupancy profiles of different TFs in the same cell were generally more similar than those of the same TF across all cells. In particular, GM12878, K562, and HepG2 each showed sets of HOT regions that were only occupied in one cell type, and these tended to be depleted for initiating Pol II and for histone 3 lysine 4 trimethylation vs. monomethylation; these regions might represent cell line-specific enhancers.

Because of these cell-specific signals and because of the incomplete overlap among the sets of TFs tested in different cells (Additional file 1: Table S1), we also used `UniPeak` to detect enriched regions in each of the 5

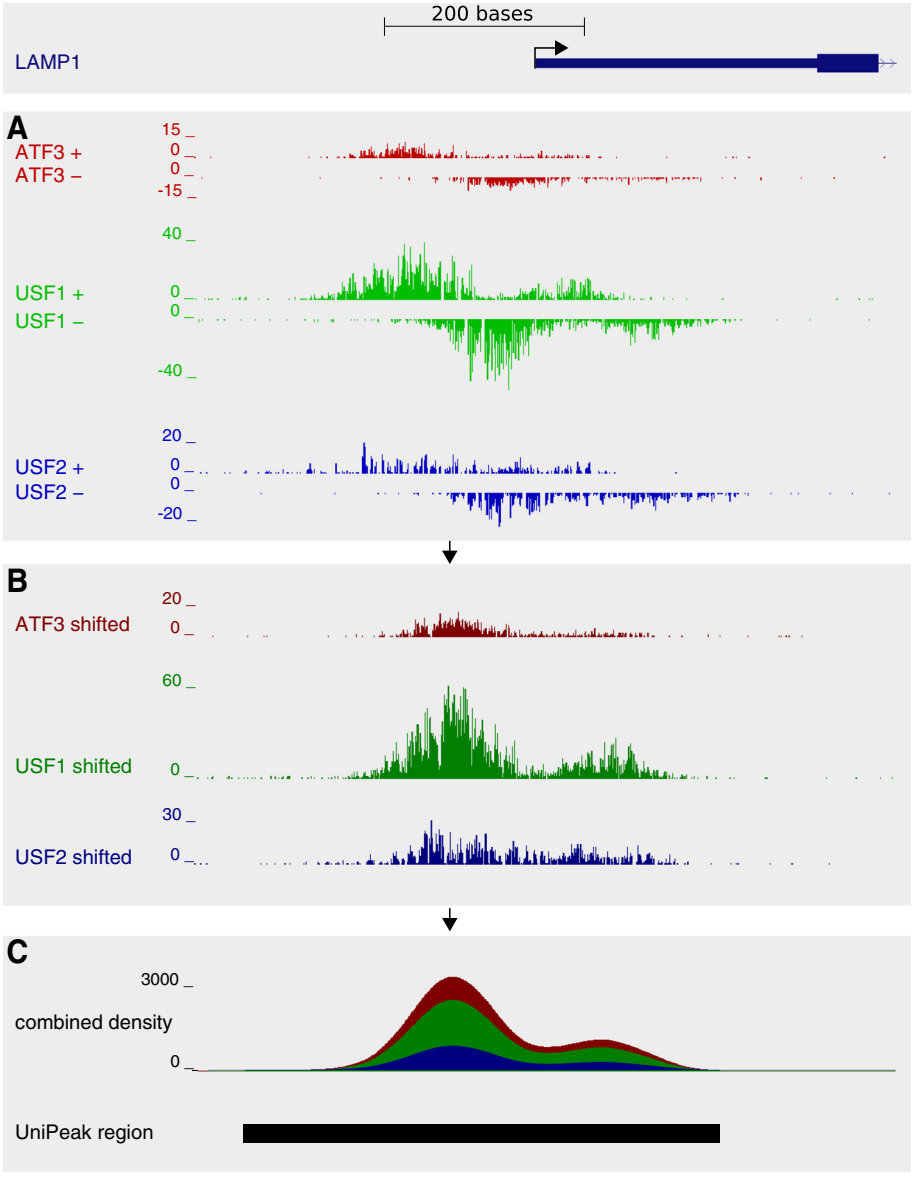

Figure 1 The `UniPeak` workflow. A: Sequence reads are considered as hits at their 5′ start positions, strand-specifically. **B**: A global read-shift value is computed independently for each sample to align forward and reverse reads. **C**: The shifted reads from all samples are then used to estimate a single underlying density profile. Enriched regions are identified where this density exceeds a fixed threshold, determined as a function of sequencing depth and genome size. Shifted reads from each sample are counted within these regions, providing a read count for each sample within each genomic region.

cell lines individually. This yielded 12,312–14,578 HOT regions from each data set, except H1-hESC with only 3,392 (Additional file 1: Figure S4). The generally higher number of detected regions may reflect higher sensitivity to cell-specific binding than in the pooled analysis, and a general lack of active cell-specific sites in H1-hESC (perhaps differentiated lineage-specific enhancers, since H1-hESC showed much higher promoter enrichment (50% consensus promoters vs. 22–39% in other cell types); this is consistent with a model in which tissue-specific enhancers are inactive or "poised" in undifferentiated cells [17]).

Many HOT regions are promoters

Since transcription factors occupy regulatory elements in the genome, we expected HOT regions to align with these elements. We compared the positions of these HOT regions with those of known or inferred promoters, according to three lines of evidence. First, we detected initiating RNA polymerase II (serine 5-phosphorylated

Table 1 Results of region calling

Data set	Proteins	Samples	Reads	`UniPeak` regions	Reads in regions	RNA polymerase II initiation sites	CAGE peaks	RefSeq TSS	Consensus promoters
All proteins	96	503	8.8B	11,239	213M (2%)	5,631 (50%)	4,951 (44%)	4,703 (42%)	4,189 (37%)
TFs only	75	357	6.3B	7,227	118M (2%)	5,745 (79%)	5,128 (71%)	4,813 (67%)	4,441 (61%)
GM12878	46	102	1.8B	12,887	61M (3%)	7,477 (58%)	6,315 (49%)	5,011 (39%)	4,522 (35%)
K562	41	93	1.5B	14,578	70M (5%)	10,174 (70%)	7,188 (49%)	6,589 (45%)	5,743 (39%)
HepG2	32	76	1.5B	12,312	48M (3%)	6,557 (53%)	5,232 (42%)	4,199 (34%)	3,791 (31%)
H1-hESC	25	52	985M	3,392	8M (1%)	2,180 (64%)	2,303 (68%)	2,127 (63%)	1,700 (50%)
HeLa-S3	16	34	498M	13,199	18M (4%)	5,499 (42%)	4,056 (31%)	3,243 (25%)	2,893 (22%)

[18]; Pol II-S5P) enrichment sites from an independent `UniPeak` analysis, again using ENCODE ChIP-seq data. Second, we used a strand-specific `UniPeak` analysis to detect enriched regions from CAGE, a form of RNA-seq that captures short tags at the 5′ end of the transcript [19]. Finally, we used transcription start site (TSS) positions from RefSeq [20], the most robust and most stringent annotation.

Of the 7,227 HOT regions called using the set of canonical TFs in all cells, 79% were within 500 bp of Pol II-S5P occupancy peaks, 71% within 500 bp of CAGE enrichment peaks, and 67% within 500 bp of RefSeq TSSs, with 61% "consensus promoters", i.e. within 500 bp of all three features (Figure 3A). Of HOT regions with occupancy peaks within 500 bp of one of these annotations, most fell within 200 bp of the given annotation (83% Pol II-S5P, 88% CAGE, 85% RefSeq TSS), with a bias toward being upstream rather than downstream of annotated TSSs (68% upstream; Figure 3B). Among the regions called in the five cell-specific analyses, 42–70% were near Pol II-S5P sites, 31–68% near CAGE peaks, and 25–63% near annotated TSSs (Table 1); the variation in these ranges reflects the different sets of TFs tested in the different cells. RefSeq TSS was consistently the least common annotation, perhaps because the database represents an incomplete set of true promoters, whereas Pol II-S5P ChIP-seq and CAGE enrichment signals occur at active TSSs regardless of whether they are annotated.

We further characterized HOT regions at consensus promoters in terms of several quantitative genomic features associated with promoters (Additional file 1: Figure S5). Most known human promoters are enriched for GC content and especially CpG dinucleotides [21], and their sequences are typically under evolutionary constraint [22]. In addition, several histone modifications are associated with regulatory genome elements: histone 3 lysine 4 trimethylation (H3K4me3) is enriched at active promoters [13], while monomethylation (H3K4me1) is enriched at enhancers [23], and histone 3 lysine 27 acetylation (H3K27ac) is enriched at both active promoters [24,25] and active enhancers [17,25,26]. Consistent with active promoters, our consensus promoters showed higher GC content, CpG content, evolutionary constraint, H3K27ac, and H3K4me3 vs. H3K4me1 (Figure 3C). Since these regions showed strong evidence of being promoters and could be associated with specific genes, we restricted all subsequent analyses to the consensus promoters in each of the five cell-specific HOT region lists, treating them as independent replicate experiments.

Similar occupancy profiles suggest binding partners

We reasoned that TFs with correlated occupancy profiles (more abundant at some sites and less abundant at others) may share mechanistic or functional relationships. To search for such relationships, we used neighbor-joining [27] to cluster TFs by the similarity of their occupancy profiles across consensus promoters in each cell (Figure 4). This analysis detected some known binding partners and gene families as well as novel relationships among TFs. For example, subunits of multimeric complexes often had very similar binding profiles, such as NFE2 and MAFF or MAFK [28]; MAX and MYC or MXI1 [29,30], NFYA and NFYB [31,32]; and USF1 and USF2 [33]; to a lesser extent, so did family members that share a DNA-binding motif, such as the ETS family (ELF1, ETS1, GABPA, SPI1) and the E-box family (MYC, USF1, USF2).

The AP-1 transcription factor is a heterodimer composed of a member of the JUN family and a member of either the FOS family or the ATF family [34]. However, in our analysis, FOS itself never clustered with a JUN family member, but JUN and JUND's occupancy profiles were correlated with those of their alternative binding partners BATF, FOSL1, FOSL2, and ATF2, and to a lesser extent CEBPB and CEBPD, which are not documented to interact with AP-1 subunits. Unlike ATF2, ATF3 reproducibly clustered with USF1 and USF2, which have no documented interaction with ATF3. The occupancy profiles of SIX5 and ZNF143 were also correlated in multiple cell

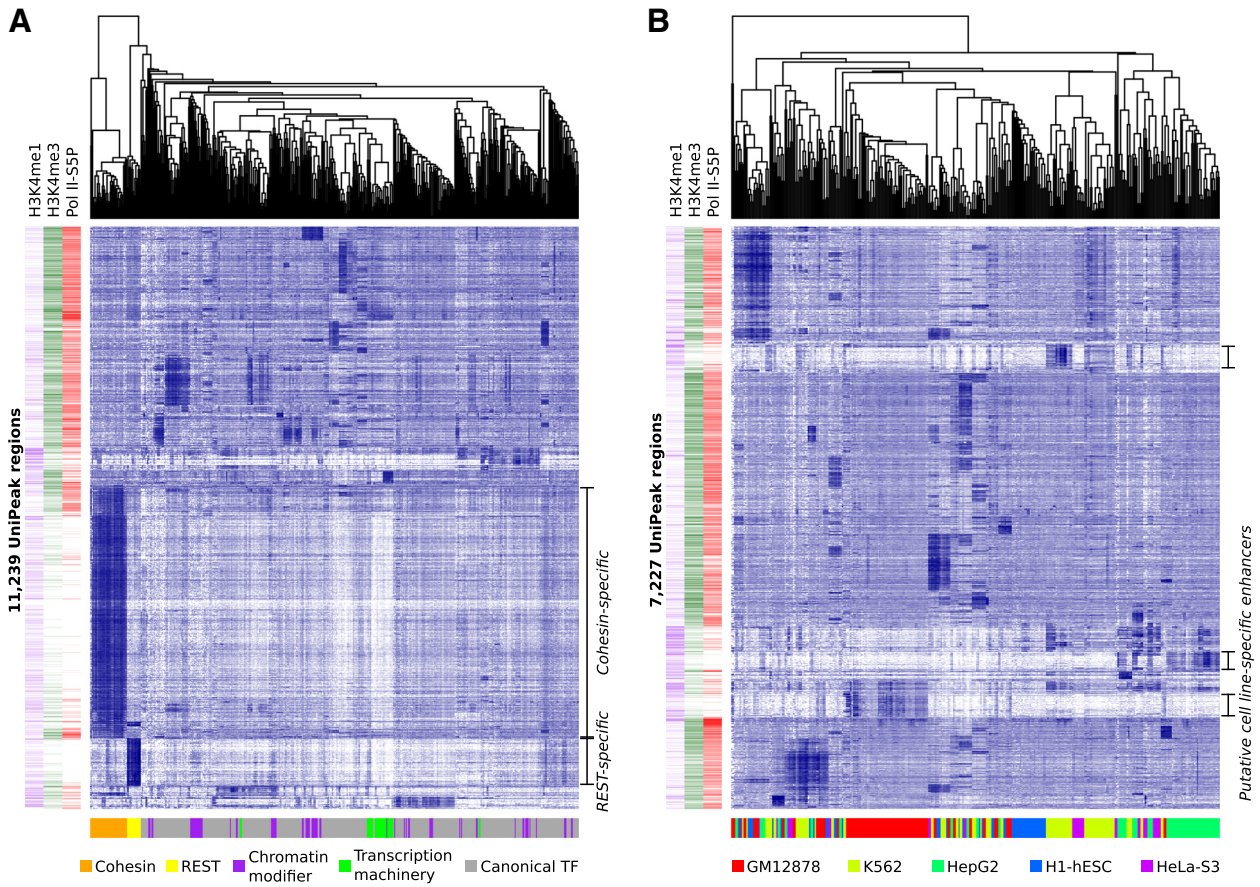

Figure 2 Clustering of protein occupancy data. A: Hierarchically clustered heatmap of all DNA-associated proteins assayed. Each row represents a HOT region called on the pooled data by `UniPeak` ($m = 11,239$), and each column represents a single replicate of a ChIP-seq experiment ($n = 503$). The sidebars show histone modification ChIP-seq signals within the same regions, normalized separately, and initiating RNA polymerase II signal and the nearest peak within 500 bp (zero if there were no peaks in this range). Color intensity in the heatmaps corresponds to variance-stabilized values for occupancy. **B**: Canonical TFs only (7,227 regions × 357 samples).

types despite no documented interaction. In fact, mammalian two-hybrid assays found no direct binding activity between these proteins [35].

One large branch in GM12878 included ATF2 [36], BATF [37], BCL3 [38], BCL11A [39], BCLAF1 [40], BHLHE40 [41], EBF1 [42], IRF4 [43], JUND [44], MEF2C [45], MTA3 [46], NFATC1 [47], PAX5 [48], PML [49], POU2F2 [50], RUNX3 [51], RXRA [52], SPI1 [53], STAT3 [54], STAT5A [55], TCF3 [56], and TCF12 [57], which are all known to be involved in the differentiation of lymphocyte lineages. This branch also included MEF2A, which, unlike its highly similar family member MEF2C, is not known to be involved in lymphocyte differentiation [58]. Thus, this analysis both recovered known functional relationships between TFs and discovered novel associations.

To test whether protein-protein interactions predict similarites in occupancy patterns, we compared our results with a comprehensive database of mammalian two-hybrid screens; data were available for all TFs in this study except FAM48A and THAP1 [35] Within each cell type, we split pairwise correlations of samples' occupancy profiles across all HOT regions into those from binding TF pairs and those from non-binding TF pairs. Pairs of replicates of the same TF were not used. On average, the occupancy profiles of binding TFs were more correlated than those of non-binding TFs (Additional file 1: Figure S6). The difference was only large in the HeLa-S3 data, perhaps due to the selection of TFs tested in that cell type; in other words, potential direct interactions between TF pairs (which may not actually occur *in vivo*) generally only explain a small part of the similarity in their occupancy patterns.

Most TFs appear to be recruited to HOT regions as cofactors

Although it is difficult to use shared occupancy profiles to infer a binding mechanism, additional analysis can illuminate a critical step in the recruitment of a TF

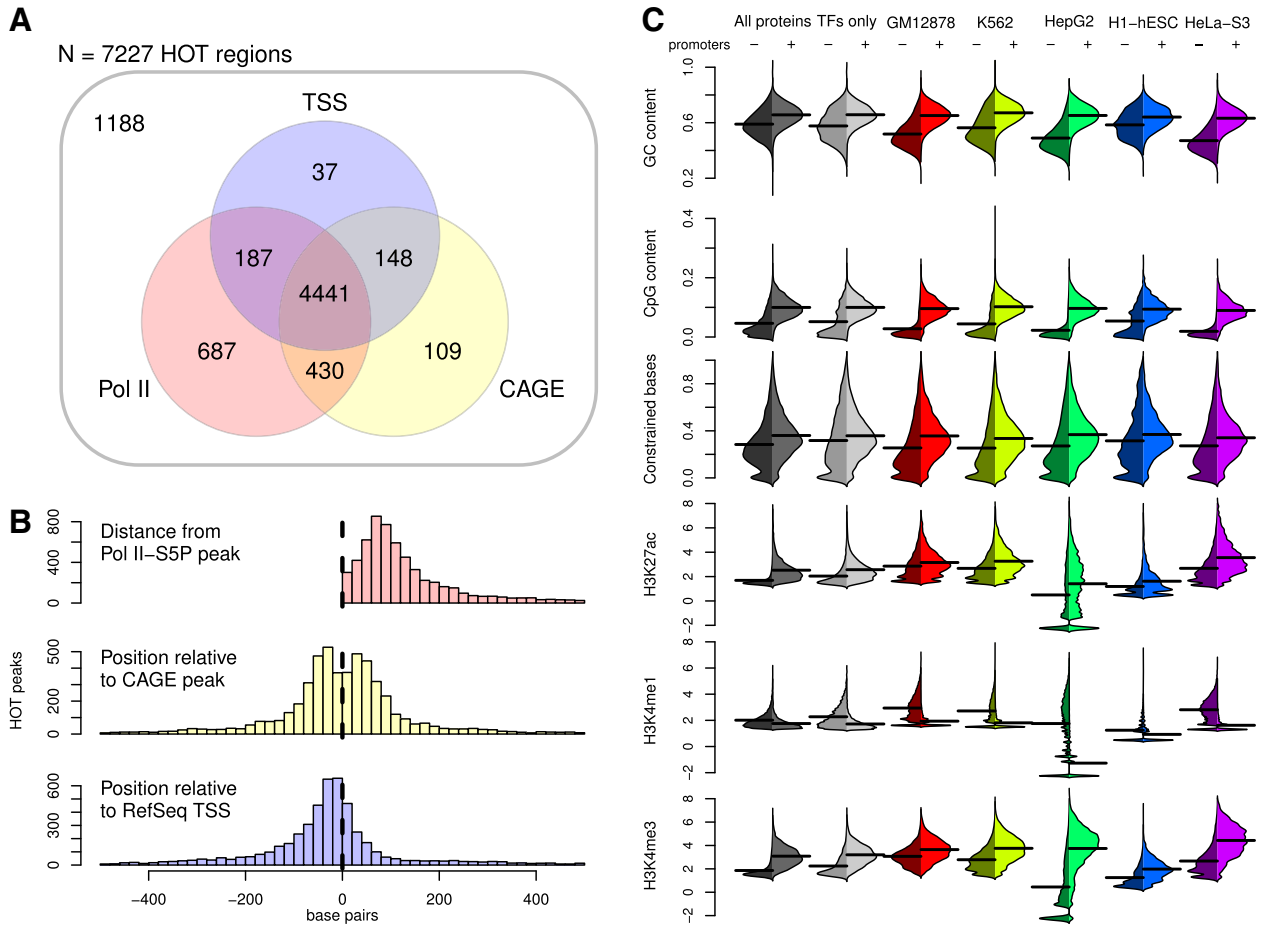

Figure 3 Characterization of HOT regions with reference annotations. A: Overlap among promoter annotations. Each HOT region from the analysis of all canonical TFs in all cells ($N = 7,227$) was annotated with the nearest initiating RNA polymerase II peak, CAGE peak, or RefSeq TSS within 500 bp. **B**: Proximity of TFs' binding regions to promoters, in base pairs, according to nondirectional RNA polymerase II ChIP-seq data, directional CAGE data, or RefSeq annotations of transcription start sites. **C**: Quantitative annotations versus presence of Pol II. Each row of figures shows distributions of a given annotation in the two combined and five cell-specific data sets, separated between HOT regions not at a consensus promoter (dark) vs. regions within 500 bp of a Pol II binding site, a CAGE peak, and a RefSeq TSS (light).

complex. A TF's observed occupancy at a given promoter might be due to either direct binding of DNA or recruitment by another protein. Most TFs in our data set have previously been annotated with DNA sequence motifs that they bind specifically. Thus, if we make the simplifying assumption that most TFs usually bind DNA at regions that contain their respective sequence motifs, then their occupancy at sites without their motifs is likely to be as cofactors recruited by other proteins.

To identify candidates for direct DNA binding, we searched across the consensus promoters from the cell-specific `UniPeak` output for occurrences of sequence motifs associated with the TFs in the ENCODE data set. We considered motifs identified *de novo* by ENCODE from analysis of each ChIP-seq experiment individually (Kheradpour P, Kellis M: ENCODE-motifs: systematic analysis of regulatory motifs associated with transcription factor binding in the human genome, under revision), and in order to avoid motifs that are not bound directly by a given TF but rather by its cofactor, we used only motifs that matched database annotations for the given TF. This yielded multiple motifs for some TFs and none for others, and some motifs were associated with TF families rather than individual TFs; thus, our analysis was based on "motif sets" that share a common annotation, rather than individual motifs. On average, any given HOT promoter contained motifs in about 4 distinct sets (Figure 5A), even though these sites are defined by the presence of many more TFs, suggesting that the majority of TFs at these sites may be recruited by other factors rather than bound to the promoter themselves. Furthermore, the number of occurrences of any set's motifs across the set of regions was too small for nearly any

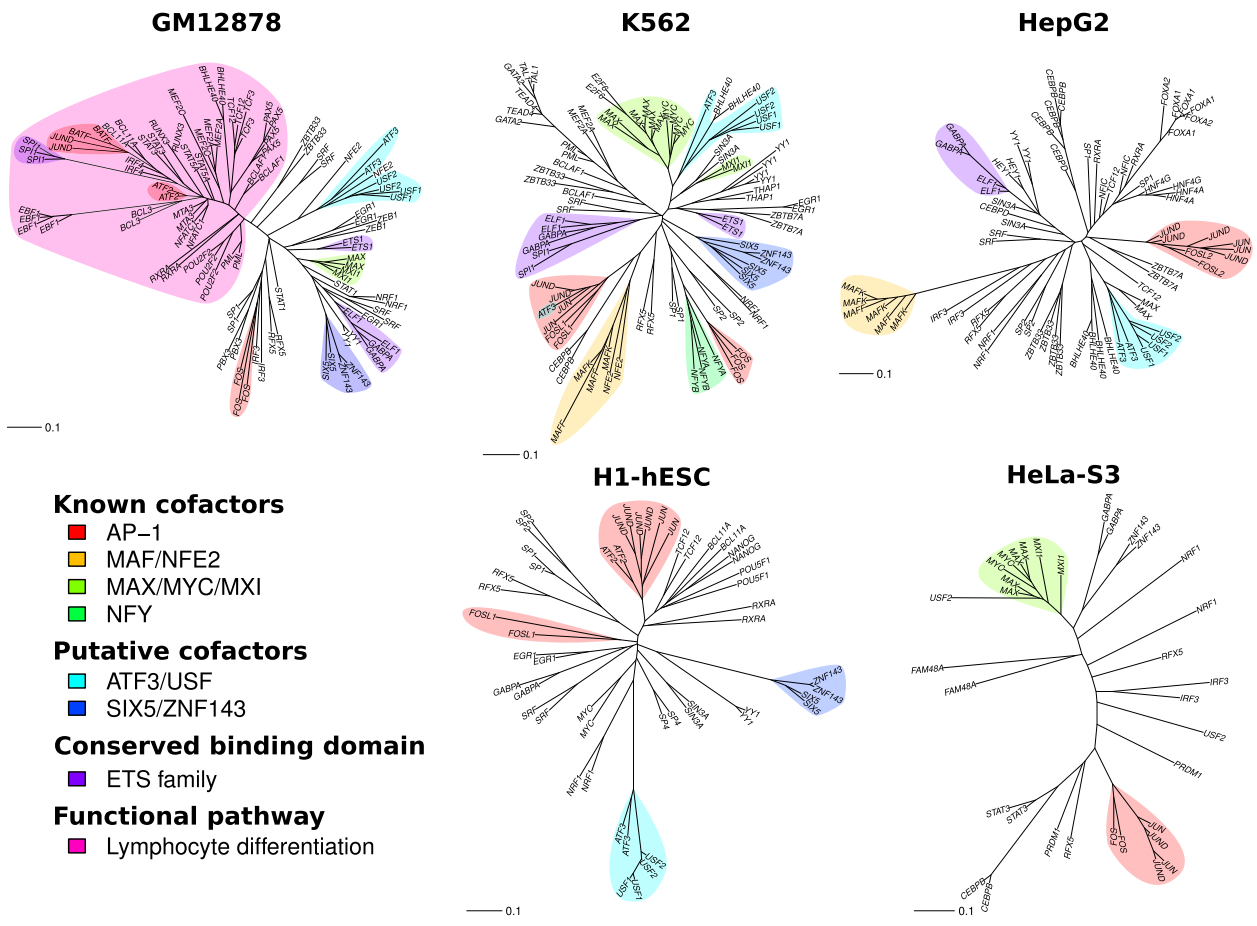

Figure 4 Neighbor-joining trees for TFs at consensus promoters in the cell-specific analyses. Clustering is by Pearson distance of occupancy profiles. Individual replicates are shown separately. Known and putative functional groups are highlighted.

particular TF to bind its motif at most HOT regions (Figure 5B).

Next we measured the association between motifs and TF occupancy at consensus promoters. For each motif, we compared the occupancy score of each TF between consensus promoters with the motif vs. those without it Figure 6, (Additional file 1: Figure S7). Many motif sets were predictive of the occupancy of other TFs besides their own, and some TFs' occupancy was better predicted by other TFs' motifs than by their own. The most promiment pattern was that ETS-family motifs were strongly predictive of many other TFs' occupancy. In particular, consensus promoters with ETS motifs were enriched for the occupancy of MAFF, MAFK, MEF2A, MEF2C, POU2F2, and SRF, suggesting that these TFs' primary mechanism of positioning at HOT regions may be recruitment by ETS family members rather than direct DNA binding.

Motifs in the E-box family are bound by TFs with a basic helix-loop-helix domain, including BHLHE40 [59], MYC-MAX [29], MXI1-MAX [30], TCF12 [60], and USF1/USF2 [33]. Other TFs enriched at E-box sites included ATF3, E2F6, NFE2, and SIN3A; of these, the only previously documented interaction with a E-box-binding TF is between SIN3A and MAD-MAX [61].

Subunits of the AP-1 transcription factor were only weakly enriched at promoters containing motifs for the FOS-JUN heterodimer; however, JUN and JUND, but not FOS, were more strongly enriched at sites with motifs associated with their alternate binding partner ATF3. The ATF3 motifs were also predictive of the occupancy of CEBPB, RFX5, and SRF, none of which are documented to interact with AP-1 directly; however, although neither is enriched at the other's motif sites, CEBPB and SRF are known binding partners [62,63]. On the other hand, FOS, but not FOSL1 or FOSL2, was very strongly enriched at sites with the NF-Y motifs, as were IRF3 (but not IRF4), PBX3, RFX5, SP1, and SP2. Of these, only SP1 is known to interact with NFYA/NFYB [64].

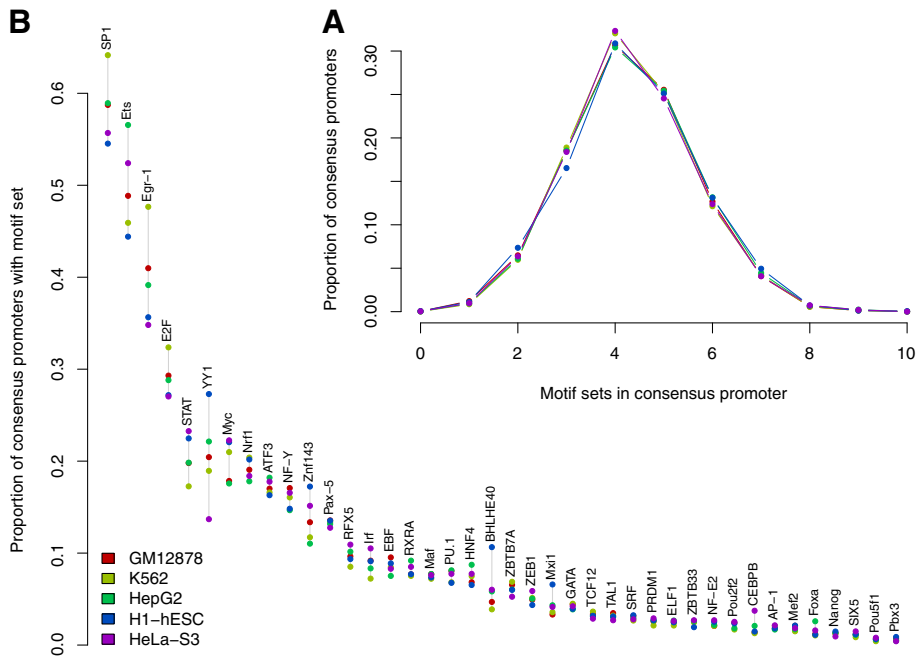

Figure 5 Frequency of DNA sequence-motif sets. Each set contains one or more motifs annotated to the same TF. **A**: Number of distinct motif sets represented at each consensus promoter peak. **B**. For each motif set, the number of consensus promoter peaks with at least one motif from the set.

Other relationships between a motif set and TFs not annotated with it include one particular MAF motif with SPI1; the NRF1 motifs with ATF3; certain STAT motifs with ELF1, ETS1, SIX5, SPI1, and ZNF143; TAL1 motif with TCF3 and TCF12; and ZNF143 motif with ETS1 and SIX5. Of these relationships, all but the last can be explained by motif sequence similarity; no interaction among ZNF143, ETS1, and SIX5 is documented. Some of the most common motifs, the GC-rich EGR1 and SP1 sets, were associated with depletion of most TFs. The NRF1 and NF-Y motifs were associated with depletion of many TFs except the few that were strongly enriched at those sites.

The occurrence of TF-associated DNA sequence motifs in HOT regions was so low, relative to the number of TFs present, that most TFs probably do not directly bind the DNA at these regions but are instead recruited as cofactors, consistent with other analyses of these data [65]. Reinforcing this, many TFs' occupancy was well predicted by motifs known to be bound directly by different TFs, and in some cases a TF showed a stronger preference for a different TF's motif than for its own.

These results are corroborated by a previous analysis of the same data [66]; however, most of the putative transcription-factor interactions inferred in that analysis are not supported by ours. Our analysis may be more stringent because it considers the strength of the ChIP-seq signal at each site rather than just presence or absence of a peak called at arbitrary thresholds.

A small number of TFs explain a large proportion of Pol II recruitment

The general role of TFs is to recruit the preinitiation complex and ultimately Pol II, which then transcribes RNA from the gene body; thus, the presence of these downstream factors and the abundance of the transcript should be partially explained by the combination of TFs at promoters. We also expect a relationship between TF occupancy and histone modifications associated with active promoters, though the causality may work in either direction. Since we have quantitative enrichment values for all these markers of gene regulation and for all TFs' occupancy, at all consensus promoters, we can measure the strength of the relationship between them statistically.

We constructed a linear regression model that treated each TF ChIP-seq sample as an independent variable, and gene regulation as the dependent variable, with each HOT consensus promoter as one observation. This model necessarily contains redundant signals, not just as strong correlations between replicates, but also as weaker correlations between factors with similar behaviors, such as sets of TFs that bind in complex; because of the number of predictors and their nested multicollinearity, standard multiple linear regression would produce uninterpretable results and suffer from overfitting or reduced power.

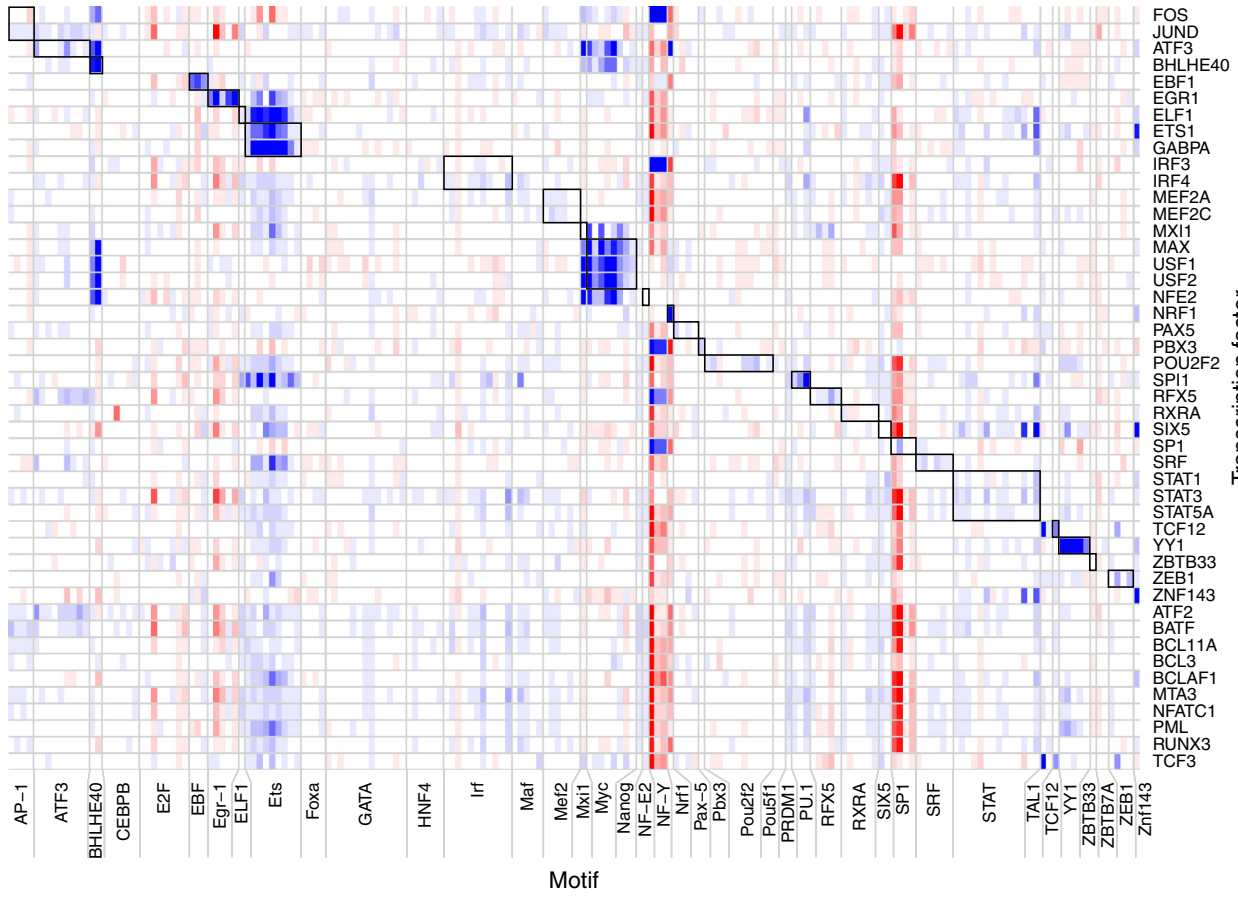

Figure 6 Predictive value of all motifs for all TFs, in GM12878 cells. Each row represents a single TF, with replicates pooled. Each column represents one sequence motif from the previous analysis, and motif sets associated with the same TF are grouped. Color intensity represents the significance of the difference in occupancy of the given TF at consensus promoters with vs. without the given motif, where significance is the logarithm of the *p*-value from a *t*-test for the difference between means of occupancy scores, which are the means of variance-stabilized scores across replicates. Blue signifies enrichment of the given TF at sites with the given motif, and red signifies depletion of the given TF at sites with the given motif.

We instead applied partial least-squares regression, which performs a rotation and dimensional reduction on the covariance matrix in order to isolate latent orthogonal signals underlying patterns from multiple observations. This method also allows both the independent and dependent variables to be matrices rather than single vectors, so only two models (v.i.) were fit for each cell type, encompassing all available data at once. The dependent variables we used were the occupancy of PIC subunits within the region, the enrichment of histone modifications (H3K4me1, H3K4me3, H3K27ac) within the region, the occupancy of initiating Pol II at its nearest enriched region, the occupancy of elongating Pol II (serine 2-phosphorylated [18]; Pol II-S2P) in the gene body corresponding to the nearest RefSeq TSS, the CAGE signal at its nearest enriched region, and the RNA-seq signal for the gene corresponding to the nearest RefSeq TSS. Since all the signals from the available experiments were required for a full observation, this analysis was restricted to consensus promoters; genes with no TFs bound were not used to train the model.

As a null model, we considered that any explanatory power from the TF signals that could also be contributed from "input" controls (total chromatin, IgG pulldown) was likely a ChIP-seq artifact rather than a meaningful TF effect. Therefore, for each cell type we compared two models: gene regulation as a function of both TF ChIP and input signal, and gene regulation as a function of input signal alone.

The presence of PIC subunits was well predicted by aggregated TF occupancy (Figure 7, Additional file 1: Figure S10; cross-validation $R^2 \approx 0.7$ for the cells with the most TFs tested), though with somewhat high contribution from input alone. Histone marks H3K4me3 and H3K27ac were somewhat well predicted (CV $R^2 \approx 0.4$), but with even higher relative contribution from input,

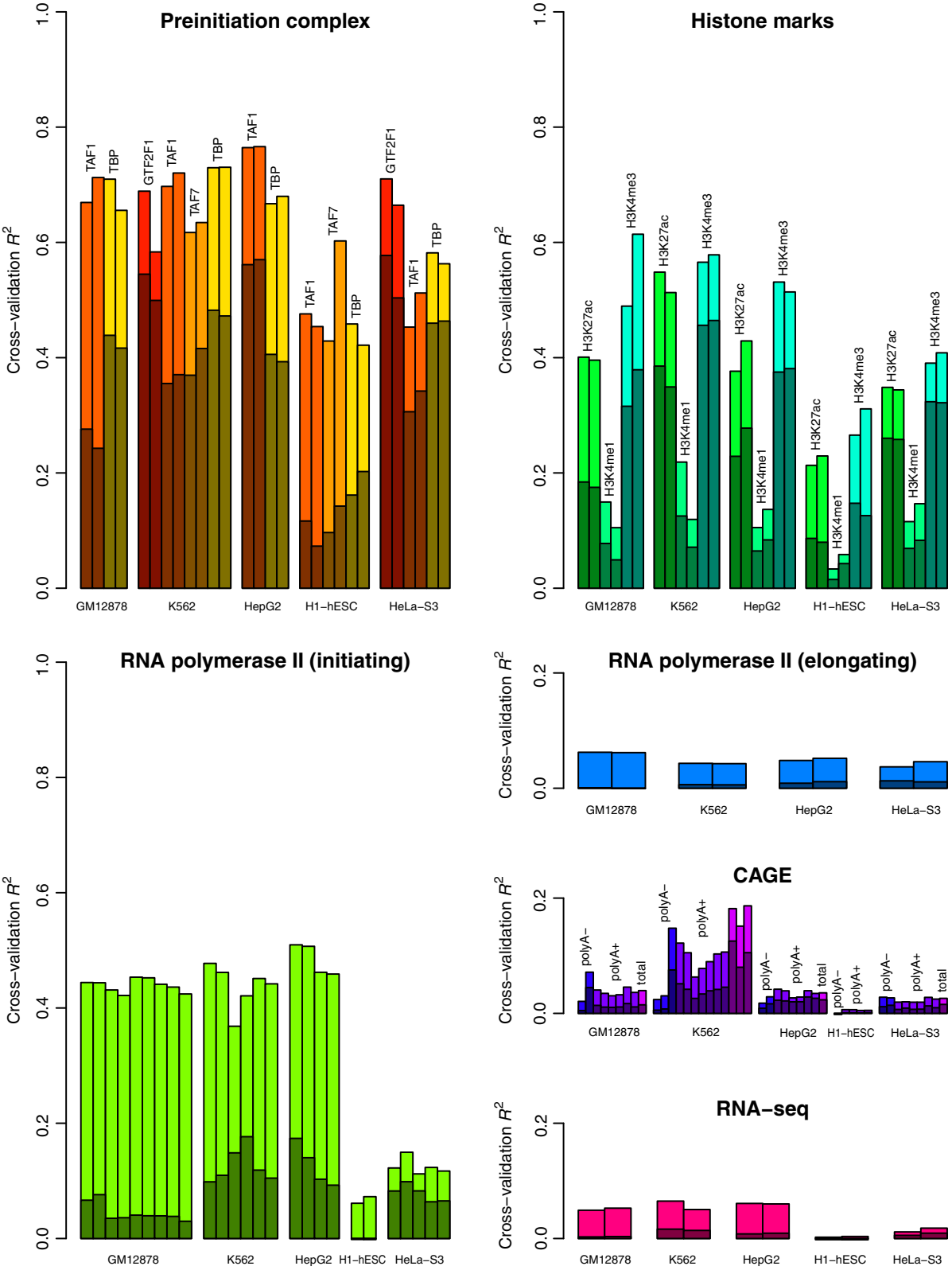

Figure 7 Functional analysis of TF binding. For each cell type, a model was fit to predict various measures of gene regulation as a function of the occupancy of all tested TFs. Cross-validation R^2 values are shown separately for each replicate. Light-colored bars correspond to the full $\sum$TFs + input models, dark-colored bars to the input-only models.

perhaps because these controls are sensitive to open chromatin, which is associated with active promoters; H3K4me1 was not well predicted by the model (CV $R^2 < 0.2$), likely because of very low signal at these regions, as expected for a mark depleted at active promoters. Pol II-S5P occupancy was also well predicted by TF occupancy (CV $R^2 \approx 0.4$), and input was not very predictive (CV $R^2 < 0.2$); the results were slightly worse in the cells with fewer TFs tested. On the other hand, Pol II-S2P occupancy was not well predicted by TF occupancy, nor was transcript abundance as measured by either CAGE or RNA-seq (CV $R^2 < 0.2$); there was no consistent difference between CAGE signals from polyadenylated (mature) and unpolyadenylated transcripts. Thus we found that the presence of these TFs is strongly associated with immediately subsequent steps in gene regulation, but only weakly associated with later steps.

Conclusions

We present a quantitative analysis of a large volume of ChIP-seq data, constituting the genome-wide occupancy profiles of a large number of TFs in five human cell types, from the ENCODE consortium [8]. The new software package `UniPeak` facilitates the comparison of binding profiles from an unlimited number of samples at a consistent set of genome regions, eliminating the difficulty of reconciling many independent lists of peak calls and producing a regions × samples matrix of signal strengths, similar to those generated by microarray experiments. Here we bring matrix analysis and sample clustering back to the forefront of a high-throughput genomics investigation. Since we view DNA-associated protein occupancy as a fundamentally quantitative phenomenon, which may have quantitative functional effects [2], we avoid applying premature thresholds and dequantification of the peak intensities. Our approach may become even more useful as improved technology allows greater sequencing depths and therefore higher quantitative precision, and perhaps also as new molecular protocols increase the signal-to-noise ratio of protein-associated DNA capture [67].

Assessing the relevance of this study to our understanding of transcriptional regulation, we found that about 40% of variance in initiating Pol II occupancy at HOT promoters can be explained by the entire set of available TF occupancy data in the cells with the most experiments. The predictive value is higher for PIC subunits, and much lower for elongating Pol II and transcript abundance. These results are also consistent with our knowledge of biological mechanisms, because there are many additional regulatory interactions between PIC recruitment and the production of an elongated, mature, stable transcript that do not involve TFs. It is important to note that these models would have shown a better fit if we had surveyed all promoters instead of just those occupied by many TFs, because the inclusion of inactive promoters would add many points near the origin (no TFs bound, no gene expression), which would make the trend more linear [7,68-70]. Finally, this analysis represents fewer than 50 TFs tested in any individual cell line, compared with the 1,400–1,900 TFs estimated to exist in the human genome [71]; in that context, 40% of variance explained represents substantial explanatory power.

Regions occupied by many different TFs are common in the human genome. Even our strictest definition finds several thousand HOT promoters, likely a considerable fraction of the active genes in any given cell line. Especially since there are far too few known DNA sequence motifs to account for all the TF occupancy at these sites, we propose that TFs collaborate combinatorially through protein-protein interactions to regulate Pol II recruitment (Figure 8), concordant with similar evidence from *Drosophila* enhancers [72]. Interactions of this nature have not previously been examined on such a large scale, due to the greater challenges of high-throughput peptide assays compared to high-throughput nucleotide assays.

This analysis yields several hypotheses that may be validated by future experiments. Based on the similarity of their occupancy profiles, reproduced independently in multiple cell lines, we infer that ATF3 and USF1/USF2 may be part of a novel protein complex; furthermore, since DNA sequence motifs associated with USF1/2 are predictive of ATF3's occupancy but not vice versa, we predict that USF1 or USF2 is the subunit of this complex that directly binds the promoter, while ATF3 is a cofactor recruited by USF1/2 despite having its own DNA-binding domain that is used in other complexes. By the same logic we also predict that SIX5 and ZNF143 are members of a novel complex in which ZNF143 is the DNA-binding subunit. In both cases, previous experimental evidence shows the two partners are not capable of binding each other alone, suggesting that a chaperone is required to enable binding, or that these interactions require both proteins to bind to a common intermediate or complex of intermediates. On the other hand, we find that in HOT promoters, FOS seems only rarely to perform its well-known role in the AP-1 heterodimer with JUN, generally supplanted by a FOSL or ATF protein, though of course FOS is known to have several alternate cofactors [34].

It should not be surprising that a TF with a functional DNA-binding domain and even a well-demonstrated sequence motif might often be recruited as a cofactor by some other protein. Indeed, this is true of TBP, the so-called TATA-binding protein, which is required for the assembly of the preinitiation complex at all

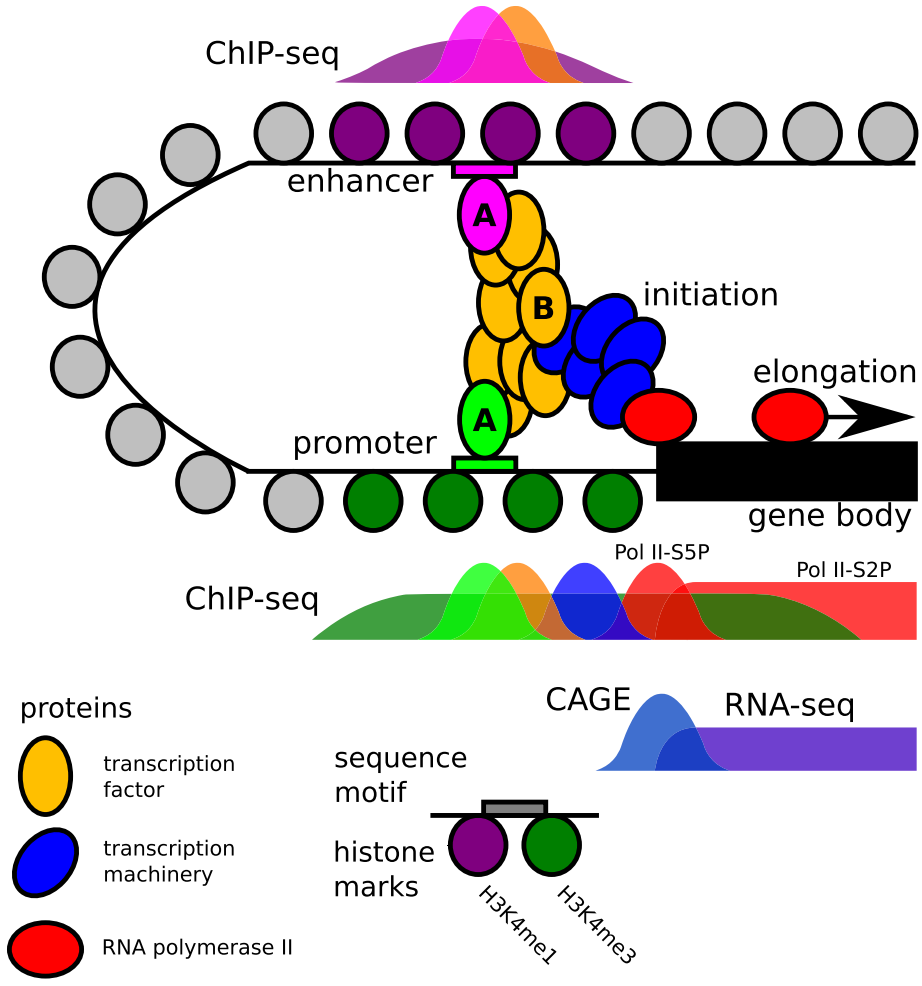

Figure 8 Proposed model of interaction between two HOT regions and resulting signals from high-throughput sequencing experiments. An enhancer is looped over and connected by a complex of many TFs, which recruit the preinitiation complex and ultimately RNA polymerase II to the TSS. Proteins A are directly bound to DNA, while protein B is recruited to the site through only protein-protein interactions. These mechanisms are indistinguishable from the ChIP-seq signal, but presence of a known sequence motif suggests which protein binds directly.

loci even though only 10–24% of human promoters have a TATA box [73,74]. One possible paradigm for gene-regulatory evolution might be the emergence of a DNA-binding TF that uses protein-protein interactions to recruit other TFs to its own recognition sites, harnessing their existing regulatory pathways without sequence motifs for the other TFs. Over evolutionary time, this additional layer of regulatory interactions between the steps of protein-DNA binding and recruitment of polymerase might remove the constraint of requiring the "downstream" TFs' sequence motifs in new regulatory elements or even conserving them in existing elements, so that TFs capable of autonomously binding DNA and recruiting the PIC become primarily cofactors for other TFs with more specialized target loci and finer regulatory control. Thus the large TF complexes, or interchangeable interactions, that we observe at HOT regions might represent multiple levels of gene regulation and therefore of evolutionary history.

Materials and methods

Read alignment

ChIP-seq nucleotide sequence reads and base qualities were obtained from the ENCODE database, and truncated to the first 25 nt, the shortest length in the data set, to prevent biases in mapping due to different read lengths. `BWA 0.6.1-r104` [75] was used to map reads to the hg19 reference assembly. Unique best hits were filtered to confident alignments with posterior probability ≥ 0.9.

Density profiles

Similarly to the robust `QuEST` algorithm [10], a smooth density profile was created using the frequency of 5′ read

starts per reference base as the input to kernel density estimation (KDE), so that the density at any given position i on one strand was given by

$$H(i) = \frac{\sum_{j=i-h}^{i+h} K\left(\frac{i-j}{h}\right) C(j)}{\sum_{k=-h}^{h} K(\frac{k}{h})}$$

where $K(x) = \frac{3}{4}(1 - x^2)\mathbf{1}_{\{|x|\leq 1\}}$ is the Epanechnikov (quadratic) kernel density function, h is the kernel bandwidth, and $C(j)$ gives the number of 5′ read starts at position j.

Enriched region calling

Any region where the smooth density profile exceeded a fixed threshold, relative to the uniform background of the total confident read count divided by the genome size, was considered enriched. 5′ read starts were then counted inside each region. The kurtosis of the distribution of 5′ read starts within each region was calculated, and leptokurtic regions were filtered out to remove technical artifacts.

Strand shift estimation

To estimate the shift between enrichment maxima from the forward and reverse strands flanking each binding site, a byproduct of 5′ end-directed sequencing and the genomic fragment size, KDE was performed separately on each strand, and preliminary enriched regions were called from the sum of the two density profiles. Among the regions containing the highest read counts, the Pearson correlation between the strand-specific density profiles was calculated for each of a spectrum of 5′ to 3′ shift values. The distribution of correlation-maximizing shift values across the top regions was smoothened with a small bandwidth and the global maximum was chosen as the sample-wide shift value. Density profiles from opposite strands were shifted by this value and added together for a unified, strand-independent profile. Regions with a low Pearson correlation between the two strands' density profiles were discarded as artifacts.

UniPeak workflow

The new software package `UniPeak` was written to automate the steps above. Starting with confidently aligned reads, strand shift was estimated independently for each sample (with the exception of negative controls, whose shift was inferred from corresponding ChIP samples, as they did not yield enough preliminary enriched regions to estimate a shift value), using the top 1,000 regions called with smoothing bandwidth 50 nt, region-calling fold-enrichment threshold 25X, kurtosis threshold 50, minimum strand correlation 0.3, minimum shift 25 nt each strand, maximum shift 150 nt each strand, and correlation vs. shift smoothing bandwidth 5 nt. The samples were then shifted accordingly and kernel smoothing was performed with bandwidth 100 nt to capture binding sites in close proximity to each other; density profiles from both strands of all samples were summed and enriched regions were called and filtered as before. Enriched regions on sex chromosomes and the mitochondrial genome were removed, along with those overlapping false-positive genome regions identified by ENCODE and those greater than 500 bp in size.

Normalization

The read-count matrix from `UniPeak 1.0` was normalized by the variance-stabilizing transformation in `DESeq 1.7.7` [11], determining the dispersion-mean relationship with local fitting, pooling all samples to estimate a single empirical dispersion value per analysis, and using only the fitted dispersion-mean relationship values. Replicate experiments from different laboratories were treated as separate classes.

Clustering analysis

Clustering was performed on normalized data as described above. Distances were calculated with the Pearson metric $(1 - r)$. Rooted, ultrametric trees were generated by hierarchical clustering with UPGMA as implemented in the `fastcluster 1.0.4` package in `R 2.12.1` [76]. Unrooted trees were generated by neighbor-joining [27] as implemented in `RapidNJ 2.1.0` [77].

Comparison with annotations and independent data

Initiating RNA polymerase II ChIP-seq data were treated in the same manner as TF data, but independently from that analysis, with 50 nt smoothing bandwidth for region calling and no region size filter. A TF-enriched region was matched to a Pol II-S5P-enriched region if the maxima of the regions' respective density profiles were within 500 bp of each other; when more than one Pol II-S5P site was near a TF site, the nearest Pol II-S5P site was used.

Transcription start site coordinates for the `hg19` reference assembly were obtained from the RefSeq database [20]. A TF-enriched region was matched to a RefSeq TSS if the TSS was within 500 bp of the local maximum of the density profile within the region; when more than one TSS fell within this range, the nearest was used. Single-end, 75 nt RNA-seq reads from the ENCODE database were aligned to the `hg19` RefSeq transcriptome by DNAnexus, which computed the count per transcript [78]. Elongating RNA polymerase II ChIP-seq reads were aligned to the `hg19` genome, and for each annotated TSS, reads were

counted between 100 bp upstream of the TSS and 100 bp downstream of the TES for the longest isoform.

CAGE reads were obtained from the ENCODE database after alignment to `hg19` with `Delve` (T Lassmann, in prep.). CAGE-enriched regions were called via `UniPeak` in the same manner as TF binding sites, using 50 nt smoothing bandwidth, separate strands, and no shifting. CAGE regions were matched with TF regions in the same manner as Pol II-S5P regions.

Evolutionary constraint within a region was calculated as the proportion of positions with a rejected substitution (RS) score greater than 2, according to `GERP++` [22].

From ENCODE's database we retrieved 226 sequence motifs that were both inferred *de novo* from ChIP-seq data and matched to similar motifs in other databases, such that there was a variably sized set of motifs annotated to each individual TF. These were aligned to reference sequence in a 201 bp window centered at each HOT region peak by `MAST 4.6.0` [79]. A HOT region peak was considered to have a hit for a given TF's motif set if any motif in the set had a `MAST` hit of $E < 10$.

Modeling gene regulation

For each cell line, we constructed a model of the general form

$$\sum Y_{\text{PIC}} + \sum Y_{\text{histone}} + \sum Y_{\text{pol2}} + \sum Y_{\text{CAGE}} + \sum Y_{\text{RNA-seq}} \sim \sum X_{\text{input}} + \sum X_{\text{TF}}$$

where the Y terms form a matrix of the individual replicates of the dependent variables, normalized together by `DESeq` as before: preinitiation-complex occupancy within the region, histone-mark occupancy within the region, Pol II occupancy at the nearest `UniPeak` site, CAGE signal at the nearest `UniPeak` site, and RNA-seq signal for the gene corresponding to the nearest RefSeq TSS; and the X terms form the matrix of all the individual replicates of TF occupancy scores plus the signal from negative-control samples (input, IgG, reverse-crosslinked chromatin) within the regions, normalized together. Since both the independent and depent variables were highly multicollinear, we used the `pls 2.3-0` package in `R 2.12.1` [80] to reduce this model to latent variables by partial least-squares regression. The number of LVs used in each model was determined as the first LV plus all subsequent LVs that subtracted at least 0.01 from the average RMSEP (Additional file 1: Figures S8, S9). Cross-validation used the leave-one-out method: the R^2 values were calculated by validating with each `UniPeak` region after re-training the model on the remainder of the data.

Additional files

Additional file 1: Contains a supplementary table and supplementary figures. All data used in this work are available from the ENCODE data portal, http://genome.ucsc.edu/ENCODE/. Scripts and processed data can be obtained at http://mendel.stanford.edu/sidowlab/downloads/hot/.

Additional file 2: Describes a study validating the `UniPeak` method.

Competing interests

The authors declare no conflict of interest.

Authors' contributions

JWF and AS conceived the study and prepared the manuscript. JWF performed all analysis. Both authors read and approved the final manuscript.

Acknowledgements

We thank Cheryl Smith, Ed Grow, Noah Spies, and Erik Lehnert for critical reading of this manuscript. We are also grateful to Anshul Kundaje and Phil Lacroute for invaluable technical advice. This work was supported by the Stanford Genome Training Program (NIH/NHGRI T32 HG000044), a subcontract to ENCODE grant HG004695, and a KAUST AEA grant.

Author details

[1]Department of Genetics, Stanford University, 300 Pasteur Drive, Stanford, California 94305, USA. [2]Current address: Douglas Mental Health University Institute, McGill University, 6875 Boulevard LaSalle, Verdun, Québec H4H 1R3, Canada. [3]Department of Pathology, Stanford University, 300 Pasteur Drive, Stanford, California 94305, USA.

Received: 13 May 2013 Accepted: 4 October 2013
Published: 20 October 2013

References

1. Moorman C, Sun LV, Wang J, de Wit E, Talhout W, Ward LD, Greil F, Lu XJ, White KP, Bussemaker HJ, van Steensel B: **Hotspots of transcription factor colocalization in the genome of Drosophila melanogaster.** *P Natl Acad Sci USA* 2006, **103**(32):12027–12032. [http://dx.doi.org/10.1073/pnas.0605003103]
2. MacArthur S, Li XY, Li J, Brown J, Chu HC, Zeng L, Grondona B, Hechmer A, Simirenko L, Keranen S, Knowles D, Stapleton M, Bickel P, Biggin M, Eisen M: **Developmental roles of 21 Drosophila transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions.** *Genome Biol* 2009, **10**(7):R80+. [http://dx.doi.org/10.1186/gb-2009-10-7-r80]
3. Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, Liu T, Yip KY, Robilotto R, Rechtsteiner A, Ikegami K, Alves P, Chateigner A, Perry M, Morris M, Auerbach RK, Feng X, Leng J, Vielle A, Niu W, Rhrissorrakrai K, Agarwal A, Alexander RP, Barber G, Brdlik CM, Brennan J, Brouillet JJ, Carr A, Cheung MS, Clawson H, Contrino S, et al.: **Integrative analysis of the Caenorhabditis elegans genome by the modENCODE Project.** *Science* 2010, **330**(6012):1775–1787. [http://dx.doi.org/10.1126/science.1196914]
4. Roy S, Ernst J, Kharchenko PV, Kheradpour P, Negre N, Eaton ML, Landolin JM, Bristow CA, Ma L, Lin MF, Washietl S, Arshinoff BI, Ay F, Meyer PE, Robine N, Washington NL, Di Stefano L, Berezikov E, Brown CD, Candeias R, Carlson JW, Carr A, Jungreis I, Marbach D, Sealfon R, Tolstorukov MY, Will S, Alekseyenko AA, Artieri C, et al.: **Identification of functional elements and regulatory circuits by Drosophila modENCODE.** *Science* 2010, **330**(6012):1787–1797. [http://dx.doi.org/10.1126/science.1198374]
5. Negre N, Brown CD, Ma L, Bristow CA, Miller SW, Wagner U, Kheradpour P, Eaton ML, Loriaux P, Sealfon R, Li Z, Ishii H, Spokony RF, Chen J, Hwang L, Cheng C, Auburn RP, Davis MB, Domanus M, Shah PK, Morrison CA, Zieba J, Suchy S, Senderowicz L, Victorsen A, Bild NA, Grundstad AJ, Hanley D, MacAlpine DM, Mannervik M, Venken K, et al.: **A cis-regulatory map of the Drosophila genome.** *Nature* 2011, **471**(7339):527–531. [http://dx.doi.org/10.1038/nature09990]
6. Kvon EZ, Stampfel G, Yáñez Cuna, J O, Dickson BJ, Stark A: **HOT regions function as patterned developmental enhancers and have a distinct cis-regulatory signature.** *Gene Dev* 2012, **26**(9):908–913. [http://dx.doi.org/10.1101/gad.188052112]

7. Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, Epstein CB, Frietze S, Harrow J, Kaul R, Khatun J, Lajoie BR, Landt SG, Lee BK, Pauli F, Rosenbloom KR, Sabo P, Safi A, Sanyal A, Shoresh N, Simon JM, Song L, Trinklein ND, Altshuler RC, Brown JB, Cheng C, Djebali S, Dong X, Ernst J, Furey T S, et al.: **An integrated encyclopedia of DNA elements in the human genome.** *Nature* 2012, **489**(7414):57–74. [http://dx.doi.org/10.1038/nature11247]
8. Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan KK, Cheng C, Mu XJ, Khurana E, Rozowsky J, Alexander R, Min R, Alves P, Abyzov A, Addleman N, Bhardwaj N, Boyle AP, Cayting P, Charos A, Chen DZ, Cheng Y, Clarke D, Eastman C, Euskirchen G, Frietze S, Fu Y, Gertz J, Grubert F, Harmanci A, Jain P, Kasowski M, et al.: **Architecture of the human regulatory network derived from ENCODE data.** *Nature* 2012, **489**(7414):91–100. [http://dx.doi.org/10.1038/nature11245]
9. Johnson DS, Mortazavi A, Myers RM, Wold B: **Genome-wide mapping of in vivo protein-DNA interactions.** *Science* 2007, **316**(5830):1497–1502. [http://dx.doi.org/10.1126/science.1141319]
10. Valouev A, Johnson DS, Sundquist A, Medina C, Anton E, Batzoglou S, Myers RM, Sidow A: **Genome-wide analysis of transcription factor binding sites based on ChIP-seq data.** *Nat Methods* 2008, **5**(9):829–834. [http://dx.doi.org/10.1038/nmeth.1246]
11. Anders S, Huber W: **Differential expression analysis for sequence count data.** *Genome Biol* 2010, **11**(10):R106+. [http://dx.doi.org/10.1186/gb-2010-11-10-r106]
12. Wendt KS, Yoshida K, Itoh T, Bando M, Koch B, Schirghuber E, Tsutsumi S, Nagae G, Ishihara K, Mishiro T, Yahata K, Imamoto F, Aburatani H, Nakao M, Imamoto N, Maeshima K, Shirahige K, Peters JMM: **Cohesin mediates transcriptional insulation by CCCTC-binding factor.** *Nature* 2008, **451**(7180):796–801. [http://dx.doi.org/10.1038/nature06634]
13. Ng HH, Robert F, Young RA, Struhl K: **Targeted recruitment of Set1 histone methylase by elongating Pol II provides a localized mark and memory of recent transcriptional activity.** *Mol Cell* 2003, **11**(3):709–719. [http://dx.doi.org/10.1016/S1097-2765(03)00092-3]
14. Chong JA, Tapia-Ramírez J, Kim S, Toledo-Aral JJ, Zheng Y, Boutros MC, Altshuller YM, Frohman MA, Kraner SD, Mandel G: **REST: a mammalian silencer protein that restricts sodium channel gene expression to neurons.** *Cell* 1995, **80**(6):949–957. [http://view.ncbi.nlm.nih.gov/pubmed/7697725]
15. Schoenherr CJ, Paquette AJ, Anderson DJ: **Identification of potential target genes for the neuron-restrictive silencer factor.** *P Natl Acad Sci USA* 1996, **93**(18):9881–9886. [http://www.pnas.org/content/93/18/9881.abstract]
16. Ballas N, Grunseich C, Lu DD, Speh JC, Mandel G: **REST and its corepressors mediate plasticity of neuronal gene chromatin throughout neurogenesis.** *Cell* 2005, **121**(4):645–657. [http://dx.doi.org/10.1016/j.cell.2005.03.013]
17. Rada-Iglesias A, Bajpai R, Swigut T, Brugmann SA, Flynn RA, Wysocka J: **A unique chromatin signature uncovers early developmental enhancers in humans.** *Nature* 2011, **470**(7333):279–283. [http://dx.doi.org/10.1038/nature09692]
18. Komarnitsky P, Cho EJ, Buratowski S: **Different phosphorylated forms of RNA polymerase II and associated mRNA processing factors during transcription.** *Gene Dev* 2000, **14**(19):2452–2460. [http://dx.doi.org/10.1101/gad.824700]
19. Takahashi H, Lassmann T, Murata M, Carninci P: **5′ end-centered expression profiling using cap-analysis gene expression and next-generation sequencing.** *Nat Protoc* 2012, **7**(3):542–561. [http://dx.doi.org/10.1038/nprot.2012.005]
20. Pruitt KD, Tatusova T, Brown GR, Maglott DR: **NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy.** *Nucleic Acids Res* 2012, **40**(Database issue):D130–D135. [http://dx.doi.org/10.1093/nar/gkr1079]
21. Saxonov S, Berg P, Brutlag DL: **A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters.** *Proc Natl Acad Sci USA* 2006, **103**(5):1412–1417. http://dx.doi.org/10.1073/pnas.0510310103]
22. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S: **Identifying a high fraction of the human genome to be under selective constraint using GERP++.** *PLoS Comput Biol* 2010, **6**(12):e1001025+. [http://dx.doi.org/10.1371/journal.pcbi.1001025]
23. Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, Wang W, Weng Z, Green RD, Crawford GE, Ren B: **Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome.** *Nat Genet* 2007, **39**(3):311–318. [http://dx.doi.org/10.1038/ng1966]
24. Waki H, Nakamura M, Yamauchi T, Wakabayashi Ki Yu, J, Hirose-Yotsuya L, Take K, Sun W, Iwabu M, Okada-Iwabu M, Fujita T, Aoyama T, Tsutsumi S, Ueki K, Kodama T, Sakai J, Aburatani H, Kadowaki T: **Global mapping of cell typespecific open chromatin by FAIRE-seq reveals the regulatory role of the NFI family in adipocyte differentiation.** *PLoS Genet* 2011, **7**(10):e1002311+. [http://dx.doi.org/10.1371/journal.pgen.1002311]
25. Bonn S, Zinzen RP, Girardot C, Gustafson EH, Perez-Gonzalez A, Delhomme N, Ghavi-Helm Y, Wilczynski B, Riddell A, Furlong EEM: **Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development.** *Nat Genet* 2012, **44**(2):148–156. [http://dx.doi.org/10.1038/ng.1064]
26. Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA, Frampton GM, Sharp PA, Boyer LA, Young RA, Jaenisch R: **Histone H3K27ac separates active from poised enhancers and predicts developmental state.** *P Natl Acad Sci USA* 2010, **107**(50):21931–21936. [http://dx.doi.org/10.1073/pnas.1016071107]
27. Saitou N, Nei M: **The neighbor-joining method: a new method for reconstructing phylogenetic trees.** *Mol Biol Evol* 1987, **4**(4):406–425. [http://mbe.oxfordjournals.org/content/4/4/406.abstract]
28. Igarashi K, Kataokat K, Itoh K, Hayashi N, Nishizawa M, Yamamoto M: **Regulation of transcription by dimerization of erythroid factor NF-E2 p45 with small Maf proteins.** *Nature* 1994, **367**(6463):568–572. [http://dx.doi.org/10.1038/367568a0]
29. Blackwood EM, Eisenman RN: **Max: a helix-loop-helix zipper protein that forms a sequence-specific DNA-binding complex with Myc.** *Science* 1991, **251**(4998):1211–1217. [http://view.ncbi.nlm.nih.gov/pubmed/2006410]
30. Zervos AS, Gyuris J, Brent R: **Mxi1, a protein that specifically interacts with Max to bind Myc-Max recognition sites.** *Cell* 1993, **72**(2):223–232. [http://view.ncbi.nlm.nih.gov/pubmed/8425219]
31. Chodosh LA, Olesen J, Hahn S, Baldwin AS, Guarente L, Sharp PA: **A yeast and a human CCAAT-binding protein have heterologous subunits that are functionally interchangeable.** *Cell* 1988, **53:**25–35.
32. Hooft van Huijsduijnen R, Li XY, Black D, Matthes H, Benoist C, Mathis D: **Co-evolution from yeast to mouse: cDNA cloning of the two NF-Y (CP-1/CBF) subunits.** *EMBO J* 1990, **9**(10):3119–3127. [http://www.ncbi.nlm.nih.gov/pmc/articles/PMC552040/]
33. Sirito M, Walker S, Lin Q, Kozlowski MT, Klein WH, Sawadogo M: **Members of the USF family of helix-loop-helix proteins bind DNA as homo- as well as heterodimers.** *Gene Expression* 1992, **2**(3):231–240. [http://view.ncbi.nlm.nih.gov/pubmed/1450663]
34. van Dam H, Castellazzi M: **Distinct roles of Jun : Fos and Jun : ATF dimers in oncogenesis.** *Oncogene* 2001, **20**(19):2453–2464. [http://dx.doi.org/10.1038/sj.onc.1204239]
35. Ravasi T, Suzuki H, Cannistraci CVV, Katayama S, Bajic VB, Tan K, Akalin A, Schmeier S, Kanamori-Katayama M, Bertin N, Carninci P, Daub CO, Forrest AR, Gough J, Grimmond S, Han JHH, Hashimoto T, Hide W, Hofmann O, Kamburov A, Kaur M, Kawaji H, Kubosaki A, Lassmann T, van Nimwegen, E, MacPherson CRR, Ogawa C, Radovanovic A, Schwartz A, Teasdale R D, et al.: **An atlas of combinatorial transcriptional regulation in mouse and man.** *Cell* 2010, **140**(5):744–752. [http://dx.doi.org/10.1016/j.cell.2010.01.044]
36. Feuerstein N, Firestein R, Aiyar N, He X, Murasko D, Cristofalo V: **Late induction of CREB/ATF binding and a concomitant increase in cAMP levels in T and B lymphocytes stimulated via the antigen receptor.** *J Immunol* 1996, **156**(12):4582–4593. [http://view.ncbi.nlm.nih.gov/pubmed/8648100]
37. Ise W, Kohyama M, Schraml BU, Zhang T, Schwer B, Basu U, Alt FW, Tang J, Oltz EM, Murphy TL, Murphy KM: **The transcription factor BATF controls the global regulators of class-switch recombination in both B cells and T cells.** *Nat Immunol* 2011, **12**(6):536–543. [http://dx.doi.org/10.1038/ni.2037]
38. Zhang Q, Didonato JA, Karin M, McKeithan TW: **BCL3 encodes a nuclear protein which can alter the sulymphocyteular location of NF-kappa B proteins.** *Mol Cell Biol* 1994, **14**(6):3915–3926. [http://www.ncbi.nlm.nih.gov/pmc/articles/PMC358758/]

39. Liu P, Keller JR, Ortiz M, Tessarollo L, Rachel RA, Nakamura T, Jenkins NA, Copeland NG: **Bcl11a is essential for normal lymphoid development.** *Nat Immunol* 2003, **4**(6):525–532. [http://dx.doi.org/10.1038/ni925]
40. McPherson JP, Sarras H, Lemmers B, Tamblyn L, Migon E, Matysiak-Zablocki E, Hakem A, Azami SA, Cardoso R, Fish J, Sanchez O, Post M, Hakem R: **Essential role for Bclaf1 in lung development and immune system function.** *Cell Death Differ* 2008, **16**(2):331–339. [http://dx.doi.org/10.1038/cdd.2008.167]
41. Seimiya M, Bahar R, Wang Y, Kawamura K, Tada Y, Okada S, Hatano M, Tokuhisa T, Saisho H, Watanabe T, Tagawa M, O-Wang J: **Clast5/Stra13 is a negative regulator of B lymphocyte activation.** *Biochem Bioph Res Co* 2002, **292:**121–127. [http://view.ncbi.nlm.nih.gov/pubmed/11890681]
42. Hagman J, Travis A, Grosschedl R: **A novel lineage-specific nuclear factor regulates mb-1 gene transcription at the early stages of B cell differentiation.** *EMBO J* 1991, **10**(11):3409–3417. [http://www.ncbi.nlm.nih.gov/pmc/articles/PMC453069/]
43. Klein U, Casola S, Cattoretti G, Shen Q, Lia M, Mo T, Ludwig T, Rajewsky K, Dalla-Favera R: **Transcription factor IRF4 controls plasma cell differentiation and class-switch recombination.** *Nat Immunol* 2006, **7**(7):773–782. [http://dx.doi.org/10.1038/ni1357]
44. Meixner A, Karreth F, Kenner L, Wagner EF: **JunD regulates lymphocyte proliferation and T helper cell cytokine expression.** *EMBO J* 2004, **23**(6):1325–1335. [http://dx.doi.org/10.1038/sj.emboj.7600133]
45. Wilker PR, Kohyama M, Sandau MM, Albring JC, Nakagawa O, Schwarz JJ, Murphy KM: **Transcription factor Mef2c is required for B cell proliferation and survival after antigen receptor stimulation.** *Nat Immunol* 2008, **9**(6):603–612. [http://dx.doi.org/10.1038/ni.1609]
46. Fujita N, Jaye DL, Geigerman C, Akyildiz A, Mooney MR, Boss JM, Wade PA: **MTA3 and the Mi-2/NuRD Complex Regulate Cell Fate during B Lymphocyte Differentiation.** *Cell* 2004, **119:**75–86. [http://dx.doi.org/10.1016/j.cell.2004.09.014]
47. Peng SL, Gerth AJ, Ranger AM, Glimcher LH: **NFATc1 and NFATc2 together control both T and B cell activation and differentiation.** *Immunity* 2001, **14:**13–20. [http://view.ncbi.nlm.nih.gov/pubmed/11163226]
48. Cobaleda C, Schebesta A, Delogu A, Busslinger M: **Pax5: the guardian of B cell identity and function.** *Nat Immunol* 2007, **8**(5):463–470. [http://dx.doi.org/10.1038/ni1454]
49. Wang ZG, Delva L, Gaboli M, Rivi R, Giorgio M, Cordon-Cardo C, Grosveld F, Pandolfi PP: **Role of PML in Cell Growth and the Retinoic Acid Pathway.** *Science* 1998, **279**(5356):1547–1551. [http://dx.doi.org/10.1126/science.279.5356.1547]
50. Corcoran LM, Karvelas M: **Oct-2 is required early in T cell-independent B cell activation for G1 progression and for proliferation.** *Immunity* 1994, **1**(8):635–645. [http://view.ncbi.nlm.nih.gov/pubmed/7600291]
51. Woolf E, Xiao C, Fainaru O, Lotem J, Rosen D, Negreanu V, Bernstein Y, Goldenberg D, Brenner O, Berke G, Levanon D, Groner Y: **Runx3 and Runx1 are required for CD8 T cell development during thymopoiesis.** *P Natl Acad Sci USA* 2003, **100**(13):7731–7736. [http://dx.doi.org/10.1073/pnas.1232420100]
52. Stephensen CB, Borowsky AD, Lloyd KKC: **Disruption of Rxra gene in thymocytes and T lymphocytes modestly alters lymphocyte frequencies, proliferation, survival and T helper type 1/type 2 balance.** *Immunology* 2007, **121**(4):484–498. [http://dx.doi.org/10.1111/j.1365-2567.2007.02595.x]
53. McKercher SR, Torbett BE, Anderson KL, Henkel GW, Vestal DJ, Baribault H, Klemsz M, Feeney AJ, Wu GE, Paige CJ, Maki RA: **Targeted disruption of the PU.1 gene results in multiple hematopoietic abnormalities.** *EMBO J* 1996, **15**(20):5647–5658. [http://www.ncbi.nlm.nih.gov/pmc/articles/PMC452309/]
54. Takeda K, Kaisho T, Yoshida N, Takeda J, Kishimoto T, Akira S: **Stat3 activation is responsible for IL-6-dependent T cell proliferation through preventing apoptosis: generation and characterization of T cell-specific Stat3-deficient mice.** *J Immunol* 1998, **161**(9):4652–4660. [http://view.ncbi.nlm.nih.gov/pubmed/9794394]
55. Welte T, Leitenberg D, Dittel BN, al Ramadi BK, Xie B, Chin YE, Janeway CA, Bothwell ALM, Bottomly K, Fu XY: **STAT5 Interaction with the T Cell Receptor Complex and Stimulation of T Cell Proliferation.** *Science* 1999, **283**(5399):222–225. [http://dx.doi.org/10.1126/science.283.5399.222]
56. Bain G, Maandag EC, Izon DJ, Amsen D, Kruisbeek AM, Weintraub BC, Krop I, Schlissel MS, Feeney AJ, van Roon: **E2A proteins are required for proper B cell development and initiation of immunoglobulin gene rearrangements.** *Cell* 1994, **79**(5):885–892. [http://view.ncbi.nlm.nih.gov/pubmed/8001125]
57. Zhuang Y, Cheng P, Weintraub H: **B-lymphocyte development is regulated by the combined dosage of three basic helix-loop-helix genes, E2A, E2-2, and HEB.** *Mol Cell Biol* 1996, **16**(6):2898–2905. [http://mcb.asm.org/content/16/6/2898.abstract]
58. Swanson BJ, Jäck HM, Lyons GE: **Characterization of myocyte enhancer factor 2 (MEF2) expression in B and T cells: MEF2C is a B cell-restricted transcription factor in lymphocytes.** *Mol Immunol* 1998, **35**(8):445–458. [http://dx.doi.org/10.1016/S0161-5890(98)00058-3]
59. Shen M, Kawamoto T, Yan W, Nakamasu K, Tamagami M, Koyano Y, Noshiro M, Kato Y: **Molecular characterization of the novel basic helix-loop-helix protein DEC1 expressed in differentiated human embryo chondrocytes.** *Biochem Bioph Res Co* 1997, **236**(2):294–298. [http://dx.doi.org/10.1006/bbrc.1997.6960]
60. Hu JS, Olson EN, Kingston RE: **HEB, a helix-loop-helix protein related to E2A and ITF2 that can modulate the DNA-binding ability of myogenic regulatory factors.** *Mol Cell Biol* 1992, **12**(3):1031–1042. [http://www.ncbi.nlm.nih.gov/pmc/articles/PMC369535/]
61. Ayer DE, Lawrence QA, Eisenman RN: **Mad-Max transcriptional repression is mediated by ternary complex formation with mammalian homologs of yeast repressor Sin3.** *Cell* 1995, **80**(5):767–776. [http://view.ncbi.nlm.nih.gov/pubmed/7889570]
62. Sealy L, Malone D, Pawlak M: **Regulation of the cfos serum response element by C/EBPβ.** *Mol Cell Biol* 1997, **17**(3):1744–1755. [http://www.ncbi.nlm.nih.gov/pmc/articles/PMC231899/]
63. Hanlon M, Sealy L: **Ras regulates the association of serum response factor and CCAAT/enhancer-binding protein β.** *J Biol Chem* 1999, **274**(20):14224–14228. [http://dx.doi.org/10.1074/jbc.274.20.14224]
64. Roder K, Wolf SS, Larkin KJ, Schweizer M: **Interaction between the two ubiquitously expressed transcription factors NF-Y and Sp1.** *Gene* 1999, **234:**61–69. [http://dx.doi.org/10.1016/S0378-1119(99)00180-8]
65. Yip K, Cheng C, Bhardwaj N, Brown J, Leng J, Kundaje A, Rozowsky J, Birney E, Bickel P, Snyder M, Gerstein M: **Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors.** *Genome Biol* 2012, **13**(9):R48+. [http://dx.doi.org/10.1186/gb-2012-13-9-r48]
66. Wang J, Zhuang J, Iyer S, Lin X, Whitfield TW, Greven MC, Pierce BG, Dong X, Kundaje A, Cheng Y, Rando OJ, Birney E, Myers RM, Noble WS, Snyder M, Weng Z: **Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors.** *Genome Res* 2012, **22**(9):1798–1812. [http://dx.doi.org/10.1101/gr.139105112]
67. Rhee HSS, Pugh BF: **Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution.** *Cell* 2011, **147**(6):1408–1419. [http://dx.doi.org/10.1016/j.cell.2011.11.013]
68. Ouyang Z, Zhou Q, Wong WH: **ChIP-seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells.** *P Natl Acad Sci USA* 2009, **106**(51):21521–21526. [http://dx.doi.org/10.1073/pnas.0904863106]
69. Cheng C, Gerstein M: **Modeling the relative relationship of transcription factor binding and histone modifications to gene expression levels in mouse embryonic stem cells.** *Nucleic Acids Res* 2011, **40**(2):553–568. [http://dx.doi.org/10.1093/nar/gkr752]
70. Cheng C, Alexander R, Min R, Leng J, Yip KY, Rozowsky J, Yan KK, Dong X, Djebali S, Ruan Y, Davis CA, Carninci P, Lassman T, Gingeras TR, Guigó R, Birney E, Weng Z, Snyder M, Gerstein M: **Understanding transcriptional regulation by integrative analysis of transcription factor binding data.** *Genome Res* 2012, **22**(9):1658–1667. [http://dx.doi.org/10.1101/gr.136838111]
71. Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM: **A census of human transcription factors: function, expression and evolution.** *Nat Rev Genet* 2009, **10**(4):252–263. [http://dx.doi.org/10.1038/nrg2538]
72. Junion G, Spivakov M, Girardot C, Braun M, Gustafson EH, Birney E, Furlong EEM: **A transcription factor collective defines cardiac cell fate and reflects lineage history.** *Cell* 2012, **148**(3):473–486. [http://dx.doi.org/10.1016/j.cell.2012.01.030]
73. Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CAM, Taylor MS, Engstrom PG, Frith MC, Forrest ARR, Alkema WB,

Tan SL, Plessy C, Kodzius R, Ravasi T, Kasukawa T, Fukuda S, Kanamori-Katayama M, Kitazume Y, Kawaji H, Kai C, Nakamura M, Konno H, Nakano K, Mottagui-Tabar S, Arner P, Chesi A, Gustincich S, Persichetti F, Suzuki H, Grimmond SM, Wells CA, Orlando V, Wahlestedt C, Liu ET, Harbers M, Kawai J, Bajic VB, Hume DA, Hayashizaki Y: **Genome-wide analysis of mammalian promoter architecture and evolution.** *Nat Genet* 2006, **38**(6):626–635. [http://dx.doi.org/10.1038/ng1789]

74. Yang C, Bolotin E, Jiang T, Sladek FM, Martinez E: **Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters.** *Gene* 2007, **389:**52–65. [http://dx.doi.org/10.1016/j.gene.2006.09.029]
75. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics* 2009, **25**(14):1754–1760. [http://dx.doi.org/10.1093/bioinformatics/btp324]
76. Müllner D: **fastcluster: fast hierarchical clustering routines for R and Python.** 2011. [http://math.stanford.edu/~muellner/fastcluster.html]
77. Simonsen M, Mailund T, Pedersen CNS: **Rapid neighbor-joining algorithms in bioinformatics.** In *Algorithms in Bioinformatics, Volume 5251 of Lecture Notes in Computer Science*. Edited by Crandall KA, Lagergren J. Berlin, Heidelberg: Springer Berlin/Heidelberg; 2008:113–122. [http://dx.doi.org/10.1007/978-3-540-87361-7_10]
78. DNAnexus Inc: *RNA-seq/3SEQ transcriptome based quantification*; 2010.
79. Bailey TL, Gribskov M: **Combining evidence using p-values: application to sequence homology searches.** *Bioinformatics* 1998, **14:**48–54. [http://dx.doi.org/10.1093/bioinformatics/14.1.48]
80. Mevik BH, Wehrens R: **The pls package: principal component and partial least squares regression in R.** *J Stat Softw* 2007, **18**(2):1–24. [http://www.jstatsoft.org/v18/i02]

doi:10.1186/1471-2164-14-720
Cite this article as: Foley and Sidow: Transcription-factor occupancy at HOT regions quantitatively predicts RNA polymerase recruitment in five human cell lines. *BMC Genomics* 2013 **14**:720.